COMPUTATIONAL MORPHOLOGY

UNIVERSITY OF THE BASQUE COUNTRY

DEPARTMENT OF INFORMATICS

A PyFOMA Grammar for Quechua Verbal Morphology

Author:
Luis Antonio VASQUEZ REINA

Instructors:
Maite ORONOZ
Hans MULDEN

Contents

List of Figures

List of Tables

Abstract

This report presents a Grammar for Quechua verbal morphology written in PyFOMA. Specifically, for the verbal morphology the Ayacucho and Cuzco varieties of Quechua. It begins introducing the distribution and history of the Quechuan languages and peoples. Afterwards an overview of the Quechua grammar (phonology and morphology) and annotation schema (UniMorph) needed for this work is given. Finally, the implementation of the grammar in PyFOMA is explained. The results are discussed and the implementation is found to be successful.

1 A quick introduction to Quechua

1.1 Typology and distribution

Following Cerrón-Palomino (2003), Quechua is best understood as a family of closely related languages mainly spoken in the central Andes mountains. Quechua is the most spoken native language family of the Americas, with approximately 8.5 to 10 million speakers (Adelaar, 2004). It has two main branches, distributed across Colombia, Ecuador, Peru, Bolivia, Argentina, and small regions of Chile (see Figure 1). Quechua I languages are spoken in central Peru, while Quechua II languages are spoken in all other areas of Quechua. The Quechua II branch can be further divided into Quechua IIA, IIB, and IIC. Also known as Southern Quechua, Quechua IIC comprises the largest group of Quechua varieties, with approximately 6 to 7 million speakers (Cerrón-Palomino, 2003). The best documented Quechua varieties are all inside the Quechua IIC branch, specifically, Ayacucho Quechua, Cuzco Quechua, and Bolivian Quechua. This work will focus on Cuzco Quechua and Ayacucho Quechua.

1.2 A quick history of Quechuan people

This section provides a basic overview of Adelaar (2004) and Cerrón-Palomino (2003).

Like several ethnic groups of the Americas, there were no Quechua speaking groups with their own writing system. Therefore, the data on the status of Quechuan languages before the arrival of the Spanish is very sparse and has to be reconstructed from documents (including those from the colonial and republican periods and from modern sources), toponymy, and current field work. Though Quechua is most widely known as the language of the Inca empire, there were already different Quechuan languages even before the rise of the Kingdom of Cuzco (the precursor to the Incans). In fact, the Inca empire used a specific variety of Quechua for administration which coexisted with other Quechua varieties and non-Quechua languages like Aymara. Given the short life of the Inca empire (from the mid 15th century to mid 16th century), this administrative variety did not result in a standardised version of Quechua and had little impact on other varieties.

As a consequence of the arrival of the Conquistador Francisco Pizarro, epidemics, and the establishment of Spanish colonial rule, most subjugated native groups of western South America experienced a reduction of their populations. Silver mining and farming resulted in the highest mortality rates among all professions practiced by the native population. In the new colonies, Spanish became the official language and Catholicism the official religion. Catholic priests began documenting the languages they found and their writings constitute the earliest data we can find for several native languages of the Americas including Quechuan languages. With the

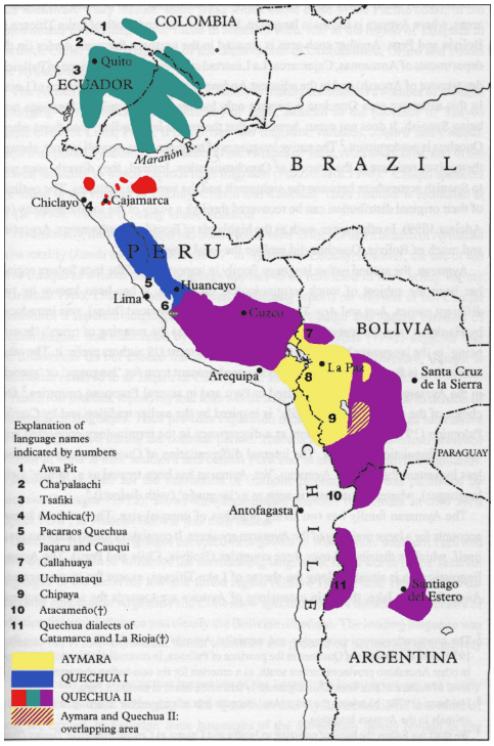

Figure 1: Branches of Quechua (Rios, 2010)

