Overcoming Resistance: The Normalization of an Amazonian Tribal Language

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

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

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

Abstract

Languages can be considered endangered for many reasons. One of the principal reasons for endangerment is the disappearance of its speakers. Another, more identifiable reason, is the lack of written resources. We present an automated sub-segmentation system called **AshMorph** that deals with the morphology of an Amazonian tribal language called Ashaninka which is at risk of being endangered due to the lack of availability (or resistance) of native speakers and the absence of written resources. We show that by the use of a cross-lingual lexicon and finite state transducers we can increase accuracy by more than 30% when compared to other modern sub-segmentation tools. Our results, made freely available on-line, are verified by an Ashaninka speaker and perform well in two distinct domains, everyday literary articles and the bible. This research serves as a first step in helping to preserve Ashaninka by offering a sub-segmentation process that can be used to normalize any Ashaninka text which will serve as input to a machine translation system for translation into other high-resource languages spoken by higher populated locations like Spanish and Portuguese in the case of Peru and Brazil where Ashaninka is mostly spoken.

1 Introduction

In South America, there are hundreds, if not thousands, of low-resource languages. In Brazil alone, one can quickly find a list[1] on-line of languages that range from *vulnerable* to *critically* endangered. Some languages, like Quechua, a low-resource languages mostly spoken in Peru and Bolivia, have gained more attention in recent work (Cardenas et al., 2018; Cotterell et al., 2018; Ortega and Pillaipakkamnatt, 2018; Hintz and Hintz, 2017) and, while not 100% translatable to their high-resource counterparts, are on the path to digital preservation. Other South American languages are closer to extinction. One such language, called Ashaninka, is spoken by nearly 70,000 people in the Amazon forests shared by Peru and Brazil (see Figure 1). While there are plenty of Ashaninka speakers when compared to other tribal languages like Guarasu which is spoken by only hundreds, native Ashaninka speakers are generally less willing to cooperate in the digitization of the language (Varese, 2004).

We have located a near-native Ashaninka person to help validate our work which consists of multiple findings and improvements. Our research ultimately presents a normalization tool that will completely automate the processing of Ashaninka text as input for a machine translation (MT) system. More specifically, we complete three main tasks: (1) alphabet normalization, (2) morphological disambiguation, and (3) a tagging system which includes part-of-speech (POS) and morphological tagging. In addition to creating an automation tool[2], called **AshMorph**, we make a human-verified sub-segmentation development and test corpus freely available.

We present our findings using **AshMorph** on a development set in one domain (articles from story books, plays, and educational text) and; then, on a test set in another domain, common biblical translations — often times the only resource low-resource languages in South America. We detail our results by comparing them to the use

[1] https://en.wikipedia.org/wiki/List_of_endangered_languages_in_Brazil

[2] https://github.com/hinantin/AshMorph

Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, pages 1–13

of the latest sub-segmentation techniques, namely Subword-NMT (Sennrich et al., 2015) (or byte-pair encoding–BPE), Morfessor (Smit et al., 2014) and SentencePiece (Kudo and Richardson, 2018) (BPE with a unigram model).

Figure 1: Ashaninka speakers covering the Amazon region between Peru and Brazil.

The structure of this paper has been created to first provide an overview of our in-depth normalization process and then provide convincing results that serve as reasoning to use **AshMorph** for future tasks. In Section 2, we compare and contrast related work on Ashaninka's nearest neighboring language, Quechua, and provide citations for other work that led up to this paper. Then, in Section 3, we dive deep into the details of **AshMorph** to show how it takes advantage of another system currently used to normalize Quechua by extending it to cover morphology and grammatical context from Ashaninka. After that, we cover our experimental settings for two stages (development and test) of **AshMorph's** development. Next, we provide detailed results along with sample output from **AshMorph** in Section 5. Lastly, in Section 6, we give insight into what we plan as our next set of experiments towards a final product for translation.

2 Related Work

Our work is novel due to its overall linguistic coverage of Ashaninka in several ways, serving specifically as a first step for translation by normalization and sub-segmentation. However, there are other works that perform sub-segmentation on low-resource languages, including Ashaninka's Peruvian counterpart, Quechua. Since Quechua and Ashaninka are both polysynthetic languages and, in some ways, have a similar morphological makeup, we present those works along with the biblical corpus translations that are used in our experiments.

