Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5450–5459
Marseille, 20-25 June 2022

Preliminary Results on the Evaluation of Computational Tools for the Analysis of Quechua and Aymara

Marcelo-Yuji Himoro*, **Antonio Pareja-Lora***,**
* ATLAS (UNED - Universidad Nacional de Educación a Distancia)
** Universidad de Alcalá (UAH)
Madrid, Spain
mhimoro1@alumno.uned.es, antonio.parejal@uah.es

Abstract

This research has focused on evaluating the existing open-source morphological analyzers for two of the most widely spoken indigenous macrolanguages in South America, namely Quechua and Aymara. Firstly, we have evaluated their performance (precision, recall and F1 score) for the individual languages for which they were developed (Cuzco Quechua and Aymara). Secondly, in order to assess how these tools handle other individual languages of the macrolanguage, we have extracted some sample text from school textbooks and educational resources. This sample text was edited in the different countries where these macrolanguages are spoken (Colombia, Ecuador, Peru, Bolivia, Chile and Argentina for Quechua; and Bolivia, Peru and Chile for Aymara), and it includes their different standardized forms (10 individual languages of Quechua and 3 of Aymara). Processing this text by means of the tools, we have (i) calculated their coverage (number of words recognized and analyzed) and (ii) studied in detail the cases for which each tool was unable to generate any output. Finally, we discuss different ways in which these tools could be optimized, either to improve their performances or, in the specific case of Quechua, to cover more individual languages of this macrolanguage in future works as well.

Keywords: natural language processing, annotation, minority language, Quechua, Aymara

1. Introduction

Among the over 6,000 languages spoken in the world, only a minority of them have been provided so far with adequate computational resources, e.g. for their processing. Therefore, there is a clear need for new digital resources to be developed for these languages. This comprises, for instance, the creation of tools for their analysis or the improvement of the existing ones, if any. Another problem to be faced is linguistic variation, which usually requires adapting or extending the existing resources for different varieties or related languages, or even to develop new specific versions of these resources.

One of the possible approaches towards the creation of new tools for the processing of these under-resourced languages would be to build some state-of-the-art machine learning models (based on, e.g., deep learning). However, the scarcity of annotated and/or bilingual data that can be used for training is one of the limiting factors for the development of language processing and analyzing tools. Therefore, in most cases, improving, re-targeting and/or fine-tuning already developed tools (when they exist) seems an unavoidable task. This is in particular the reality of Quechua and Aymara, two languages which, despite being spoken by millions of people in South America, are still under-resourced languages in this sense, due to the few resources developed for their processing to date. More specifically, they are widely spoken in several countries and comprise a number of languages or varieties, and some computational tools have been developed for their processing hitherto. However, their actual coverage (as for variation) and precision (also wrt. the different varieties and/or languages in the family or macrolanguage), and thus, also their suitability, has not been assessed so far. Hence, the aim of this work has been to evaluate some morphological analyzers developed for Quechua and Aymara, originally developed for one of the languages or varieties of these two macrolanguages, and their potential suitability for some of their related languages or varieties. It also discusses the results obtained in the experiments performed for their evaluation, and how these tools can be extended to process more linguistic varieties.

The rest of the paper has been structured as follows. Section 2 briefly introduces the macrolanguages (i.e., Quechua and Aymara) and Section 3 presents the morphological analyzers available for both of them. Section 4 and Section 5 describe the evaluation experiments. Section 6 discusses the results found in the evaluation. Finally, Section 7 wraps up the final conclusions.

2. The Macrolanguages – an Overview

2.1. Quechua

Quechua is one of the world's major language families (Torero, 1983) or macrolanguages[1]. It is spoken in a large area of South America, stretching southwards from Colombia through Ecuador, Peru, Bolivia, Chile and Argentina. It is a highly agglutinating language (as opposed to isolating or fusional languages) and its word order is predominantly SOV (Subject-Object-Verb).

Traditionally, the Quechua macrolanguage or family is divided into two main branches: Quechua I (Central or

[1] https://iso639-3.sil.org/code/que

QI) and Quechua II (Peripheral). Branch II, in turn, is further divided into three other branches: Quechua IIA (Yungay or QIIA), Quechua IIB (Northern Chinchay or QIIB) and Quechua IIC (Southern Chinchay or QIIC) (Torero, 1983).

