PRELIMINARY RESULTS ON THE EVALUATION OF COMPUTATIONAL TOOLS FOR THE ANALYSIS OF QUECHUA AND AYMARA

Marcelo Yuji Himoro, Antonio Pareja-Lora

ATLAS Applying Technology to LAnguageS

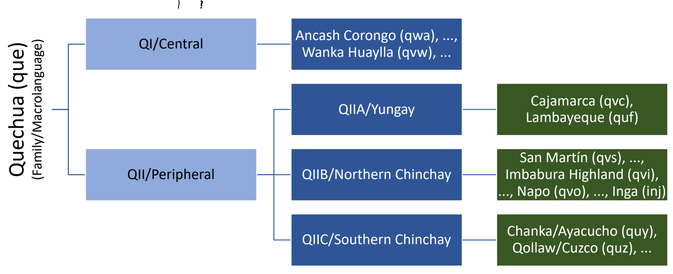
Universidad de Alcalá

Map adapted from:
Rios, A. (2010). *Applying Finite-State Techniques to a Native American Language: Quechua.* Master's thesis: Universitaät Zürich.

Quechua & Aymara

Aymara
Quechua I
Quechua II (A, B, C)
Aymara/Quechua IIC (overlapping area)

- Two of the most spoken languages in South America.
- Under-resourced languages, despite their number of speakers.
- Quechua is one of the world's language families according to some linguists (Torero, 1983) or a macrolanguage according to ISO 639-3:2007. Its varieties or languages are classified as follows:



- Aymara is considered a language according to some linguists (Hardman, 2001), but a macrolanguage according to ISO 639-3:2007. Although its subclassification is controversial, both agree in that it (as a language or a macrolanguage) belongs to the Aymara language family.

The goal of this research was:

(1) to evaluate existing morphological analyzers for Cuzco Quechua and Peruvian/Bolivian Aymara.

(2) to evaluate the suitability of the tools for other varities of these languages.

TOOLS

SQUOIA

- Project aimed at building a Spanish-Quechua and Spanish-German hybrid MT (Machine Translation) system (Rios, 2005).
- Includes a morphological analyzer implementing a finite-state transducer in xfst (Xerox Finite State Tool) language.
- For **Qullaw/Cuzco Quechua (QIIC)**, but also fairly robust regarding some phonological and orthographic differences found in other variants/languages in the QIIC branch.
- Implements a "guesser" that infers roots not found in the dictionary.
- Available at: https://github.com/a-rios/squoia.

Antimorfo

- Python package for Quechua and Spanish word analysis and generation (Gasser, 2009; 2011).
- Uses finite-state transducers in Python.
- For **Qullaw/Cuzco Quechua (QIIC)**.
- Two kinds of output: raw and human-friendly.
- Available at: https://cgi.luddy.indiana.edu/~gasser/Research/software.html.

Aymara Morph Analyzer

- Morphological analyzer for Aymara (Beesley, 2003).
- Uses finite-state transducers in xfst.
- For **Peruvian and Bolivian Aymara** (variant not specified).
- Implements a "guesser" that infers roots not found in the dictionary.
- Previously available at: https://github.com/hinantin/AymaraMorph.
- Archived at: https://code.google.com/archive/p/hinantin/source/default/source.

FUTURE WORKS

QUECHUA

- Fine-tuning the Quechua tools to process other Quechua languages.
- Combining SQUOIA and AntiMorfo to increase the overall performance.
- Manually annotating data of a variant from a different branch and evaluating SQUOIA and AntiMorfo on it (both segmentation and annotation tags).

AYMARA

- Expanding AymaraMorph internal dictionary.
- Evaluating the annotation tags of AymaraMorph and studying in what cases it was unable to provide a correct analysis.

BOTH

- Building and trying unsupervised machine learning-based morpheme segmentation methods.
  - Possible limitation: unavailability of huge amounts of Quechua and Aymara data.