Since the bible has been translated in several languages (Christodouloupoulos and Steedman, 2015) and is available on-line in a parallel format by Opus[3], we include it here as related work because it has been used in several machine translation tasks on other low-resource languages, including Quechua. Contrastingly, while the bible corpus for Quechua (acronym QUW on Opus) has 12,400 total sentences, it has 19,100 sentences for Ashaninka (acryonym CNI on Opus). Nonetheless, Quechua by far has received more attention in experimentation related to MT.

The work most similar to ours performs normalization on Quechua (Rios, 2010). It consists of the development of a morphological analyzer and normalization technique similar to ours which allows it to analyze several Quechua dialects. Using her work as a guide and the help of a near-native Ashaninka reviewer, we present a normalization pipeline for Ashaninka which, mainly due to resistance and lack of resources, is still in the early stages of a normalized writing system. Unlike Quechua, which possesses a normalized lexicon for all its southern dialect varieties (Cerrón-Palomino, 1994), Ashaninka only possesses a few candidates for a normalized alphabet, among them is the one developed by Elena Mihas (Mihas, 2010) which, based on our observations, is best suited for use as a normalized alphabet Ashaninka and its dialects.

While there are several works that perform morphological segmentation and normalization, we do not consider them especially related to our work. However, one MT system, Apertium (Forcada et al., 2011), has been used in a similar way to produce rule-based translations based on morphology. As a comparison, Apertium was used for translating Quechua to Spanish in the past (Larico Uchamaco et al., 2013). It has a unique way of rule creation in its "lttoolbox" that allows for the creation of a finite state transducer (FST), similar to our work described in Section 3, that then serves as input to its MT component. But, we additionally introduce other tagging strategies, such as an alphabetical mapping device, specific to Ashaninka, that are based on its syntactic structure

[3] http://opus.nlpl.eu/bible-uedin.php

as was done in another system used to translate Spanish to Quechua based on the same normalization technique used in **AshMorph** originally created for Quechua (Rios, 2010).

As for the other systems used for comparison in our results (Subword-NMT (Sennrich et al., 2015), Morfessor (Smit et al., 2014) and SentencePiece (Kudo and Richardson, 2018)), we do not consider them directly related to our work because they are generic segmentation system that lack normalization techniques based on linguistic knowledge. They are included here as means of comparison due to the lack of sub-segmentation systems available for Ashaninka.

Other work, while similar in nature, does not dive into the complexities of Ashaninka. In the next section, we describe our implementation of **AshMorph** in detail with an attempt to show why our work is novel and unlike research in other languages, except for Quechua, which this work is based upon.

3 Methodology

Our system has been developed so that MT systems can use its normalized output as input. In this section, we describe the normalization pipeline that consists of several steps. We first provide linguistic details about Ashaninka which include a description of the alphabet, details on the morphology, and parts of speech (POS). Then, we cover the two tag sets that cover the POS and morphology.

3.1 Alphabetical Normalization

Ashaninka has several dialects and is known to be both polysynthetic and agglutinative (Bustamante et al., 2020). The language is highly inflective and contains several suffixes that are added on, much like Quechua, to the end of a root word. The ambiguous nature of its grammatical construct has most likely led to the high amount of dialects known as *pan-Ashaninka*. With that in mind, we decided to use one alphabet, called *Asheninka Perene* (Mihas, 2010), as the main alphabet for mapping pan-Ashaninka to one alphabet for creating lexicons and other grammatical constructs in **AshMorph**. To our knowledge, Asheninka Perene is the most recent alphabet available for Ashaninka and it extends the original alphabet created by Payne (1981) as illustrated in Table 4.

3.2 Morphological normalization

Ashaninka, much like its Peruvian counterpart (Quechua), is a polysynthetic language whose inflection, which changes the meaning of a word, depends on the head of a phrase. This makes sub-segmentation of Ashaninka different than Quechua despite them being both agglutinating languages. Ashaninka typically inflects on a transparent noun and verb agreement and, more often than not, one can find a word that combines several stems (noun and verb classifiers) that make up a specific semantic meaning. For example, we denote the incorporation of a noun classifier (*tsapya*) and a verbal classifier (*ha*) below:

3.2.1 apaani asheninka isaika**tsapya**atziro inkaare

apaani *asheninka*
one man
*i-saik-a-**tsapya**-atz-i-ro*
3m.A-to.live-EP-**river.bank**-PROG-IRR-3n.m.O
inkaare
lake
"a man who lived near a lake"

3.2.2 katsinkahari

katsinka *-**ha*** *-ri*
to.be.cold -cl:**liquid** -rel
"cold water"

Examples 3.2.1 and 3.2.2 show how inflection works in Ashaninka. We note that the word heads, or roots, are marked when verbs and nouns agree with properties, such as the gender, of their arguments. In Ashaninka, verbs are often times marked by gender and that property is transferable; this is seen where both the subject and object of a phrase are cross-referenced. The cross-referencing is illustrated clearly in Example 3.2.1 where its subject (*apaani asheninka* – one man) is masculine and its corresponding verb's prefix (the first letter *i* in i-saik-a-tsapya-atz-i-ro) is also masculine. This cross-referencing is furthermore replicated in the direct object of the sentence (*inkaare* – lake). That means that not only do we see gender agreement in the corresponding verb; but, we also see the gender-based dependency between the object of the sentence (*inkaare* – lake).

