Love Thy Neighbor: Combining Two Neighboring Low-Resource Languages for Translation

John E. Ortega jortega@cs.nyu.edu
New York University, New York, New York, USA

Richard Alexander Castro Mamani rcastro@hinant.in
Universidad Nacional de San Antonio Abad, Cusco, Perú

Jaime Rafael Montoya Samame jaime.montoya@pucp.edu.pe
Pontificia Universidad Católica del Perú, Lima, Perú

Abstract

Low-resource languages sometimes take on similar morphological and syntactic characteristics due to their geographic nearness and shared history. Two low-resource neighboring languages found in Peru, Quechua and Ashaninka, can be considered, at first glance, two languages that are morphologically similar. In order to translate the two languages, various approaches have been taken. For Quechua, neural machine transfer-learning has been used along with byte-pair encoding. For Ashaninka, the language of the two with fewer resources, a finite-state transducer is used to transform Ashaninka texts and its dialects for machine translation use. We evaluate and compare two approaches by attempting to use newly-formed Ashaninka corpora for neural machine translation. Our experiments show that combining the two neighboring languages, while similar in morphology, word sharing, and geographical location, improves Ashaninka–Spanish translation but degrades Quechua–Spanish translations.

1 Introduction

Low-resource languages (LRL) can be defined as languages that suffer from the presences of insufficient parallel source-target data. Until recently, in order to translate LRLs, rule-based (RBMT) or statistical-based machine translation (SMT) systems have been used with a combination of features and heuristic approaches to create a model that could predict target-side translations based on probability techniques given a source sentence (also known as a segment). With the rebirth of neural machine translation (NMT) in recent years thanks to higher-compute system availability, neural approaches have been used to jointly learn from several source and target segments (Zoph et al., 2016; Gu et al., 2018; Lakew et al., 2018b) avoiding the highly laborious process of creating rules and features to translate using previous RBMT and SMT systems. The majority of research that uses NMT for LRLs tends to show how the combining of two or more source-side languages to one target-side language can help translate low-resource languages by imputing word-level features from a higher-resource language to a lower-resource language.

One such case (Ortega et al., 2021) translates Quechua, a Peruvian LRL, to Spanish, an HRL, using Finnish, another HRL, in an approach called *BPE-Guided* based on glossary def-

initions found from suffixes on Wikipedia[1]. Their work which uses an NMT system based on byte-pair encoding (BPE) (Sennrich et al., 2015b) with a long short term memory (LSTM) (Graves, 2012) was found to outperform other systems measured according to human and system evaluations using BLEU (Papineni et al., 2002).

Research has also been performed by Ortega et al. (2020) on a neighboring Peruvian language called Ashaninka. Ashaninka has less resources than Quechua and is spoken by fewer people. There are nearly 70,000 native Ashaninka speakers (Gordon and Grimes, 2005) as compared to around 5 million native Quechua speakers[2] and both languages can be broken down into different dialects. The amount of resources available for Ashaninka is on the order of 8,000 sentences (or segments) whereas Quechua data is about 40,000 segments and growing. Ortega et al. (2020) dedicated their initial work on Ashaninka to language normalization by creating a finite-state transducer based on previous Quechua work (Rios, 2010). They left for future work the inclusion of Ashaninka in an NMT system.

In order to advance the work by Ortega et al. (2020, 2021), we use resources from their published articles available online[3] to extend their experiments which, in turn, marks the first time, to our knowledge, that an Ashaninka–Spanish machine translation (MT) system is introduced to the MT research community. Our hope is that, since Finnish and Quechua were found to be successful in previous work (Ortega and Pillaipakkamnatt, 2018) due to their highly-similar morphology, the addition of Ashaninka as source-side input should increase performance since Quechua and Ashaninka are from the same region, display similar morphological constructs, and even share loaned vocabulary words where higher-resource languages (Quechua and Spanish) are found in the lower-resource language (Ashaninka).

Our effort is a three-fold, novel, experimental introduction for the two Peruvian languages as seen below:

1. Introduce for the first time a Ashaninka–Spanish MT system.

2. Show how two neighboring South American languages with low resources perform when combined as training data for a NMT system.

