# Peru is Multilingual, Its Machine Translation Should Be Too?

**Arturo Oncevay**
School of Informatics, ILCC
University of Edinburgh
a.oncevay@ed.ac.uk

## Abstract

Peru is a multilingual country with a long history of contact between the indigenous languages and Spanish. Taking advantage of this context for machine translation is possible with multilingual approaches for learning both unsupervised subword segmentation and neural machine translation models. The study proposes the first multilingual translation models for four languages spoken in Peru: Aymara, Ashaninka, Quechua and Shipibo-Konibo, providing both many-to-Spanish and Spanish-to-many models and outperforming pairwise baselines in most of them. The task exploited a large English-Spanish dataset for pretraining, monolingual texts with tagged back-translation, and parallel corpora aligned with English. Finally, by fine-tuning the best models, we also assessed the out-of-domain capabilities in two evaluation datasets for Quechua and a new one for Shipibo-Konibo[1].

## 1 Introduction

Neural Machine Translation (NMT) has opened several research directions to exploit as many and diverse data as possible. Massive multilingual NMT models, for instance, take advantage of several language-pair datasets in a single system (Johnson et al., 2017). This offers several advantages, such as a simple training process and enhanced performance of the language-pairs with little data (although sometimes detrimental to the high-resource language-pairs). However, massive models of dozens of languages are not necessarily the best outcome, as it is demonstrated that smaller clusters still offer the same benefits (Tan et al., 2019; Oncevay et al., 2020).

Peru offers a rich diversity context for machine translation research with 47 native languages (Simons and Fenning, 2019). All of them are highly distinguishing from Castilian Spanish, the primary official language in the country and the one spoken by the majority of the population. However, from the computational perspective, all of these languages do not have enough resources, such as monolingual or parallel texts, and most of them are considered endangered (Zariquiey et al., 2019).

In this context, the main question then arises: shouldn't machine translation be multilingual for languages spoken in a multilingual country like Peru? By taking advantage of few resources, and other strategies such as multilingual unsupervised subword segmentation models (Kudo, 2018), pretraining with high resource language-pairs (Kocmi and Bojar, 2018), back-translation (Sennrich et al., 2016a), and fine-tuning (Neubig and Hu, 2018), we deployed the first many-to-one and one-to-many multilingual NMT models (paired with Spanish) for four indigenous languages: Aymara, Ashaninka, Quechua and Shipibo-Konibo.

## 2 Related work

In Peru, before NMT, there were studies in rule-based MT, based on the Apertium platform (Forcada et al., 2011), for Quechua Eastern Apurimac (*qve*) and Quechua Cuzco (*quz*) (Cavero and Madariaga, 2007). Furthermore, Ortega and Pillaipakkamnatt (2018) improved alignments for *quz* by using an agglutinative language as Finnish as a pivot. Apart from the Quechua variants, only Aymara (Coler and Homola, 2014) and Shipibo-Konibo (Galarreta et al., 2017) have been addressed with rule-based and statistical MT, respectively.

Ortega et al. (2020b) for Southern Quechua, and Gómez Montoya et al. (2019) for Shipibo-Konibo, are the only studies that employed sequence-to-sequence NMT models. They also performed transfer learning experiments with potentially related language pairs (e.g. Finnish or Turkish, which are agglutinative languages). However, as far as we know, this is the first study that trains a multilingual model for some language spoken in Peru. For

[1] Available in: https://github.com/aoncevay/mt-peru

*Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 194–201
June 11, 2021. ©2021 Association for Computational Linguistics

related work on multilingual NMT, we refer the readers to the survey of Dabre et al. (2020).

## 3 Languages and datasets

To enhance replicability, we only used the datasets provided in the AmericasNLP Shared Task[2].

- **Southern Quechua**: with 6+ millions of speakers and several variants, it is the most widespread indigenous language in Peru. AmericasNLP provides evaluation sets in the standard Southern Quechua, which is based mostly on the Quechua Ayacucho (quy) variant. There is parallel data from dictionaries and Jehovah Witnesses (Agić and Vulić, 2019). There is parallel corpus aligned with English too. We also include the close variant of Quechua Cusco (quz) to support the multilingual learning.
- **Aymara** (aym): with 1.7 million of speakers (mostly in Bolivia). The parallel and monolingual data is extracted from a news website (Global Voices) and distributed by OPUS (Tiedemann, 2012). There are aligned data with English too.
- **Shipibo-Konibo** (shp): a Panoan language with almost 30,000 speakers in the Amazonian region. There are parallel data from dictionaries, educational material (Galarreta et al., 2017), language learning flashcards (Gómez Montoya et al., 2019), plus monolingual data from educational books (Bustamante et al., 2020).
- **Ashaninka** (cni): an Arawakan language with 45,000 speakers in the Amazon. There is parallel data from dictionaries, laws and books (Ortega et al., 2020a), plus monolingual corpus (Bustamante et al., 2020).