One of the interesting phenomena of Ashaninka is that it contains a suffix (–paye) that can inflect the meaning of a word to be plural and nominal. Nonetheless, Ashaninka can still be highly inflective and ambiguous when dealing with plural or singular nouns. One example where there is a nominal root that could refer to more than one entity occurs in a word like *koya – woman*:

- ≥ 1 one or more entities, e.g.: *koya* 'woman or women'.
- $= 1$ or one entity, e.g.: *aparoni koya* 'one woman'.
- > 1 more than one entity, e.g.: *koyapaye* 'women'.

In our morphological analysis, we collected several morphological rules and assigned them operations. The operations generally depend on boundary symbols (marked as $\sim$ in our rule set). We make all rules publicly available in our FST. Here's an example of a morphological rule used to differentiate verbal roots where a personal pronoun prefix is both partially and totally duplicated:

{ (*n* '1SG.S/A' + *oirink* 'to.lower' + a 'EP' –> *noi∼noirinka* "I get lower and lower") } (partial duplication of "lower")

{ (*n* ' 1SG.S/A' + *ak* 'to.answer.back' + a 'EP' –> *naka∼naka* 'I answer back constantly') } (total duplication of "answer back").

3.3 Grammatical Tagging

The normalization of most languages where there are not enough resources to learn linguistic features using an automated method, such as deep learning, typically requires a tagging approach. One of the most common tagging approaches is POS tagging; however, for languages like Quechua and Ashaninka, initial morphological tagging is needed before POS tagging can be performed. The implementation of a dual-faceted approach to cover both POS and morphology is what makes **AshMorph** unique.

While earlier tagging efforts (Payne, 1981) were originally based on three types of morphology (nouns, verbs, and adverbs), our work takes advantage of two more main classes (adjectives and pronouns). The original argument (Payne, 1981) for having only three classes was that adjectives and pronouns were indistinguishable; however, we have found the contrary by taking into account more seminal work (Mihas, 2010) that points towards the advantage of having major word classes (nouns and verbs) along with several smaller classes of adjective, adverbs, pronouns, and more. Our work uses their work (Mihas, 2010) to assist in the creation of a normalization technique that covers both POS and low-level morphological anomalies. Many of the final rules (described in further detail below) were gotten during the development stage in several trial-and-error experiments to expand initial lexicons.

Part of Speech Tagging One of the more interesting approaches for our tag set development is the use of rules developed using the Czech language, specifically those mentioned in previous work (Hlaváčová, 2017). We adapt those rules, as done by others who worked on similar Peruvian languages (Pereira-Noriega et al., 2017; Cardenas et al., 2018), for matching Ashaninka in a quick fashion as described in the following. First, our tag set, as (Hlaváčová, 2017) does, consists of POS tags, i.e. labels used to indicate the POS and other grammatical categories such as case or tense for each *token* – a token is defined as a sequence of non-blank characters between blanks, handling punctuation as separate tokens. Second, we group multi-token units, such as proper names or numbers, at the structural level as done in previously (Brants et al., 2003, pag. 77). Table 1 provide a more in-depth illustration the Ashaninka tag set that we developed for **AshMorph**.

Deep Morphological Tagging Our morphological tagger has been created to perform a deep analysis on Ashaninka, deeper than POS tagging alone. It takes several lexical classes into account that have nothing to do with POS. We use *strict* tagging for word forms which takes into account inflection, tense, gender, and more. Additionally, we convert our tagging output into a comprehensive, human-readable, tag set similar to another morphological analyzer created for Quechua (Rios, 2010).