EXPERIMENT 1

- Evaluate the morphological analyzers on the individual languages for which they were developed.
- Limited to morpheme segmentation evaluation (no evaluation of annotation tags).
- A token is considered correctly analyzed if the segmented parts match 100% the gold standard (no partial scoring).

DATA (QUECHUA)

- Full sentences, randomly selected.
- **Source:** gold standard – Quechua Treebank (Rios, 2015).
- **Variant:** Qullaw/Cuzco Quechua
- **Size:** ~1,000 words.
- **Annotation:** morphological, manual.
- **Genre:** texts about agriculture, development aid, economy, education, media, culture and biographic texts.
- **Register:** mainly formal.

DATA (AYMARA)

- Full sentences, randomly selected.
- Source: "Aymara On The Internet" website (Beck et al., 2008).
- Size: ~1,000 words.
- Annotation: morphological, manual.
- Genre: dialogues.
- Register: colloquial

EXPERIMENT 2

- Assesses the coverage of the tools for analyzing the different language variants (regardless of their correctness).
- Due to the lack of annotated data, correctness is not evaluated (only assesses the number of processed tokens).
- If the tool returns any output for a token, it gets a 1 score. Otherwise, it gets a 0 score.

DATA (QUECHUA)

- Full texts, randomly selected.
- **Source:** books published by the Ministry of Education for Peru (pe), Bolivia (bo), Chile (cl) and Ecuador (ec), New Testament for Colombia (co) and Guillín et al. (2021) for Argentina (ar).
- **Variants:** (1) Chawpi/Central (QI-pe); (2) Inkawasi-Kañaris (QIIA-pe); (3) Kichwa (QIIB-pe); (4) Kichwa Unificado (QIIB-ec); (5) Inga (QIIB-co); (6) Qullaw/Cuzco (QIIC-pe1); (7) Chanka/Ayacucho (QIIC-pe2); (8) Quechua Normalizado (QIIC-bo); (9) Qishwa (QIIC-cl); (10) Santiagueño (QIIC-ar).
- **Size:** ~500 words/sample.
- **Genre:** oral literature (except for Colombia [QIIC-co]).

DATA (AYMARA)

- Full texts, randomly selected.
- **Source:** books published by the Ministry of Education (Peru, Bolivia, Chile)
- **Variants:** Peruvian (aym-pe), Bolivian (aym-bo), Chilean (aym-cl).
- **Size:** ~500 words/sample.
- **Genre:** oral literature.

RESULTS (EXPERIMENT 1)

SQUOIA

error	occurrences
diverging segmentation criteria	**207**
output needs adjustments	(37)
others	(170)
incorrectly segmented	**8**
named entities	(4)
Spanish loans/borrowings	(4)
not processed	**41**
named entities	(23)
Spanish loans/borrowings	(12)
English terms	(6)
total	**256**

	SQUOIA
Precision	70.63%
Recall	67.27%
F1	68.91%

- In many cases, the tool and the gold standard did not agree, but both could be considered as correct.
- Named entities and foreign terms are the main cause for incorrect processed or unprocessed tokens.

AntiMorfo

error	occurrences
not processed	**387**
root not in the dictionary	(21)
named entities	(34)
Spanish loans/borrowings	(127)
English terms	(7)
others	(198)
incorrectly segmented	**13**
diverging segmentation criteria	**76**
output needs adjustments	(9)
others	(67)
total	**476**

	AntiMorfo
Precision	79.35%
Recall	41.80%
F1	54.76%

- Unable to process many tokens – many for which no reason could be identified.
- Like with SQUOIA, named entities and foreign terms were the main cause identified for incorrecttly processed or unprocessed tokens.

Aymara Morph Analyzer

error	occurrences
not processed	**84**
root not in the dictionary	(43)
named entity	(25)
Spanish loans/borrowings	(9)
others	(7)
incorrect segmentation	**136**
root not in the dictionary	(89)
named entity	(11)
Spanish loans/borrowings	(7)
others	(29)
diverging segmentation criteria	**108**
total	**327**

	AymaraMorph
Precision	64.9%
Recall	57.85%
F1	61.17%

- Unlike for the other two tools, the main reason for incorrect processed or unprocessed tokens was that roots were not included in the internal dictionary.
- Some misspellings and unannotated tokens were found in the gold standard.