The four languages are highly agglutinative or polysynthetic, meaning that they usually express a large amount of information in just one word with several joint morphemes. This is a real challenge for MT and subword segmentation methods, given the high probability of addressing a "rare word" for the system. We also note that each language belongs to a different language family, but that is not a problem for multilingual models, as usually the family-based clusters are not the most effective ones (Oncevay et al., 2020).

| Language | Mono. | es | en |
|---|---|---|---|
| aym - Aymara | 8,680 | 5,475 | 5,045 |
| cni - Ashaninka | 13,193 | 3,753 | |
| quy - Quechua | | 104,101 | 14,465 |
| shp - Shipibo-Konibo | 23,593 | 14,437 | |
| quz - Quechua Cusco | | 97,836 | 21,760 |

Table 1: Number of sentences in monolingual and parallel corpora aligned with Spanish (es) or English (en). The latter are used for en→es translation and we only noted non-duplicated sentences w.r.t. the *–es corpora.

**Pre-processing** The datasets were noisy and not cleaned. Lines are reduced according to several heuristics: Arabic numbers or punctuation do not match in the parallel sentences, there are more symbols or numbers than words in a sentence, the ratio of words from one side is five times larger or shorter than the other, among others. Table 5 in the Appendix includes the original and cleaned data size per language-pair, whereas Table 1 presents the final sizes.

**English-Spanish datasets** We consider the EuroParl (1.7M sentences) (Koehn, 2005) and the NewsCommentary-v8 (174k sentences) corpora for pre-training.

## 4 Methodology

### 4.1 Evaluation

The train data have been extracted from different domains and sources, which are not necessarily the same as the evaluation sets provided for the Shared Task. Therefore, the official development set (995 sentences per language) is split into three parts: 25%-25%-50%. The first two parts are our custom dev and devtest sets[3]. We add the 50% section to the training set with a sampling distribution of 20%, to reduce the domain gap in the training data. Likewise, we extract a sample of the training and double the size of the development set. The mixed data in the validation set is relevant, as it allows to evaluate how the model fits with all the domains. We used the same multi-text sentences for evaluation, and avoid any overlapping of the Spanish side with the training set, this is also important as we are going to evaluate multilingual models. Evaluation for all the models used BLEU (Papineni et al., 2002) and chrF (Popović, 2015) metrics.

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

[3] We are also reporting the results on the official test sets after the finalisation of the Shared Task.

| BLEU | Aymara | | | Ashaninka | | | Quechua | | | Shipibo-Konibo | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **→Spanish** | dev | devtest | test | dev | devtest | test | dev | devtest | test | dev | devtest | test |
| (a) Multilingual | **11.11** | **9.95** | 3.70 | **8.40** | **9.37** | **5.21** | 12.46 | 11.03 | 8.04 | **10.34** | **12.72** | **10.07** |
| (b) Multi+BT | 10.76 | 8.39 | 2.87 | 7.30 | 5.34 | 3.44 | 11.48 | 8.85 | 7.51 | 9.13 | 10.77 | 7.58 |
| (c) Multi+BT[t] | 10.72 | 8.42 | 2.86 | 7.45 | 5.69 | 3.15 | 11.37 | 10.02 | 7.12 | 8.81 | 10.73 | 7.18 |
| (d) Pairwise | 9.46 | 7.66 | 2.04 | 4.23 | 3.96 | 2.38 | **15.21** | **14.00** | **8.20** | 7.72 | 9.48 | 4.44 |
| **Spanish→** | dev | devtest | test | dev | devtest | test | dev | devtest | test | dev | devtest | test |
| (e) Multilingual | 8.67 | 6.28 | 2.19 | 6.74 | 11.72 | **5.54** | 10.04 | 5.37 | **4.51** | 10.82 | 10.44 | 6.69 |
| (f) Multi+BT | 3.31 | 2.59 | 0.79 | 1.29 | 3.38 | 2.82 | 1.36 | 2.02 | 1.73 | 1.63 | 3.76 | 2.98 |
| (g) Multi+BT[t] | **10.55** | **6.54** | **2.31** | **7.36** | **13.17** | 5.40 | **10.77** | 5.29 | 4.23 | **11.98** | **11.12** | **7.45** |
| (h) Pairwise | 7.08 | 4.96 | 1.65 | 4.12 | 8.40 | 3.82 | 10.67 | **6.11** | 3.96 | 8.76 | 7.89 | 6.15 |