The words forms that we analyze in Ashaninka are formed, much like in English, by the concatenation of letters. For example, the English words "write", "writes", "sisters", and "where" in Ashaninka are *sankenataantsi*, *isankenate*, *choenipaeni*, and *tsika*, repectively. The heads, or nuclei, of Ashaninka words are created in **AshMorph** by using their corresponding lemmas. For

Category	Type	Abbrev.
Verb (V)	A-class I-class Copula	A I COP
Noun (N)		
Adjective (Adj)	Adjective Underived Adjective Derived Adjective	
Adverb (Adv)	Time Place Locative Others	
Pronoun (Prn)	Personal Possessive Demonstrative Interrogative Indefinite Forms	Pers Poss Dem Wh Indef
Numeral (NUM)		
Particle (C)	Connective Interjection Discourse	Conn Interj
Ideophone (IDEO)		
Unknown (UNK)		
Punctuation	$., $", $', $-, $?, $!	

Table 1: **AshMorph's** POS Tag Set – originally based on structural information from Czech language (Hlaváčová, 2017).

the preceding example, the heads are: *sankena*, *sankena*, *choeni*, and *tsika*. By using the lemma, we are able to expand the understanding of inflection into what we call a **paradigm** – a set of word-forms that can be created by means of "inflection" from their lemma. We have found that lemmas in Ashaninka are often times different according to the dialect; so, in future iterations, some input may differ for words where lemmas are distinct much like would occur in English with the words "color" and "colour". Nonetheless, in our corpora (described further in Section 4), there were no major variants found despite the corpora contents which consisted of several different dialects.

Our morphological tagging system is based on two main concepts: category and value. Morphological category in our system refers to the properties of a word such as gender (masculine–+m, or non-masculine–n.m.). The tense, on the other hand, is based on what's known as the reality status systems (RSS), a binary verbal distinction between "realized" and "unrealized" situations established in previous work (Michael, 2014). We deem morphological value as the actual value of a morphological category. For example, the morphological **values** for the morphological **category** "number" could be singular (SG) or plural (PL). In our system, the final morphological tag that **AshMorph** outputs is a sequence of (X, Y) pairs annotated with the format $+X@Y$, denoting that Y is the morphological value of the morphological category X. The example below shows how **AshMorph** tags the verbal root *tash* ("to be hungry" in English) followed by the suffixes *agantsi* and *antsi* which denote the infinitive tense. When lemmas, or heads, take on different morphological values for the same morphological category $(X, Y_1), (X, Y_2), ..., (X, Y_n)$, we annotate them as $+X@Y_1; Y_2...; Y_n.$. The lemma *tash-* 'to.be.hungry' may take either of the infinitive 'INF' suffixes: *-agantsi* or *-antsi* resulting in the annotation `[+INF@agantsi;antsi.]` as seen in Section 7.1.1 (Supplemental Material).

Since South American languages with complex morphology like Quechua and Ashaninka contain a lot of information inside of their suffixes, we use a *maximal* set of morphological values to cover what can be considered a complex RSS (Mihas, 2010) which provides a lot of contextual information by using what are known as *realis*, or 'REAL', and *irrealis*, or 'IRR', suffixes. **AshMorph's** maximal set of morphological values can be defined as the set that is sufficient for morphological description of a single word form given its lemma that typically depends on the POS classification for the lemma. For the **REAL** suffixes, the annotation model is $+X@Y_1Y_2Y_3.$; and, for the **IRR** suffixes, the annotation model is $+X@Y_1Y_2Y_3Y_4Y_5.$. In order to group the morphological values in said format, we use what is known as the EAGLES[4] format which, amongst other formats, is used for those cases that need to take position of a sequence of values into account.

EAGLE output can be difficult to read; so, we prepared a method to modify the output in such a way that a beginner, not trained on Ashaninka, can use them. Our method follows previous work (Hulden and Francom, 2012; Rios, 2010, pag. 2115) on lemmas and suffixes; and, is nearly identical to the Quechua normalizer from their work that produces human-readable labels based on a lexicon as seen in Sections 7.1.1 - 7.1.4 (Supplemental Material).

[4] http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html

3.4 Combined Normalization

The output of **AshMorph** is a combination of alphabetical and morphological normalization that produces a readable tag set that explains the POS, morphology, and grammatical construct for each word. It allows us to separate each word from the Ashaninka input into parsed informational chunks. The process explained in this section can be considered somewhat more complex than others; but, it is advantageous at this stage because it provides deep insight into how Ashaninka is linguistically structured. The finalized output is straightforward to read from other software packages. We provide as an example in Figure 2 ("OUTPUT WITH TAGS:").