2.2. Aymara

Aymara is spoken in Peru, Bolivia and Chile. However, conceptualizing and subclassifying it is more difficult and controverted. Also a predominantly SOV highly agglutinating language, traditionally Aymara has been considered to belong to the Southern Aymara branch of the Aymaran language family, alongside with Jaqaru and Kawki, which belong to the remaining branch, namely the Central Aymara branch (spoken in Peru). The Southern Aymara branch can be further divided into the Northern, Intermediate and Southern dialects (Hardman, 2001; Cerrón-Palomino, 2000), across the three aforementioned countries.

However, Ethnologue (as well as ISO 639-2:1998; ISO 639-3:2007) provides a different classification. According to it, the Aymaran language family is divided into two branches: Tupe (containing Jaqaru[2]) and Aymara (a macrolanguage[3] containing the Central[4] and Southern Aymara[5] languages) provides a different classification. These two sub-branches do not completely match with any of the sub-branches found in other more traditional classifications, such as the one previously mentioned. Nevertheless, both classifications agree on the fact that the different Aymara languages are spoken in Peru, Bolivia and Chile. As for this paper, thus, we will not refer to any of the classifications, as more experiments must be carried out in order to determine how they relate (or not) to the computational processing properties and needs of the different languages (or language varieties). Accordingly, whenever needed, we will refer to the country or the region where the sample text being processed corresponds to.

3. The NLP Tools Evaluated

3.1. The NLP Tools for Quechua

3.1.1. SQUOIA

SQUOIA was a project from the University of Zurich that aimed at building a hybrid Machine Translation system between Spanish-Cuzco Quechua and Spanish-German. For Cuzco Quechua, it includes a Spanish-Quechua bilingual dictionary in Apertium format, treebanks annotated with dependency trees, a morphological analyzer, a spell checker, and a pipeline for analyzing and parsing Southern Quechua text (Rios, 2015).

In this research, we are particularly interested in the morphological analyzer included in the toolkit and, hence, we will refer to it henceforward simply as "SQUOIA". It was developed in xfst (Xerox Finite State Tool) and, consequently, it uses Finite-State Transducers to analyze words. The whole system includes 5 transducers: two of them recognize Quechua words (handling, respectively, regular orthographies and some orthographic variations), another two deal with Spanish words (one for Spanish loans adapted to Quechua and another one for Spanish words in their original spelling), and the other one (the "guesser" henceforth) infers the root and the category of word forms not found in the dictionary.

Even though it was developed for Cuzco Quechua (QIIC sub-branch), it also takes into account phonological (and thus, also orthographic) differences found in other languages in the same sub-branch, such as the lack of glottalized and aspirated stops in the Ayacucho (Peru) and Santiagueño (Argentina) Quechua languages (Rios, 2015). As such, the tools should be able to analyze text from other QIIC varieties.

3.1.2. AntiMorfo

AntiMorfo is a morphological analyzer for Cuzco Quechua (QIIC sub-branch) and Spanish. It was developed in Python using finite state transducers (Gasser, 2009; Gasser, 2011). It can be imported as a Python package and provides methods to analyze words. Two kinds of output are possible, namely a human-friendly one, describing grammatical features of the word, and the more detailed, raw output of the transducers.

3.2. The NLP Tools for Aymara

3.2.1. Aymara Morph Analyzer

Aymara Morph Analyzer is a morphological analyzer for Aymara. It was developed also in xfst by Beesley (2003), who only specifies in his paper that the language is spoken in Peru and Bolivia without giving further details on what Aymara language the tool is meant to process. The version evaluated in this research has been retrieved from an archived Google Code repository[6], since the previous repository where it was previously available[7] has been taken down. This implementation also includes a guesser function for roots not found in its internal dictionary.

4. First Evaluation Experiment

The goal of this first experiment was to evaluate the performance of the morphological analyzers on the individual languages for which they were developed (SQUOIA and AntiMorfo for Cuzco Quechua and Aymara Morph Analyzer for Aymara as spoken in Peru and Bolivia).

4.1. Description of the Experiment

The annotated data used in this evaluation were drawn from the Quechua Treebank (Rios, 2015) for Quechua and from the "Aymara On The Internet" website (Beck

[2] https://iso639-3.sil.org/code/jqr

[3] https://iso639-3.sil.org/code/aym

[4] https://iso639-3.sil.org/code/ayr

[5] https://iso639-3.sil.org/code/ayc

[6] https://code.google.com/archive/p/hinantin/

[7]

et al., 2008)[8] for Aymara. The Quechua data were encoded using UTF-8 character set, while the Aymara data were originally converted from ISO 8859-1 (latin1) to UTF-8. About 1,000 words, comprising only full sentences, were randomly selected and extracted from both sources (1,058 words in Quechua and 1,059 words in Aymara). While the former contains texts about agriculture, development aid, economy, education, media, and culture as well as biographic texts (i.e., a more formal register), the latter consists of annotated dialogues (hence, they represent mainly the colloquial register of the language). This selection is justified by the low availability of manually annotated data that could be used as a reference in both languages.