Table 2: BLEU scores for the dev and devtest custom partitions and the official test set, including all the multilingual and pairwise MT systems into and from Spanish. BT = Back-translation. BT[t] = Tagged back-translation.

| chrF | Aymara | | | Ashaninka | | | Quechua | | | Shipibo-Konibo | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **→Spanish** | dev | devtest | test | dev | devtest | test | dev | devtest | test | dev | devtest | test |
| (a) Multilingual | **31.73** | **28.82** | **22.01** | **26.78** | **26.82** | **22.27** | 32.92 | 32.99 | 29.45 | **31.41** | **33.49** | **31.26** |
| (d) Pairwise | 28.77 | 25.03 | 19.79 | 20.43 | 20.40 | 18.83 | **36.01** | **36.06** | **30.90** | 27.25 | 29.91 | 25.31 |
| **Spanish→** | dev | devtest | test | dev | devtest | test | dev | devtest | test | dev | devtest | test |
| (g) Multi+BT[t] | **37.32** | **35.17** | **26.70** | **38.94** | **38.44** | **30.81** | 44.60 | 38.94 | **37.80** | **40.67** | **39.47** | **33.43** |
| (h) Pairwise | 28.89 | 28.23 | 21.13 | 32.55 | 32.29 | 27.10 | **45.77** | **39.68** | 36.86 | 34.97 | 34.96 | 27.09 |

Table 3: chrF scores for the dev and devtest custom partitions and the official test sets for the best multilingual setting and the pairwise baseline in each direction.

### 4.2 Multilingual subword segmentation

Ortega et al. (2020b) used morphological information, such as affixes, to guide the Byte-Pair-Encoding (BPE) segmentation algorithm (Sennrich et al., 2016b) for Quechua. However, their improvement is not significant, and according to Bostrom and Durrett (2020), BPE tends to oversplit roots of infrequent words. They showed that a unigram language model (Kudo, 2018) seems like a better alternative to split affixes and preserve roots (in English and Japanese).

To take advantage of the potential lexical sharing of the languages (e.g. loanwords) and address the polysynthetic nature of the indigenous languages, we trained a unique multilingual segmentation model by sampling all languages with a uniform distribution. We used the unigram model implementation in SentencePiece (Kudo and Richardson, 2018) with a vocabulary size of 32,000.

### 4.3 Procedure

For the experiments, we used a Transformer-base model (Vaswani et al., 2017) with the default configuration in Marian NMT (Junczys-Dowmunt et al., 2018). The steps are as follows:

**Pre-training** We pre-trained two MT models with the Spanish–English language-pair in both directions. We did not include an agglutinative language like Finnish (Ortega et al., 2020b) for two reasons: it is not a must to consider highly related languages for effective transfer learning (e.g. English–German to English–Tamil (Bawden et al., 2020)), and we wanted to translate the English side of en–aym, en–quy and en–quz to augment their correspondent Spanish-paired datasets. The en→es and es→en models achieved 34.4 and 32.3 BLEU points, respectively, in the newsdev2013 set.

**Multilingual fine-tuning** Using the pre-trained en→es model, we fine-tuned the first multilingual model many-to-Spanish. Following established practices, we used a uniform sampling for all the datasets (quz–es included) to avoid under-fitting the low-resource language-pairs[4]. Results are in Table 2, row (a). We replicated this to the es→many direction (row (e)), using the es→en model.

**Back-translation** With model (a), we back-translated (BT) the monolingual data of the indigenous languages and train models (b) and (f): original plus BT data. However, the results with BT data underperformed or did not converge. Potential reasons are the noisy translation outputs of model (a) and the larger amount of BT than human-translated sentences for all languages, even though

[4] Temperature-based sampling or automatically learned data scorers are more advanced strategies (Wang et al., 2020). However, we left that analysis for further work.

we sampled BT and human translations uniformly.

**Tagged back-translation (BT[t])** To alleviate the issue, we add a special tag for the BT data (Caswell et al., 2019). With BT[t], we send a signal to the model that it is processing synthetic data, and thus, it may not hurt the learning over the real data. Table 2 (rows (c,g)) shows the results.