3. Perform a micro-analysis on the morphology similarities and difference between Quechua and Ashaninka.

In order to realize the three points, we narrate the following. First, in Section 2, we describe related approaches not mentioned in Section 1. Next, we analyze Quechua and Ashaninka similarities and differences in Section 3. Our methodology and approach are detailed in Section 4 along with the experimental settings in Section 5. We then provide results in Section 6 that show how combining Quechua and Ashaninka together perform on both a Quechua and Ashaninka test set. Lastly, we conclude with an explanation on our findings and potential future research lines in Section 7.

2 Related Work

Ortega et al. (2020) present a system called *AshMorph* which is an approach for normalizing Ashaninkan text for machine translation use. Additionally, the corpus and MT system introduced by Ortega et al. (2021) are used. For more information on how they were used in our work, see Section 5. In this section, we describe other approaches that are similar to ours.

[1] `https://wikipedia.org`

[2] The native-speaker count includes all dialects for both languages

[3] `https://github.com/johneortega/mt_quechua_spanish` and `https://github.com/hinantin/AshMorph`

Pourdamghani and Knight (2017) use a deciphering approach which *relates* a high-resource language to a low-resource language through a character-level ciphering algorithm. Their work assumes that words are ordered similarly. We could not use this approach since, as discussed later in Section 3, word ordering is one of the key morphological differences between Quechua and Ashaninka.

Tantuğ and Adalı (2018) focus on agglutinating languages by using eight informal target-side, rule-based, edits. Their work can be considered similar to the work from Ortega et al. (2020) due to the way it handles morphology and knowledge transfer. However, they use discrete rules meant to work with a statistical disambiguation system for combining the source and target language. Our aim is to show that NMT could be used to learn similar rules without human intervention. Nonetheless, we feel that their work could be included for comparison in future iterations.

Bahdanau et al. (2014) use a neural machine translation system to first learn aligned words that form an encoded vector and then translate them. This work is similar to ours in its approach; however, our work is for an extremely low-resource language (Ashaninka) and depends on character-level differences not performed in their work.

We mirror Zoph et al. (2016)'s approach by using the "OpenNMT-LSTM" system mentioned in Ortega et al. (2021). Zoph et al. (2016)'s results show an increase of 5 BLEU when combining languages; our results are similar when using Quechua as the high-resource language.

Other work (Gu et al., 2018; Karakanta et al., 2018) tend to focus on the addition of several languages with high resources as was done by Ortega et al. (2021) with the inclusion of Finnish, a high-resource language. In this case, we are adding the lower-resource language, Ashaninka, with hopes to better the higher-resource language, Quechua. Additionally, other work (Lakew et al., 2018a) points out that bilingual NMT models may require adjustments when multilingual models perform better. Their work is considered helpful; but, at this early stage of investigation, we lean on the work from Zoph et al. (2016) for guidance.

3 Morphology

Quechua and Ashaninka are morphologically similar at first glance. However, the deeper differences explained here help to understand the results presented in Section 6. In this section, we provide an in-depth analysis of both languages based on previous work (Cerrón-Palomino, 1987; Mihas, 2015). The comparative analysis of the two language's grammatical makeup and morphology, to our knowledge, has not been taken into account by other research, specifically for machine translation.

Like many native North and South American languages, Ashaninka and Quechua are both *polysynthetic* and *agglutinating* (Bustamante et al., 2020), they add prefixes or suffixes to a root morpheme which expand or change a word's meaning. An example follows of the two languages agglutinating similarity.

"the child's hand"	
Quechua	**Ashaninka**
warmapa makin	irako eentsi
warma-**pa** maki-**n**	**ir**-ako eentsi
child-**GEN** hand-**3SG**	**3M**-hand child

At first glance, it is clear that the two languages form words by agglutination. Yet, Quechua and Ashaninka vastly differ when examined closer. This is seen with possessive noun phrases like "the child's hand" above where Quechua adds a suffix (–pa) for genitive (GEN) noun possession and adds a suffix for the possessive person (–n marks the third-person singular (3SG) for

"maki" (hand)). While Quechua double marks possession, Ashaninka only marks the entity being possessed ("ako" (hand) is marked with the third-person masculine possessive (3M) prefix "ir") leaving the possessive person ("eentsi" (child)) unchanged. Additionally, it is worthwhile to note that ordering of words in Quechua is typically of type *possessor–possessed*, while in Ashaninka the order is reversed to *possessed–possessor*.