CONCLUSIONS (EXPERIMENT 1)

- Substantial disagreement between our gold standards (annotated data) and the output produced by the tools.
  - ➢ More than one possible correct analysis.
  - ➢ Even though one gold standard (Quechua Treebank) and one of the tools (SQUOIA) were developed by the same author.
  - ➢ SQUOIA could reach a F1 value as high as 95% if both applied the same segmentation criteria.
- AntiMorfo and Aymara Morph Analyzer performance need improvements.
  - ➢ AntiMorfo requires further investigation on problematic cases for which no cause has been identified.
  - ➢ According to the results found, expanding Aymara Morph Analyzer internal dictionary could naturally improve its performance.

RESULTS (EXPERIMENT 2)

SQUOIA/AntiMorfo

VARIETY	SQUOIA	SQUOIA with guesser	AntiMorfo	SQUOIA + AntiMorfo
QI-pe	53.98%	77.27%	40.53%	77.46%
QIIA-pe	51.47%	72.69%	42.44%	74.66%
QIIB-pe	58.41%	68.58%	50.09%	69.32%
QIIB-ec	54.04%	64.50%	45.76%	66.27%
QIIB-co	40.27%	53.59%	24.70%	55.39%
QIIC-pe1	94.66%	97.24%	84.16%	98.53%
QIIC-pe2	96.27%	99.07%	70.52%	99.25%
QIIC-bo	86.85%	94.82%	78.49%	96.61%
QIIC-cl	92.48%	97.88%	84.58%	99.42%
QIIC-ar	71.18%	79.50%	42.17%	81.62%

- As expected, both SQUOIA and AntiMorfo processed more tokens for QIIC languages, the group to which Qullaw/Cuzco Quechua also belongs.
  - ➢ However, SQUOIA was able to process more tokens from Chanka/Ayacucho (QIIC-pe2) than from Qullaw/Cuzco Quechua.
- Unexpectedly, SQUOIA processed more tokens from Chawpi/Central, a QI variety, than for varieties in the QII branch (QIIA and QIIB) to which Qullaw/Cuzco Quechua (a QIIC variety) also belongs.
  - ➢ No conclusions can be drawn from these figures, as no evaluation on correcteness has been performed.
  - ➢ The cause might be merely orthographic (refer to the Conclusions).
- Being written in less standardized varieties, the data in Inga (QIIB-co) and Argentinean (QIIC-ar) samples yield results that do not agree with other varieties in the same QIIB and QIIC branches, respectively.
- There is very little gain in combining both SQUOIA and AntiMorfo.

Aymara Morph Analyzer

VARIETY	AymaraMorph	AymaraMorph with guesser
aym-pe	44.51%	100%
aym-bo	53.53%	100%
aym-cl	55.38%	100%

- No noticeable difference noted for the different varieties.
- No unprocessed tokens with the guesser.
  - ➢ Possibly due to the use of more standardized forms.

CONCLUSIONS (EXPERIMENT 2)

- **AntiMorfo:** expected behaviour – more tokens processed for QIIC varieties.
- **SQUOIA:** despite being developed for Cuzco Quechua (QIIC branch), can process more tokens for Central Quechua (QI branch) than for QIIB and QIIC varieties.
  - ➢ Possibly due to phonetic/ortographic specificities in QIIB and QIIC.

English	QIIA-pe	QIIB-ec	QIIC-pe1	QI-pe
you	qam	kan	qam	qam
like that	shina	shina	hina	- (chaynaw)
new	mushuk	mushuk	musuq	mushuq
teacher	yaĉachikuk	yachachikuk	yachachikuq	yachachikuq

- **Aymara Morph Analyzer:** despite the expected linguistic differences, no noticeable difference observed for the different samples.
  - ➢ A similar standard is likely to be currently in use in Peru, Bolivia and Chile.