**Pairwise baselines** We obtained pairwise systems by fine-tuning the same pre-trained models (without any back-translated data). For a straightforward comparison, they used the same multilingual SentencePiece model.

## 5 Analysis and discussion

One of the most exciting outcomes is the deteriorated performance of the multilingual models using BT data, as we usually expect that added back-translated texts would benefit performance. Using tags (BT[t]) to differentiate which data is synthetic or not is only a simple step to address this issue; however, there could be evaluated more informed strategies for denoising or performing online data selection (Wang et al., 2018).

Besides, in the translation into Spanish, the multilingual model without BT data outperforms the rest models in all languages but Quechua, where the pairwise system achieved the best translation accuracy. Quechua is the "highest"-resource language-pair in the experiment, and its performance is deteriorated in the multilingual setting[5]. A similar scenario is shown in the other translation direction from Spanish, where the best multilingual setting (+BT[t]) cannot overcome the es→quy model in the devtest set.

Nevertheless, the gains for Aymara, Ashaninka and Shipibo-Konibo are outstanding. Moreover, we note that the models are not totally overfitted to any of the evaluation sets. Exceptions are es→aym and es→quy, with a significant performance dropping from dev to devtest, meaning that it started to overfit to the training data. However, for Spanish→Ashaninka, we observe that the model achieved a better performance in the devtest set. This is due to oversampling of the same-domain dev partition for training (§4.1) and the small original training set.

| Stories (shp) | shp→es | | | es→shp | | |
|---|---|---|---|---|---|---|
| | full | half | Δt | full | half | Δt |
| BestMulti | 1.90 | 1.43 | 0 | 0.56 | 0.68 | 0 |
| BestMulti+FT | - | **5.73** | -1.66 | - | **5.82** | -1.93 |
| Magazine (quy) | quy→es | | | es→quy | | |
| | full | half | Δt | full | half | Δt |
| Pairwise | 2.96 | 2.32 | 0 | 2.17 | 1.59 | 0 |
| Pairwise+FT | - | **9.14** | -0.83 | - | **2.92** | +0.78 |
| Apertium | 5.82 | | | - | - | |
| Ortega et al. | 0.70 | | | - | - | |

Table 4: Out-of-domain BLEU scores. Best model is fine-tuned (+FT) with half of the dataset and evaluated in the other half. Δt = original test score variation.

Concerning the results on the official test set, the performance is lower than the results with the custom evaluation sets. The main potential reason is that the official test is four times bigger than the custom devtest, and therefore, offers more diversity and challenge for the evaluation. Another point to highlight is that the best result in the Spanish–Quechua language-pair is obtained by a multilingual model (the scores between the model (e) and (g) are not significantly different) instead of the pairwise baseline.

Decoding an indigenous language is still a challenging task, and the relatively low BLEU scores cannot suggest a translation with proper adequacy or fluency. However, BLEU works at the word-level, and other character-level metrics should be considered to better assess the highly agglutinative nature of the languages. For reference, we also report the chrF scores in Table 3 for the best multilingual setting and the pairwise baseline. As for the Spanish decoding, fluency is preserved from the English→Spanish pre-trained model[6], but more adequacy is needed.

## 6 Out-of-domain evaluation

It is relevant to assess out-of-domain capabilities, but more important to evaluate whether the models are still capable to fine-tune without overfitting. We use a small evaluation set for Quechua (*Kallpa*, with 100 sentences), which contains sentences extracted from a magazine (Ortega et al., 2020b). Likewise, we introduce a new evaluation set for Shipibo-Konibo (*Kirika*, 200 sentences), which contains short traditional stories.

We tested our best model for each language-pair, fine-tune it (+FT) with half of the out-of-domain

[5] In multilingual training, this behaviour is usually observed, and other approaches, such as injecting adapter layers (Bapna and Firat, 2019), might help to mitigate the issue. We left the analysis for further work.

[6] This might be confirmed by a proper human evaluation

dataset, and evaluate it in the other half. To avoid overfitting, we controlled cross-entropy loss and considered very few updates for validation steps. Results are shown in Table 3, where we observe that it is possible to fine-tune the multilingual or pairwise models to the new domains without loosing too much performance in the original test.

The Quechua translations rapidly improved with the fine-tuning step, and there is a small gain in the original test for es→quy, although the scores are relatively low in general. Nevertheless, our model could outperform others (by extrapolation, we can assume that the scores for the rule-based Apertium system (Cavero and Madariaga, 2007) and Ortega et al. (2020b)'s NMT system are similar in half of the dataset).