Verbal conjugation generally inflects and agglutinates in both languages. In Quechua, verbs use suffixes to express the present, past, or future tense. On the other hand, in Ashaninka, most verbs do not take tense inflection into account, instead they use a category called the "reality status" which distinguishes between two types of events: (1) past and present (*real*) events or (2) future (*unreal*) events. (Michael, 2014)

	"to come"	
Quechua	**Ashaninka**	**Conjugation**
hamu-ni	no-**pok**-i	"I come"
hamu-rqa-ni	no-**pok**-i	"I came"
hamu-saq	no-m-**pok**-e	"I will come"

Above, we see how the verb "to come" is conjugated for Quechua and Ashaninka. There is a clear distinction between present (hamuni), past (hamurqani), and future (hamusaq) tenses for the root Quechua morpheme **hamu**. Contrastingly, we see how Ashaninka uses the real/unreal method described, present and past (nopoki) are the same but the future (nompoke) is different for the root Ashaninka morpheme **pok**.

Other linguistic differences also exist with respect to suffixes and their order. More specifically, the phrasal order differs such that Quechua usually takes a subordinate clause proceeded by the verb while Ashaninka is the opposite. Additionally, the three languages (Quechua, Ashaninka, and Spanish) contain words in written texts that can be considered unknown, or *loaned*, words that are inherited from their higher-resource language where Quechua inherits from Spanish and Ashaninka inherits from both Quechua and Spanish. The overlapping words and other differences mentioned are found in the corpora from the work mentioned (Ortega et al., 2020, 2021) which contains normalized texts from corpora created in the past Mihas (2010); Cerrón-Palomino (2008).

4 Methodology

From the description in Section 3, it is clear that, while initially similar, the morphological makeup of Ashaninka is different than Quechua. Our experiments determine if it is better to use Ashaninka or Finnish as a language for transliteration in a NMT system when translating Quechua and Ashaninka to Spanish. The inclusion of Finnish as a source language in both Quechua and Ashaninka translations to Spanish is motivated by Ortega et al. (2021) which showed that neural machine translation was better when including Finnish as a source language during training.

Our experiments are based on previous work (Ortega et al., 2020, 2021) which experiments with Quechua[4], Finnish[4], and Ashaninka[5] as the source languages and Spanish as the target language. We use their translation and normalization approaches to compare the two neighboring language's (Quechua and Ashaninka) translations into Spanish using the NMT system described below.

The best performing system from Ortega et al. (2021)'s work is a NMT system first used

[4]Quechua and Finnish are the source languages in (Ortega et al., 2020).

[5]Ashaninka was not translated into another language in Ortega et al. (2021).

for development (called *OpenNMT-LSTM*) and later in testing (called *OpenNMT*).[6] We compare its performance by using Quechua, Finnish, and Ashaninka to train the NMT system in various combinations (see *Train Languages* in Table 1) with Spanish as the target language.

The results show the performance of the NMT system when using Ashaninka, a neighboring language (about 10 km away), and Finnish, a language that is of high geographic distance (about 8,000 km away), as source languages for translating Quechua to Spanish. Additionally, experiments are performed to show how well Quechua and Finnish perform as source languages when translating Ashaninka to Spanish. The implication is that since Finnish is agglutinative and polysynthetic and it has been shown to improve performance when translating Quechua to Spanish (Ortega et al., 2021), it should help when translating from both Quechua and Ashaninka to Spanish. The next section describes experimental settings for all languages.

5 Experimental Settings

Our experiments mirror previous experiments (Ortega et al., 2020, 2021) in terms of the corpora and NMT system used. Since we combine languages from both works, some of the corpora and languages used as NMT system input is different. In this section, we present those input changes and reiterate the similarities to previous work.

First, for Ashaninka text to be used as input into the NMT system, we transform it using the *AshMorph* (Ortega et al., 2020) normalization technique. For purposes of Ashaninka inclusion in the experiments, there are 521 Ashaninka training sentences (or segments), 111 development segments, and 111 test segments. They are used in three different *training* experiments: (1) Quechua+Finnish+Ashaninka, (2) Quechua+Ashaninka and (3) Ashaninka only; and, in one development and test direction (Ashaninka–Spanish). All of the corpora is randomly selected from the developoment corpora (Cushimariano Romano and Sebastián Q., 2008) used previously (Ortega et al., 2020).