Figure 2 shows how a sample input (labelled "PRIMARY INPUT") is first parsed using alphabetical and morphological normalization. Since the three words (*amõyasatzi*, *amoyasatzi*, *amoñasatzi*) are actually the same word but from different dialects, we normalize them to one word (*amonyasatzi*). After that, an FST is used to transform the normalized input into a statistical model using the method described in Section 3.2. The model is then applied to the word using a method similar to previous work (Rios, 2010) which produces output that contains POS tags along with other morphological tags such as gender.

In the next section, we cover the data used for creating our models and validation based on the methodology explained.

4 Experimental Settings

Our experiments consist of two stages, a development stage and a test stage. While it is not customary to include the development stage in experiments, we add it here due to the corpus domain difference. During the development stage, extra attention was given to the different dialects and suffixes to construct a generic system that will work with any type of Ashaninka input.

The **AshMorph** system is made publicly available[5] as are the development and test corpora. As mentioned in Section 3 , it consists of a morphological analyzer and FST model similar to those (Rios, 2010) created for another language, Quechua, most commonly spoken in Peru. Our contribution lies in the modification of the system from Quechua to Ashaninka; the applying of an alphabetical, grammatical, and POS normalization, along with the development and test corpora, validated by a reviewer.

The development stage corpus consists of a collection of various stories, plays, and education text extracted from '*Ñaantsipeta asháninkaki birakochaki*' (Cushimariano Romano and Sebastián Q., 2008) made publicly availabe[6] with permission from the authors. In total, 745 sentences were extracted by analyzing each sentence and converting them to their inflected forms which resulted in a vocabulary of 2389 words. For creating language models during development, we used 50 randomly selected sentences for evaluation by a near-native Ashaninka speaker. Evaluations were done by first providing the evaluator with the output which contained several sub-segments marked by our system (**AshMorph**). Rules were modified based on the evaluation in order to fine tune **AshMorph** during development and helped the system to gain the optimum results during the test stage.

In order to show that **AshMorph** works well with texts not found in the development stage, we chose a corpus from a reliable well-known, online, resource, Opus[7]. We randomly chose 50 of the 7774 total bible-based sentences and executed **AshMorph** on them using the model based on rules gotten from the development stage. Again, a near-native reviewer was asked to validate the **AshMorph** sub-segmentation and accuracy along with a confusion matrix were used to show how well both systems performed.

Three main systems were used to compare off-the-shelf sub-segmentation systems with **AshMorph**. We decided to use systems that have been previously used in MT tasks because our final goal, left for future work, is to use this work for translating local Ashaninka texts to Spanish using MT systems. The first system used for comparison was Subword-NMT[8] (Version 0.3.7). We used the default byte-pair encoding mechanism in Subword-NMT, in our experiments, the vocabulary was 2389 words in size. The second system used for comparison was Morfessor[9] (Version 2.0.6). Morfessor is a classifier-based system that uses statistical-based rules based on a N-

[5] https://github.com/hinantin/AshMorph

[6] https://github.com/hinantin/AshMorph/dev_corpus

[7] http://opus.nlpl.eu/bible-uedin.php

[8] https://github.com/rsennrich/subword-nmt

[9] https://github.com/aalto-speech/morfessor

best Viterbi algorithm. It has also been used in other MT projects (Liu et al., 2017; Zuters and Strazds, 2019) for translation on languages similar to Quechua or Ashaninka. The third system, SentencePiece (Kudo and Richardson, 2018), is commonly used in high-grade MT systems for tasks related to neural machine translation and rapid evaluation(Bérard et al., 2019; Neubig and Hu, 2018). With SentencePiece, we used the default byte-pair encoding with unigram model for segmentation. For all three systems, we used a corpus split similar to the development stage above used in **AshMorph** consisting of 696 train/tune sentences and 50 test sentences.

In the next section, we provide details on how well our system performed by providing both accuracy and a detailed precision score.

5 Results and Conclusion

AshMorph performs well considering the amount of resources available. When compared to other, current, sub-segmentation systems based on statistical models of some sort, it outperforms them on the order of 37 to 59% during the development stage and 26 to 45% during the test stage. The change of corpus from localized stories, plays and educational texts during development to a biblical domain does not seem to highly affect it (a difference on average of 13% between systems). We believe that the decrease in performance is mostly due to new words introduced by the biblical corpus that were hard to discern even by the near-native reviewer.

The normalization rules introduced during the development stage were created from various consultations with linguists, trained in Quechua, Ashaninka, and other South American native languages. In order to develop **AshMorph**, a cluster of dialects with a wide range of writing systems, initially without a normalized writing system, were gathered to created and served as a pivot to provide higher vocabulary coverage and reduce human effort. The normalization system presented here provided evidence that, for a low-resource language that consists of several dialects and rarities in its infancy, a system based on linguistic consensus can outperform other, more modern, statistical-based modeling systems.