To carry out the experiment, scripts were written in Python to read the annotated data, process it by means of the tools, obtain its output and compare it with the annotated data.

Due to the complexity and the problems faced to evaluate the annotations, especially in the case of AntiMorfo (for its particular output), this experiment has focused on assessing the accuracy of the morpheme segmentation (but not the accuracy of the annotation tags, which will be addressed in future works).

4.2. Results

To begin with, it should be mentioned that, when analyzing the results of the experiment, we have considered the morphological analysis to be correct only if the morpheme segmentation provided by the tool and by the annotated data matched completely.

Tables 1, 2 and 3 show the different measures values (precision, recall and F1 score) found for SQUOIA, AntiMorfo and Aymara Morph Analyzer, respectively. Precision is given by the ratio of correctly segmented tokens divided by the tokens segmented by the tool. Similarly, recall is defined by the number of tokens that the tool was able to correctly segment divided by the number of tokens in the original text. F1 score is the harmonic mean of both measures. Note that, in all these measures, reoccurring errors are counted as many times as these tokens appear in the texts.

Comparing the values obtained with the two tools for Quechua processing, it might seem like AntiMorfo outperforms SQUOIA due to its higher value for precision. However, its low recall indicates that a substantial number of tokens could not be analyzed by the tool. This recall value is quite lower than that of SQUOIA, whose F1 is clearly higher than AntiMorfo's as well. Therefore, it could be stated that SQUOIA outperforms AntiMorfo when processing the Cuzco Quechua variety.

For reference, Table 4 shows the same measures for morpheme segmentation in other languages of the same typology (i.e., agglutinative).

Precision	70.63%
Recall	67.27%
F1	68.91%

Table 1: SQUOIA performance values

Precision	79.35%
Recall	41.80%
F1	54.76%

Table 2: AntiMorfo performance values

Precision	64.9%
Recall	57.85%
F1	61.17%

Table 3: Aymara Morph Analyzer performance values

language	**Precision**	**Recall**	**F1**
Basque	28%	99.41%	43.7%
Finnish	92.39%	81%	86.32%
Japanese	93.5%	96.5%	95%
Turkish	90.8%	90.22%	90.51%

Table 4: Precision, recall and F1 score values for morphological segmentation in other agglutinative languages: Basque (Aduriz et al., 2020), Finnish (Pirinen et al., 2016), Japanese (Higashiyama et al., 2021) and Turkish (Seeker and Çetinoğlu, 2015).

Firstly, for the Quechua analyzing tools, as shown in Table 5, tokens incorrectly processed by SQUOIA were classified as "diverging segmentation criteria" (there is a disagreement between our reference data and the annotation given by the tool, as shown by Examples 1b and 1a, 2b and 2a, and 3b and 3a[9]) or "incorrect segmentation". In the cases of the former, there are a few of them which occur because the tool gives a normalized form of the root or some of the affixes instead of presenting the form that occurs in the token ("output needs adjustments"; see Examples 4a and 4b). As for unprocessed tokens, it was mainly caused by named entity occurrences (see Example 5), though it was also caused by Spanish loans or borrowings or English terms instances not found in the internal dictionary (see Example 6 and 7). Finally, as for the incorrectly segmented tokens, the causes were that it included a named entity (see Example 8) or due either to the lack of a suitable Spanish root in the internal dictionary (see Examples 9). In most of the cases, at least for native words, the guesser function of the tool was able to correctly identify and segment the token (see

[8] According to the source, the data are essentially in Aymara as spoken in La Paz (Bolivia) with some influence of Tiahuanaco (near the border between Bolivia and Peru).

[9] In all examples, where applicable, the data included in the (a) subclause refers to the original annotated data and the (b) subclause refers to the annotations obtained with the tool.

Examples 10b and 10a).