Second, for Quechua normalization as input to the NMT system, a morphological normalizer (Rios and Castro-Mamani, 2014) from previous work (Ortega et al., 2021) is used. Quechua is used as a training language in all of our training experiments except for when Ashaninka is tested in isolation. The Quechua corpora consists of 17,500 training segments, 2,500 development segments, and 5585 test segments all randomly chosen from Ortega et al. (2021)'s experiments originated from the Opus corpus[7] (Tiedemann, 2012) and used in three different training settings: (1) Quechua+Finnish, (2) Quechua+Finnish+Ashaninka, and (3) Quechua+Ashaninka; and, in one development and test direction (Quechua–Spanish).

Third, Finnish and Spanish, both considered high-resource languages, are more plentiful. Like the work from Ortega et al. (2021), we use the JW300 corpus (Agić and Vulić, 2019). Since Spanish is the target language in all cases, Finnish is the only high-resource language included for training. We use 149,251 Finnish segments for training in two systems: (1) Quechua+Finnish and (2) Quechua+Finnish+Ashaninka. Spanish is used only for parallel development for testing with Quechua–Spanish and Ashaninka–Spanish language pairs.

All segments for all languages were tokenized and true-cased using Moses (Koehn et al., 2007) after normalization.

To summarize our validation technique for the neural MT system experiments, we use two source–target pairs: Quechua–Spanish and Ashaninka–Spanish. For example, for the *qu+fi+cni* system in Table 1 is used for translating Quechua to Spanish (*qu–es*). Its validation (or dev) data consists of 2500 parallel *qu–es* segments and test data is of 5585 *qu–es* segments. The Ashaninka to Spanish (*cni–es*) experiments consist of a dev and test set of 111 parallel *cni–es* segments.

[6] Details about the hyper parameters for both systems are found in Section 5.

[7] `http://opus.nlpl.eu/`

The NMT system used for all experiments is the system described in Ortega et al. (2021)'s dev phase called *OpenNMT-LSTM*. The system is trained for 100,000 epochs and it is a 2-layer LSTM model (Hochreiter and Schmidhuber, 1997) with 500 hidden units, dropout of 0.3, and uses stochastic gradient descent as the learning optimizer along with a batch size of 64. To evaluate the NMT system, we use BLEU (Papineni et al., 2002) like the work from Ortega et al. (2020, 2021).

The next section explains how previous work (Ortega et al., 2020, 2021) was used to test the neighboring Quechua and Ashaninka languages with the NMT system proposed.

6 Results

The experiments in Table 1 show the results of combining Quechua, Finnish, and Ashaninka. There are three main training scenarios along with one Ashaninka experiment in isolation. For each training scenario, there are two experiments performed, one with Quechua to Spanish (qu–es) and one with Ashaninka to Spanish (cni–es).

Our results are aligned with what has been discussed in Section 3 section at a high level – Ashaninka and Quechua appear similar in linguistic nature at first glance; however, at a deeper evaluation, the lack of resources and complex grammatical differences decrease *qu–es* translation performance. On the other hand, similar to work from Zoph et al. (2016), we have shown that by adding Quechua resources to Ashaninka, there is a gain of 4.6 BLEU. In all other cases where Ashaninka was combined with Quechua or Finnish, the performance degraded for *qu–es* translations and only very slightly (.2 BLEU) increased in one *cni–es* case.[8] Another interesting takeaway is that Finnish remains the better language to combine with Quechua when translating Quechua to Spanish. This is due to the large amount of Finnish training examples (149,251) compared to the small amount of Ashaninka training examples (521). In actuality, the BLEU score of the *qu+cni* trained system is the same as the BLEU score of using *qu+es* alone in training reported by Ortega et al. (2021). This leads us to believe that if there were more Ashaninka training examples the potential to outperform Finnish as the transfer-learning language is high.

Train Languages	Direction	Train Count	Dev Count	Test Count	BLEU
qu+fi	*qu–es*	166751	2500	5585	22.6
qu+fi	*cni–es*	166751	111	111	0.0
qu+fi+cni	*qu–es*	167272	2500	5585	17.0
qu+fi+cni	*cni–es*	167272	111	111	0.2
qu+cni	*qu–es*	18021	2500	5585	20.1
qu+cni	*cni–es*	18021	111	111	5.9
cni	*cni–es*	521	111	111	1.3

Table 1: Translating to Spanish (es) with Quechua (qu), Finnish (fi), and Ashaninka (cni) using a neural machine translation system.