Finalized results from **AshMorph** result in a robust normalization and sub-segmentation system that display a clear path from input to output, clarity should be provided when developing a system of this nature. By providing clear rule-based boundaries (Input–>Morpheme–>Suffix–>Morphological Tag–>Essential Translation) as shown in Table 2, one can trace and modify rules accordingly to build a more powerful system, that, later, combined with other MT techniques can achieve better translations.

In order to show that a sub-segmentation system developed with linguistic knowledge can be powerful for Ashaninka, we provide a detailed review of our findings during both the development and test stages in Table 3. We measured accuracy, true positives (TP), false positives (FP) and false negatives (FN). For our evaluation, accuracy is considered to be only those sub-segments were there was complete agreement between **AshMorph** and the Ashaninka reviewer, similar to a TP in the following sentence. A TP is considered to be the number of sub-segments where **AshMorph** agreed with the Ashaninka reviewer. An FP is considered to be the number of sub-segments marked by **AshMorph** but not marked by the Ashaninka reviewer. Lastly, an FN is considered to be the number of sub-segments that the **AshMorph** did not mark but the human annotator did.

From Table 2, we clearly establish that for Ashaninka, much like previous work (Rios, 2010) on Quechua, a completely linguistic-based approach performs better than other baseline sub-segmentation systems. The addition of morphological and grammatical rules help **AshMorph** outperform the other systems tested by nearly 30% in most cases. The overall result can be considered a first step for MT or other tasks that may want to include the Ahaninka language. Here, we overcome the initial drawbacks of resistance from native speakers using previous work on a similar language that is spoken in a nearby region, Quechua.

We believe that the publicly-available rules and sub-segmentation system presented here along with the reviewed corpora and texts should be considered a principle step for low-resource languages, specifically Ashaninka. The results have shown that a more robust system with more resources could take advantage of **AshMorph's** ability to separate words into their morphological and grammatical construct by extending the system to match its needs.

The idea of using an initial Quechua-based normalization system for Ashaninka was based on

Input:	amõyasatzi			
	Morpheme	Suffix Class	Morphology Tag	Essential Translation
Output:	[=amonya/amõya]		[NRoot:CPB]	[=Amonia.river]
	[-satsi+m.]	[NS:CPB]	[+CL:provenance]	[=human.provenance, inhabitant.of.swh.]
	1	2	3	4
English:	'from the Amonia river'			

Table 2: Sample input and output from AshMorph that shows how tracing the system's decision path can be helpful when developing finite-state transducer rules for input to a machine translation system.

	dev				**test**			
	subword-nmt	**morfessor**	**sentencepiece**	**AshMorph**	**subword-nmt**	**morfessor**	**sentencepiece**	**AshMorph**
ACC	16.3 %	38.94 %	38.08 %	74.99 %	29.99 %	47.21 %	48.4 %	74.42 %
TP	82	131	135	305	668	923	907	1489
FP	152	172	131	179	1382	1270	1163	344
FN	321	265	261	91	1376	1121	1137	555

Table 3: A side-by-side comparison of accuracy(ACC), true positives (TP), false positives (FP) and false negatives(FN) for **AshMorph** and other modern sub-segmentation tools.

several similarities in the two languages, mainly agglutination and polysemy. Apart from that, the possibility of finding Ashaninka reviewers that had some background with Quechua morphology was higher. That led us to believe that by creating a system for segmenation based on morphological analysis and linguistic knowledge, we could achieve high performance, better than those used in most MT tasks. **AshMorph** is a first step in low-resource translation and could be used for other languages.

6 Future Work

We believe that the "sky is the limit" for Ashaninka. We have somewhat overcome initial resistance by native speakers to help establish the first step of translating Ashaninka, sub-segmentation and language tagging. With improvements of up to 30% in accuracy as a baseline for our initial experiments, future work could include more linguists and/or native Ashaninka speakers as the margin for improvement stands around 20% (our current highest accuracy score in the test stage is 74.42%). There is one experiment in particular for which resources (time and economic) fell short: the training of an unsupervised segmentation technique over the entire set of sentences for which we do not have annotations from Opus. The experiment would eliminate the need to constrain the amount of text available for learning segmentation. We also believe that by analyzing various combinations of hyper-parameters to statistical systems, we may achieve better results than published. In our opinion, MT systems can now easily use the output from **AshMorph** to produce an initial translation set; or, at a minimum, reproduce results from other languages with extremely low resources as has been done recently (Karakanta et al., 2018). We plan on first using **AshMorph** to normalize Ashankinka text as input to a rule-based MT system like Apertium (Forcada et al., 2011) where **AshMorph** FST rules could be compared and evaluated for performance. The hope is to create translations that perform as well as its neighboring language, Quechua, as presented in previous work (Rios, 2015).