(1) qallarispa
'starting'
a. qallari-spa
b. qalla-ri-spa

(2) runakunap
'people's'
a. runakuna-p
b. runa-kuna-p

(3) huñunakuykuna
'gatherings, events'
a. huñuna-ku-ykuna
b. huñu-na-ku-y-kuna

(4) manan
'not' (with assertive suffix)
a. mana-n
b. mana-m

(5) Perú
'Peru'

(6) miryu-pi
'in the media'

(7) people

(8) Laospi
a. Laos-pi
'in Laos'
b. lao-s-pi (lado-kuna-pi)
'beside'

(9) kirusini
'kerosene, paraffin, lamp oil'

(10) kharwayusqa
'piled up'
a. kharwayu-sqa
b. kharwa-yu-sqa

As for AntiMorfo (see Table 6), there was a significant number of unprocessed tokens. The main cause for this was that the tool was unable to correctly process the affixes, even when the root was included in the internal dictionary (see Example 11). Other reasons were the occurrence of Spanish loans and borrowings (see Example 6), the corresponding root not being included in the internal dictionary (see Example 12), followed by named entities (see Example 5) and the occurrence of English terms (see Example 7). Similar to what was observed for SQUOIA, there were cases in which the segmentation in the reference data and the one provided by the tool did not match (see Example 13a and 13b and 14a and 14b). In some cases, such as Example 15 (previously described for SQUOIA as well), a normalized one is given by the tool. There were also a few cases of incorrect segmentation (see Example 16).

(11) chaypi
'here'

(12) **ñan**pi
'on the way'

(13) rimanku
'(they) speak'
a. rima-n-ku
b. rima-nku

(14) waynakunapaq
'youngster's'
a. wayna-kuna-paq
b. waynakuna-paq

(15) karan
'[it] was' (emphatic)
a. ka-rqa-mi
b. ka-ra-n

(16) niwaq
'he/she used to tell me'
a. ni-wa-q
b. ni-waq

As for Aymara, a similar table for Aymara Morph Analyzer is provided here (see Table 7). There is a substantial number of tokens that were incorrectly analyzed by the tool, mainly because either (i) the root was not found in the internal dictionary (see Example 17), (ii) they were Spanish loans or borrowings (see Example 18), or (iii) they were named entities (see Example 19), but also a number of cases for which no specific reason could be identified (see Example 20). In many cases, similarly to what has been reported for the Quechua tools, the segmentation criteria of our reference data do not match (see Example 21). Regarding the unprocessed tokens, this problem was mostly due to the lack of a suitable root in the internal dictionary (see Example 22), but also when a Spanish loan or borrowing (see Example 23), or a named entity (see Example 24) was being handled. Eventually, we were unable to determine the reason why some other tokens could not be processed (see Example 25). Besides, some tokens missed an annotation in our reference data (see

error	occurrences	Examples
diverging segmentation criteria	**207**	-
output needs adjustments	(37)	4
others	(170)	1 2 3
incorrect segmentation	**8**	-
named entities	(4)	8
Spanish loans/borrowings	(4)	9
not processed	**41**	-
named entities	(23)	5
Spanish loans/borrowings	(12)	6
English terms	(6)	7

Table 5: Absolute frequency of errors by type with SQUOIA

error	occurrences	Examples
diverging segmentation criteria	**76**	-
output needs adjustments	(9)	15
others	(67)	13 14
incorrect segmentation	**13**	16
not processed	**387**	-
root not in the dictionary	(21)	12
named entities	(34)	5
Spanish loans/borrowings	(127)	6
English terms	(7)	7
others	(198)	11

Table 6: Absolute frequency of errors by type with AntiMorfo

Example 26).[10] Some other tokens (accounted for as misspellings in Table 7) were spelled with a diverging orthography in the annotation (see Example 27). Given that the data were manually annotated and available in two different orthographies, it is very likely that these omissions and misspellings were unintentional. These problematic cases account for about 3.5% of the tokens.

(17) **lastus**kam
‘until midday’
a. lastusi-kama
b. lastu-si-ka-m(a)

(18) **turista**sti
‘as for the tourist’
a. turista-sti
b. turi-chi-ta-sti / turi-su-ta-sti / turi-si-ta-sti

(19) **Rusaliya**
‘Rosalía’
a. Rusaliya
b. Rusali-ya

(20) phinats
‘the pile of potatoes?’
a. phina-ti-sa
b. phina-t(a)-s(a)

(21) **ururakisä**
‘(it’s) day already’
a. uru-raki-sä
b. uru-raki-sa-”

(22) khä (short form of “khaya”)
‘that’ (as a demonstrative adjective)

(23) kustal
‘bag (of cereals, seeds, etc.)’