7 Conclusion and Future Work

We have shown that while previous work combining languages may seem viable for low-resource languages, in some cases, while languages seem similar at first glance, results may differ. This is clear from our experiments with Quechua and Ashaninka that show performance loss when adding them together for transfer-based learning in an NMT system. Nonetheless, it would be advantageous to try other techniques such as back-translation (Poncelas et al., 2018;

[8]The higher resource pairs consist of 166,751 pairs of parallel data together of which the Finnish data is 149,251 parallel segments in total.

Karakanta et al., 2018; Sennrich et al., 2015a) to create more synthetic Ashaninka data since, at this point, Finnish provides more gain when combined with Quechua than Ashaninka does.

Future lines of investigation will include a supervised version of the **AshMorph** (Ortega et al., 2020) algorithm with the intent to automate sub-segment level selection. The plan is to improve Ashaninka to Spanish translations by first creating more human-evaluated training data and, second, experimenting with several other resources to create more synthetic data. Experimentation should also explore other similar languages since Quechua seems to help (not hurt) Ashaninka to Spanish translations.

References

Agić, Ž. and Vulić, I. (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.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bustamante, G., Oncevay, A., and Zariquiey, R. (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.

Cerrón-Palomino, R. (1987). *Lingüística quechua*. Centro de Estudios Rurales Andinos" Bartolomé de Las Casas".

Cerrón-Palomino, R. (2008). *Quechumara: Estructuras paralelas del quechua y del aimara*. Plural editores.

Cushimariano Romano, R. and Sebastián Q., R. C. (2008). Ñaantsipeta asháninkaki birakochaki. diccionario asháninka-castellano. versión preliminar. `http://www.lengamer.org/publicaciones/diccionarios/`. Visitado: 01/03/2013.

Gordon, R. G. and Grimes, B. F. (2005). Ethnologue : languages of the world.

Graves, A. (2012). Long short-term memory. In *Supervised sequence labelling with recurrent neural networks*, pages 37–45. Springer.

Gu, J., Hassan, H., Devlin, J., and Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Karakanta, A., Dehdari, J., and van Genabith, J. (2018). Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1-2):167–189.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W.and Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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, Prague, Czech Republic.

Lakew, S. M., Federico, M., Negri, M., and Turchi, M. (2018a). Multilingual neural machine translation for low-resource languages. *IJCoL. Italian Journal of Computational Linguistics*, 4(4-1):11–25.

Lakew, S. M., Lotito, Q. F., Negri, M., Turchi, M., and Federico, M. (2018b). Improving zero-shot translation of low-resource languages. *arXiv preprint arXiv:1811.01389*.

Michael, L. (2014). The nanti reality status system: Implications for the typological validity of the realis/irrealis contrast. *Linguistic Typology*, 18(2):251–288.

Mihas, E. (2010). *Essentials of Ashéninka Perené Grammar*. PhD thesis, The University of Wisconsin.

Mihas, E. (2015). *A grammar of Alto Perené (Arawak)*. De Gruyter Mouton.

Ortega, J., Castro-Mamani, R. A., and Montoya Samame, J. R. (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.

Ortega, J. and Pillaipakkamnatt, K. (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.

Ortega, J. E., Mamani, R. C., and Cho, K. (2021). Neural machine translation with a polysynthetic low resource language. *Machine Translation*, pages 1–22.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Poncelas, A., Shterionov, D., Way, A., Wenniger, G. M. d. B., and Passban, P. (2018). Investigating backtranslation in neural machine translation. *arXiv preprint arXiv:1804.06189*.

Pourdamghani, N. and Knight, K. (2017). Deciphering related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2513–2518.

Rios, A. (2010). Applying finite-state techniques to a native american language: Quechua. *Institut für Computer Linguistik, Universitaät Zürich*.

Rios, A. and Castro-Mamani, R. (2014). Morphological disambiguation and text normalization for southern quechua varieties.

Sennrich, R., Haddow, B., and Birch, A. (2015a). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., and Birch, A. (2015b). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Tantuğ, A. C. and Adalı, E. (2018). Machine translation between turkic languages. In *Turkish Natural Language Processing*, pages 237–254. Springer.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.