Acknowledgements

We would like to thank Liliana Fernández-Fabián (Univ. Nacional Mayor de San Marcos) for linguistic assistance. Also, Rubén Cushimariano-Romano and Richer Sebastian-Quinticuarithe – authors of the "Asháninka – Spanish Dictionary". Lastly, a special thanks to Elena Mihas for her research on Perene Asheninka which helped us develop our morphological analyzer.

References

Alexandre Bérard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs europe's systems for the wmt19 machine translation robustness task. *arXiv preprint arXiv:1907.06488*.

Thorsten Brants, Wojciech Skut, and Hans Uszkoreit. 2003. Syntactic annotation of a german newspaper corpus. In *Treebanks*, pages 73–87. Springer.

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.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*.

Rodolfo Cerrón-Palomino. 1994. Quechua sureño. diccionario unificado. *Biblioteca Básica Perúana, Biblioteca Nacionál del Peru*.

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

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, S. J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The conll-sigmorphon 2018 shared task: Universal morphological reinflection. *CoRR*, abs/1810.07125.

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

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.

Daniel J Hintz and Diane M Hintz. 2017. The evidential category of mutual knowledge in quechua. *Lingua*, 186:88–109.

Jaroslava Hlaváčová. 2017. Golden rule of morphology and variants of word forms. *Journal of Linguistics/Jazykovedný casopis*, 68(2):136–144.

Mans Hulden and Jerid Francom. 2012. Boosting statistical tagger accuracy with simple rule-based grammars. In *LREC*, pages 2114–2117.

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

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Guido-Raúl Larico Uchamaco, Hugo David Calderón Vilca, and Flor Cagniy Cárdenas Mariño. 2013. Incubation system machine translation spanish to quechua, based on free and open source platform apertium. *CEPROSIMAD*. Online version: http://ceprosimad.com/revista/ingenieria.pdf.

Chao-Hong Liu, Qun Liu, and Glasnevin Dublin. 2017. Introduction to the shared tasks on cross-lingual word segmentation and morpheme segmentation. *Proceedings of MLP*, pages 71–74.

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

Elena Mihas. 2010. *Essentials of Ashéninka Perené Grammar*. Ph.D. thesis, The University of Wisconsin.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

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.

David Lawrence Payne. 1981. *The phonology and morphology of Axininca Campa*, volume 66. Summer Institute of Linguistics Arlington Texas.

José Pereira-Noriega, Rodolfo Mercado-Gonzales, Andrés Melgar, Marco Sobrevilla-Cabezudo, and Arturo Oncevay-Marcos. 2017. Ship-lemmatagger: Building an nlp toolkit for a peruvian native language. In *International Conference on Text, Speech, and Dialogue*, pages 473–481. Springer.

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

Annette Rios. 2015. *A basic language technology toolkit for Quechua*. Ph.D. thesis, University of Zurich.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University.

Stefano Varese. 2004. *Salt of the mountain: Campa Asháninka history and resistance in the Peruvian jungle*. University of Oklahoma Press.

Jānis Zuters and Gus Strazds. 2019. Subword segmentation for machine translation based on grouping words by potential roots. *Baltic Journal of Modern Computing*, 7(4):500–509.

7 Supplemental Material

In this section we provide examples of suffixes in Section 7.1 to help understand the Ashaninka language. The examples presented are based on previous work (Hulden and Francom, 2012; Rios, 2010, pag. 2115) on lemmas and suffixes; and, is nearly identical to the Quechua normalizer from their work that produces human-readable labels based on a lexicon. In the second section (Section 7.2), a table is presented that shows the difference in alphabetical characters used as part of our normalization technique. There is a comparison between two authors that include final normalization by (Mihas, 2010) as the character set used in this paper. Lastly, there is a finalized view the **AshMorph** normalization process on a few words (*amõyasatzi*, *amoyasatzi*, *amoñasatzi*) which end up being normalized to *amonyasatzi*.

7.1 Suffix Examples

In this section, we present several tagging examples to help understand the strategy of normalization in Ashaninka. These are provide in English and Spanish counterparts to show the differences that one would encounter when attempting to normalize these words to Ashaninka's neighboring high-resource language, Spanish.