(24) **Justina**sti
‘what about Justina?’

(25) apxaruwayamxa
‘take (it) away’

(26) wal
‘well, good, carefully’

(27) *khuchhi (khuchi)
‘pig’

[10] This is due to the fact that, even though the morphological analyzers evaluated here do not pay attention to the context in which a token occurs to process it, only full sentences were extracted from the original texts, despite they contained some tokens that were not annotated.

error	occurrences	Examples
diverging segmentation criteria	**108**	21
incorrect segmentation	**136**	-
root not in the dictionary	(89)	17
named entities	(11)	19
Spanish loans/borrowings	(7)	18
others	(29)	20
not processed	**84**	-
root not in the dictionary	(43)	22
named entities	(25)	24
Spanish loans/borrowings	(9)	23
others	(7)	25
misspelling	**2**	27
not evaluated (missing annotation)	**35**	26

Table 7: Absolute frequency of errors by type with Aymara Morph Analyzer

5. Second Evaluation Experiment

The goal of this experiment was to evaluate to what degree the tools were capable of producing any kind of morphological analysis on other individual languages of the same macrolanguage, regardless of its accuracy.

5.1. Description of the Experiment

The data used to evaluate the coverage of the tools consist of about 500 words of oral literature for each sample, prioritizing full texts rather than random excerpts. It should be remarked that rather than working with each specific language, we decided to work (whenever possible) with their corresponding standardized forms, as they appear in publications edited by the corresponding Ministry of Education in the countries where they are spoken. Sometimes a standardized language may cover multiple languages, such as is the case of Ecuadorian Quechua ("Kichwa Unificado") or Bolivian Quechua ("Quechua Normalizado"), which encompasses all Quechua languages spoken in the respective countries. An exception was made for the Argentinean Santiagueño Quechua and the Colombian Inga Quechua text samples. The former comprises selected text from the book "Wawqes Pukllas" (Andreani, 2014), of a similar genre as the text in the other samples. As for the latter, due to the low availability of sources, the full text of "The Parable of the Prodigal Son" and part of "The Parable of the Good Samaritan" were extracted from the Bible, as they bear at least some similarity in genre to the other texts selected. By doing so, we aimed at guaranteeing a minimum degree of language standardization in the samples.

Hence, for Quechua, 10 different samples have been collected, namely Chawpi (QI-pe, representing the Central branch), Inkawasi-Kañaris (QIIA-pe, Yungay branch), Kichwa (QIIB-pe, Northern Chinchay branch), Qullaw (QIIC-pe1, representing Cuzco and Puno Quechua, Southern branch), Chanka (QIIC-pe2, representing Ayacucho Quechua, Southern branch) from Peru, Inga (QIIB-co, Northern Chinchay branch) from Colombia, Kichwa Unificado (QIIB-ec, Northern Chinchay branch) from Ecuador, Quechua Normalizado (QIIC-bo, Southern branch) from Bolivia, Qhishwa (QIIC-cl, Southern branch) from Chile and Santiagueño (QIIC-ar, Southern branch) from Argentina. As for Aymara, we have 3 samples, one from each country (Peru, Bolivia and Chile, denoted by aym-pe, aym-bo and aym-cl, respectively). Despite the known lexical differences found in the Aymara Chilean variety, they could not be verified in the aym-cl sample, leading us to believe that a standard similar to that used in Bolivia and Peru is also being applied in Chile.

Again, scripts were written in Python to read the samples, tokenize the text and process it by means of the tools. The coverage of the tools for each specific sampled variety is calculated by dividing the number of tokens processed the number of tokens in the sample.

Due to the lack of availability of manually annotated data for the different individual (standardized) languages of Quechua and Aymara (to the best of our knowledge), this evaluation only assesses the number of tokens processed, regardless of the correctness of the morphological analysis and the annotations tags.

5.2. Results

Tables 8 and 9 show the size (in number of tokens) of the sample of each of the languages included in this experiment.