For Shipibo-Konibo, we also observe some small gains in both directions without hurting the previous performance, but the scores are far from being robust. *Kirika* is challenging given its *old style*: the translations are extracted from an old book written by missionaries, and even when the spelling has been modernised, there are differences in the use of some auxiliary verbs for instance (extra words that affect the evaluation metric)[7].

## 7 Conclusion and future work

Peru is multilingual, *ergo*, its machine translation should be too! We conclude that multilingual machine translation models can enhance the performance in truly low-resource languages like Aymara, Ashaninka and Shipibo-Konibo, in translation from and into Spanish. For Quechua, even when the pairwise system performed better in this study, there is a simple step to give a multilingual setting another opportunity: to include a higher-resource language-pair that may support the multilingual learning process. This could be related in some aspect like morphology (another agglutinative language) or the discourse (domain). Other approaches focused on more advanced sampling or adding specific layers to restore the performance of the higher-resource languages might be considered as well. Besides, tagged back-translation allowed to take some advantage of the monolingual data; however, one of the most critical following steps is to obtain a more robust many-to-Spanish model to generate back-translated data with more quality. Furthermore, to address the multi-domain nature of these datasets, we could use domain tags to send more signals to the model and support further fine-tuning steps. Finally, after addressing the presented issues in this study, and to enable zero-shot translation, we plan to train the first many-to-many multilingual model for indigenous languages spoken in Peru.

## Acknowledgements

The author is supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 825299 (GoURMET) and the EPSRC fellowship grant EP/S001271/1 (MTStretch). The author is also thankful to the insightful feedback of the anonymous reviewers.

## 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.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Rachel Bawden, Alexandra Birch, Radina Dobreva, Arturo Oncevay, Antonio Valerio Miceli Barone, and Philip Williams. 2020. The University of Edinburgh's English-Tamil and English-Inuktitut submissions to the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 92–99, Online. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

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 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of*

[7] The dataset, with further analysis, is available at: https://github.com/aoncevay/mt-peru

*the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Indhira Castro Cavero and Jaime Farfán Madariaga. 2007. Traductor morfológico del castellano y quechua (Morphological translator of Castilian Spanish and Quechua). *Revista I+ i*, 1(1).

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

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A comprehensive survey of multilingual neural machine translation.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

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.

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.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

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.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

John Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020a. 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.

John Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like Quechua and Finnish to aid in low-resource translation. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 1–11.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020b. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Gary F. Simons and Charles D. Fenning, editors. 2019. *Ethnologue: Languages of the World. Twenty-second edition*. Dallas Texas: SIL international. Online version: http://www.ethnologue.com.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China. Association for Computational Linguistics.

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.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Roberto Zariquiey, Harald Hammarström, Mónica Arakaki, Arturo Oncevay, John Miller, Aracelli García, and Adriano Ingunza. 2019. Obsolescencia lingüística, descripción gramatical y documentación de lenguas en el perú: hacia un estado de la cuestión. *Lexis*, 43(2):271–337.

# Appendix

| | $S$ (orig.) | $S$ (clean) | % clean | $T/S$ (src) | $T/S$ (tgt) | ratio $T$ src/tgt |
|---|---|---|---|---|---|---|
| es-aym | 6,453 | 5,475 | -15.16% | 19.27 | 13.37 | 1.44 |
| es-cni | 3,860 | 3,753 | -2.77% | 12.29 | 6.52 | 1.89 |
| es-quy | 128,583 | 104,101 | -19.04% | 14.2 | 8.17 | 1.74 |
| es-shp | 14,511 | 14,437 | -0.51% | 6.05 | 4.31 | 1.4 |
| es-quz | 130,757 | 97,836 | -25.18% | 15.23 | 8.62 | 1.77 |
| en-quy | 128,330 | 91,151 | -28.97% | 15.03 | 8.68 | 1.73 |
| en-quz | 144,867 | 100,126 | -30.88% | 14.84 | 8.42 | 1.76 |
| en-aym | 8,886 | 7,689 | -13.47% | 19.36 | 13.32 | 1.45 |

Table 5: Statistics and cleaning for all parallel corpora. We observe that the Shipibo-Konibo and Ashaninka corpora are the least noisy ones. $S$ = number of sentences, $T$ = number of tokens. There are sentence alignment issues in the Quechua datasets, which require a more specialised tool to address.