7.1.1 Verbal root:

```
"[=tash+INF@agantsi;antsi.][VRoot][=to.be.hungry
(EN: have/feel.hungry)]" : {tash}
```

7.1.2 Verbal root:

```
| "[=ameet+INF@aantsi.][VRoot][=to.have.sb's.hair.cut, to.shave
(ES: cortar.(el.pelo), afeitar)]" : {ameet}
```

7.1.3 Verbal suffix:

```
| "[--][-eNpa~empaa+tns@fut.+C.A@mid.v.+RSS@00000]
[MODALITY][+IRR]" : "@EP"[{empaa}]

 | "[=kishi+gndr@n.m.+inal@tsi.+mrph.phon@LEN:^k^Ø.][NRoot]
 [=hair (ES: cabello, pelo; PT: pelo; QU: chukcha)]" : {kishi}
```

7.1.4 Nominal suffix:

```
| "[--][-satsi+gndr@m.][NS:CPB][+CL:provenance][=human.provenance;
inhabitant.of.swh., one.who.dwells.swh.]" : {satsi}
```

7.2 Alphabet Comparison

Here, we present a comparison of two main alphabets that have been derived from various written Ashaninka dialects. The final normalization (Mihas, 2010) has made it easier for **AshMorph** to incorporate Ashananika's dialects into one language for normalization purposes.

Phoneme	Writing systems	
	Payne 1981	Mihas, Elena 2010
i	i	i
e	i	e
a	a	a
o	o	o
p	p	p
p^j		**py**
t	t	t
t^j	č	ty
k	k	k
k^j		**ky**
s	s+V_a	s
g	ɡ	
ʃ	s+V_i, ç+$V_{a,i,o}$	sh
ʧ	$č^h$	ch
ts^h	$ȼ^h$, t^h	ts
ts	ȼ	**tz**
h	h	h
m	m	m
n	n	n
ɲ	ñ	**ny**
ɾ	r	r
$ɾ^h$	r^y	**ry**
w	$\bar{w}$	**w (or v)**
j	y	y
N	N	**n, m (or n)**

Table 4: The first found alphabetical list of Pan-Ashaninka phonemes (Payne, 1981) along with the most current (Mihas, 2010).

7.3 Input Normalization

In this section, we provide a visualization of how a word that is written in various ways due to Ashaninka's complexity is normalized into one word. These normalization rules are part of the finite state transducer (FST) model introduced in this paper.

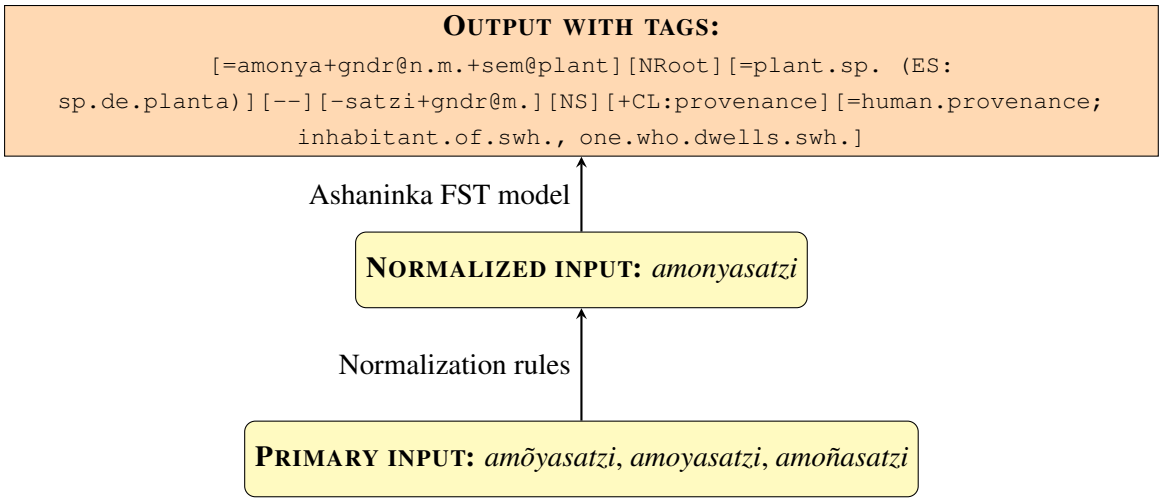

Figure 2: Normalization in the LOOKUP of the Ashaninka FST model which illustrates distinct input forms of a word being first converted into their normalized input and then into their resulting output.