As shown in Table 10 and 11, both SQUOIA and AntiMorfo were able to analyze more tokens for Southern Quechua languages, the group to which Cuzco Quechua also belongs. On the one hand, as expected, AntiMorfo reached its highest processing rate for Quechua Qullaw (QIIC-pe1), the sample representing Cuzco Quechua, thus performing better for all languages from the QII branch than for the QI branch. The only exception was the Colombian sample (QIIB-co), which uses a rather diverging orthography and it probably impacted the performance of the tool.

sample	word count
QI-pe	541
QIIA-pe	508
QIIB-pe	528
QIIB-ec	514
QIIB-co	527
QIIC-pe1	542
QIIC-pe2	546
QIIC-bo	516
QIIC-cl	518
QIIC-ar	517

Table 8: Size (in number of tokens) of the Quechua samples used in Experiment 2

sample	word count
aym-pe	505
aym-bo	500
aym-cl	505

Table 9: Size (in number of tokens) of the Aymara samples used in Experiment 2

On the other hand, surprisingly, SQUOIA was proportionally able to process more words from Quechua Chanka (QIIC-pe2) sample. We could not determine the reason for this. Many factors could have influenced these results, such as the frequency of named entities in the sample or its size itself. For both, the lowest values inside Southern Quechua (QIIC sub-branch) were found for the Argentinean sample (QIIC-ar), which could partially be explained by the high frequency of Spanish borrowings and loans in the sample and by the fact that these texts were not written by language authorities.

variety	squoia	squoia with guesser
QI-pe	53.98%	77.27%
QIIA-pe	51.47%	72.69%
QIIB-pe	58.41%	68.58%
QIIB-ec	54.04%	64.5%
QIIB-co	40.27%	53.59%
QIIC-pe1	94.66%	97.24%
QIIC-pe2	96.27%	99.07%
QIIC-bo	86.85%	94.82%
QIIC-cl	92.48%	97.88%
QIIC-ar	71.18%	79.5%

Table 10: SQUOIA performance

variety	AntiMorfo
QI-pe	40.53%
QIIA-pe	42.44%
QIIB-pe	50.09%
QIIB-ec	45.76%
QIIB-co	24.7%
QIIC-pe1	84.16%
QIIC-pe2	70.52%
QIIC-bo	78.49%
QIIC-cl	84.58%
QIIC-ar	42.17%

Table 11: AntiMorfo performance.

variety	SQUOIA+AntiMorfo
QI-pe	77.46%
QIIA-pe	74.66%
QIIB-pe	69.32%
QIIB-ec	66.27%
QIIB-co	55.39%
QIIC-pe1	98.53%
QIIC-pe2	99.25%
QIIC-bo	96.61%
QIIC-cl	99.42%
QIIC-ar	81.62%

Table 12: Performance of SQUOIA and AntiMorfo combined.

variety	AymaraMorph	AymaraMorph with guess
aym-pe	44.51%	100%
aym-bo	53.53%	100%
aym-cl	55.38%	100%

Table 13: Performance of Aymara Morph Analyzer.

Regarding the Quechua languages from the other branches, SQUOIA performed best for Quechua Chawpi (QI-pe), followed by Quechua Inkawasi-Kañaris (QIIA-pe) and all the QIIB languages. It was expected that the tools would be able to process more words in all the languages in the QII branch, as the Southern Quechua languages (like Cuzco Quechua) also belong to this branch. Nevertheless, it performed better on the QI branch. This might be counter-intuitive. However, considering that this evaluation is restricted to segmentation and it does not cover the accuracy and the correctness of the morphological analysis tags, no conclusions can be drawn from these figures.

Some factors possibly influencing the results could be the rules encoded in SQUOIA to handle the linguistic diversity found in the Southern Quechua branch. Another hypothesis is that the phonetic and orthographic differences found in the Quechua languages from other sub-branches in QII might have affected the performance of the tool. For instance, Inkawasi-Kañaris Quechua uses the grapheme “ĉ”, not found in the other Quechua languages. QIIB languages present a number of more or less systematic orthographic differences (at least in their standardized form; see Table 14), besides the absence of glottalized and aspirated stops (previously mentioned in 3.1.1 as a peculiarity of Ayacucho and Santiagueño Quechua in the QIIC sub-branch, but also shared by some Quechua languages in other branches). Should these peculiarities be taken into account when retargeting these tools, it would be expected that the performance of the tools would be highly improved when processing these varieties.

English	QIIA-pe	QIIB-ec	QIIC-pe1
you	q̲am	k̲an	q̲am
like that	s̲h̲ina	s̲h̲ina	h̲ina
new	mus̲h̲uq̲	mus̲h̲uk̲	mus̲uq̲
teacher	yaĉ̲achikuk̲	yac̲h̲achikuk̲	yac̲h̲achikuq̲

Table 14: Orthographic differences found in some Quechua languages.

As shown in Table 12, there is little gain in combining both SQUOIA and AntiMorfo. It could be argued that AntiMorfo is outperformed by SQUOIA in most cases. For Aymara Morph Analyzer, the difference in performance found across the different samples was not noticeably different, although the tool did slightly worse for the Peruvian sample. However, no obvious difference between the samples was found.

6. Discussion and future work

As discussed in Section 5.2, a reevaluation of SQUOIA could lead to precision and recall values close to 100% for Cuzco Quechua, thus entailing that it would supersede AntiMorfo at least for our test data. In any case, clearly, some fine-tuning needs to be done in order to adapt the tools to the linguistic differences among different Quechua languages. Further experimentation will be needed to find out whether separated versions would be needed for each sub-branch (QI, QIIA, QIIB and QIIC) or a single improved version would be feasible. Developing separated versions would make debugging easier, at the expense of requiring the maintenance of at least 4 different versions. Likewise, a single version that can cover different Quechua languages might be difficult to maintain. Another possible drawback is that the performance for Cuzco Quechua might be degraded by the addition of roots and morphemes from other Quechua languages. However, further experiments and evaluation will be performed in this direction. Hence, a more thorough evaluation of the tools for Quechua languages in sub-branches other than Cuzco Quechua using manually annotated data will help us have a wider picture of the situation.

For Aymara Morph Analyzer, adding more roots to its internal dictionary could be a good way to quickly improve its performance, in so far as the hitherto unprocessed tokens are concerned. A prior analysis and study of the incorrectly analyzed words are also being planned, in order to improve and/or correct the rules encoded in xfst.

Finally, future work will try and find out as well if unsupervised machine learning methods could help build some alternative tools that could outperform the ones evaluated here. This would require compiling a huge amount of unannotated Quechua and Aymara text, which might be not easy or even feasible (after all, they are under-resourced languages). However, this approach should be followed, at least to prove (if this is the case) that it does not suit this task, at least with the resources available.

7. Conclusions

In this research, our goal has been to preliminarily evaluate the existing morphological analyzers for two of the most spoken indigenous macrolanguages or languages of South America, namely Quechua and Aymara. For Quechua, we have evaluated SQUOIA and AntiMorfo, both developed for Cuzco Quechua. For Aymara, we have evaluated Aymara Morph Analyzer.

This has been achieved by carrying out two separate evaluation experiments. The first experiment aimed at evaluating the performance of these tools for the languages they were developed for. The test data consisted of about 1,000 words in both Cuzco Quechua and Aymara manually annotated, extracted from the Quechua Treebank and "Aymara on the Internet", respectively. Due to the complexity of the output of AntiMorfo and its lack of documentation, this evaluation was limited to morpheme segmentation and, hence, the correctness of the morphological analysis tags has not been assessed. While evaluating SQUOIA and AntiMorfo (for Quechua), we have identified substantial disagreement in the manually annotated data and the output of the tools. Thus, the tools have presented a precision of approximately 70% and 80% and recall of 67% and 41%, respectively. However, in most cases when a disagreement occurs, both the human annotation and the SQUOIA annotation could be regarded as correct. Accordingly, were these cases considered correct as for the tool annotations, SQUOIA would reach precision and recall values as high as 95%. In contrast, for Aymara, Aymara Morph Analyzer reached about 64% and 57% for precision and recall, respectively, with many cases of incorrect processing and unprocessed tokens.

The second evaluation experiment, in turn, was designed to evaluate how the tools perform with different Quechua languages and Southern Aymara varieties. In an attempt to limit the evaluation to standardized varieties, priority was given to resources edited by the respective Ministry of Education of the countries where both languages are spoken (except for Argentina and Colombia). Hence, 10 samples were selected: five in different Quechua languages from Peru and one for each of the other countries (Colombia, Ecuador, Bolivia, Chile and Argentina). For Aymara, three samples were collected: one for Peru, one for Bolivia and one for Chile. Due to the lack of annotated data, this evaluation only assesses the proportion of tokens processed (in other words, in what cases the tools have been able to process the token and provide at least one candidate morphological analysis). For Quechua, as expected, both tools have done best for the varieties in the Southern Quechua branch. More specifically, on the one hand, SQUOIA has presented some surprising results: despite being developed for Cuzco Quechua (QIIC branch), the tool has been able to process more tokens from the Central (QI) branch than from other varieties in the same branch (QIIA and QIIB). This has been probably caused by (1) the phonetic (and consequently orthographic) specificity of the other varieties in the same branch; besides (2) the rules encoded to handle the variations found in the Southern Quechua languages (partially shared by Central Quechua). AntiMorfo, on the other hand, has presented results according to what was expected, being able to process more tokens for Quechua languages inside the QII branch. For Aymara, no noticeable differences were found for the different samples.

In general terms, regarding the processing of Southern Quechua (QIIC), SQUOIA outperforms AntiMorfo; however, both their performances should be increased somehow. Therefore, as future work, we intend to carry out two kinds of experiments towards this end: (1) to fine-tune them in order to process other Quechua languages (especially those in the other branches); and (2) to increase their overall performance by combining their results, such as in (Pareja-Lora, 2012), both before and after fine-tuning them. Besides, and prior to (2), it would be most suitable to perform yet another experiment in order to evaluate the performance of SQUOIA for at least one Quechua language from a different branch, using manually annotated data. This would shed some light on the performance of the tool when applied to a language outside the Southern branch and how to better proceed to adapt it.

Regarding the processing of Aymara Morph Analyzer, besides expanding the internal dictionary, future work includes studying in what cases the tool has been unable to provide a correct morphological analysis. This analysis seems essential to improve its current implementation.

Last, but not least, building some other tools by means of unsupervised machine learning methods on huge amounts of unannotated Quechua and Aymara text will be done too, in order to determine if better performance can be reached this way.

8. Bibliographical References

Aduriz, I., Arriola, J. M., Artola, X., Beloki, Z., Ezeiza, N., and Gojenola, K. (2020). Morfeus+: Word parsing in basque beyond morphological segmentation. *Word Structure*, 13(3):283–315.

Andreani, H. A. (2014). Wawqes pukllas. prácticas juveniles de escritura quichua (argentina). *Bellaterra Journal of Teaching & Learning Language & Literature*, 7:38–56.

Beck, H. W., Legg, S., Hardman, M., Lord, G., Llanque-Chana, J., and Lowe, E. (2008). A collaborative multilingual database project on aymara implemented in peru and bolivia. In *2008 Providence, Rhode Island, June 29–July 2, 2008*, page 1. American Society of Agricultural and Biological Engineers.

Beesley, K. R. (2003). Finite-state morphological analysis and generation for aymara. In *Proceedings of the Workshop of Finite-State Methods in Natural Language Processing: 10th Conference of the European Chapter of the Association for Computational Linguistics)*, pages 19–26, Budapest. Association for Computational Linguistics.

Cerrón-Palomino, R. (2000). *Lingüística aimara*. Biblioteca de la tradición oral andina. Centro de Estudios Regionales Andinos "Bartolomé de Las Casas".

Gasser, M. (2009). Semitic morphological analysis and generation using finite state transducers with feature structures. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 309–317, Athens, Greece, March. Association for Computational Linguistics.

Gasser, M. (2011). Antimorfo 1.1 user's guide.

Hardman, M. (2001). *Hardmann / Aymara*. LINCOM studies in Native American linguistics. LINCOM Europa.

Higashiyama, S., Utiyama, M., Watanabe, T., and Sumita, E. (2021). User-generated text corpus for evaluating Japanese morphological analysis and lexical normalization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5532–5541, Online, June. Association for Computational Linguistics.

International Organization for Standardization (ISO). (1998a). ISO 639-2:1998 – Codes for the representation of names of languages – Part 2: Alpha-3 code. Standard, International Organization for Standardization (ISO), Geneva, CH.

International Organization for Standardization (ISO). (1998b). ISO 639-3:2007 – Codes for the representation of names of languages — Part 3: Alpha-3 code for comprehensive coverage of languages). Standard, International Organization for Standardization (ISO), Geneva, CH.

Pareja-Lora, A. (2012). *Providing Linked Linguistic and Semantic Web Annotations: The OntoTag Hybrid Annotation Model*. LAP Lambert Academic Publishing, Chisinau.

Pirinen, T., Toral, A., and Rubino, R. (2016). Rule-based and statistical morph segments in english-finnish smt. In *Proceedings of the Second Interna-*

tional Workshop on Computational Linguistics for Uralic Languages, pages 56–69, Szeged.

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

Seeker, W. and Çetinoğlu, O. (2015). A Graph-based Lattice Dependency Parser for Joint Morphological Segmentation and Syntactic Analysis. *Transactions of the Association for Computational Linguistics*, 3:359–373, 06.

Torero, A. (1983). *La familia lingüística quechua. América Latina en sus lenguas indígenas*. Monte Ávila, Caracas.