goal of evangelising the native population, only languages with many speakers and a large geographical distribution were well documented, frequently using the model by Antonio de Nebrija's Spanish Grammar. This model was heavily influenced by Latin grammars. The first grammar of Quechua was published in 1560 by Domingo de Santo Tomás, a Dominican priest from Sevilla. In the following centuries, the Spanish language expanded into the territory of the colonies. In some regions, Spanish became the only language spoken, though in general multilingualism became the norm for people of native origin. After the rebellion of Tupac Amaru II in Cuzco against the Bourbon Reforms in the mid 18th century, Spanish became the single language of Evangelising in all Spanish colonies. Afterwards in the Central Andes, Quechua lost all official recognition.

With the wave of independence movements at the beginning of the 19th century, several of the modern Republics in Central and South America were established. These movements were mainly driven by American-born Spanish descendants, who replaced colonial authorities after gaining independence. Thus, living conditions for native peoples did not improve in the new Republics and the discrimination against their cultures and languages continued. Spanish remained the only official language. In the 20th century, industrialization brought people from the countryside to the cities, where native languages had low social prestige. Modern authors reacted to this situation and produced modern literature in native languages. For Quechua, this can be reflected in the work of José María Arguedas:

> Yo no soy un aculturado; yo soy un peruano que orgullosamente, como un demonio feliz habla en cristiano y en indio, en español y en quechua.
> *I'm not a cultural upstart; I am a Peruvian who proudly, like a happy de-*

mon, speaks in Christian and Indian, in Spanish and in Quechua.

José María Arguedas, 1968

Nowadays, the majority of native speakers of Quechuan languages are urban and also speak Spanish natively. While biases, discrimination, and historical disadvantages against Quechua speakers continue to this day, there has been a revival of interest in learning and preserving Quechua in the last few decades (Zariquiey and Córdova, 2008).

2 A quick overview of Quechua grammar

As mentioned in Section 1, this work only deals with Cuzco Quechua and Ayacucho Quechua.

2.1 Phonology

Cuzco and Ayacucho Quechua are commonly considered to have three vowels:

	low central	unrounded front	unrounded front
phoneme	/a/	/i/	/u/
allophones	[a]	[i], [e]	[u], [o]

Table 1: Vowels of Cuzco Quechua and Ayacucho Quechua

Cuzco Quechua and Ayacucho Quechua have a very similar, but not identical, consonant inventory. For example, Cuzco Quechua has aspirated and ejective stops that are not present in Ayacucho Quechua. Table 2 shows the full inventory of consonants of both varieties in standardized writing and IPA, highlighting in grey consonants exclusive to Cuzco Quechua.

	Bilabial	Alveolar	Palatal	Velar	Uvular	Glottal
Plain stop	p	t	ch [$\widehat{\text{tʃ}}$]	k	q	
Aspirated stop	ph [p^h]	th [t^h]	chh [$\widehat{\text{tʃ}}^\text{h}$]	kh [k^h]	qh [q^h]	
Ejective stop	p’	t’	ch’ [$\widehat{\text{tʃ}}$’]	k’	q’	
Fricative		s				h
Nasal	m	n	ñ [ɲ]			
Lateral liquids		l	ll [ʎ]			
Vibrant liquids		r				
Semi-consonants	w		y [j]			

Table 2: Consonants of Cuzco Quechua and Ayacucho Quechua. Adapted from Zariquiey and Córdova (2008)

Words in Cuzco Quechua which include aspirated and ejective stops usually have a plain stop equivalent in Ayacucho Quechua (see Table 2.1). For this reason, syntactic software tools developed for Cuzco Quechua can be easily adapted to Ayacucho Quechua by replacing aspirated and ejective stops with their corresponding plain stop.

Cuzco Quechua	Ayacucho Quechua	Gloss
t'anta	tanta	bread
qhaway	qaway	to see
chhachu	chachu	filthy and hairy

2.2 Morphology

Quechuan languages are agglutinative, which means it adds suffixes to basic roots to form morphologically and semantically complex words. For each specific variety, the order of suffixes is highly fixed (Adelaar, 2004).

2.2.1 Pronouns

Quechua has seven pronouns depending on number and person. There are two pronouns for the first person plural: Inclusive (the speakers and the hearer) and Exclusive (the speakers without the hearer). Ayacucho and Cuzco Quechua use the following forms:

Pronoun	Person
ñuqa	1st Singuar
qam	2nd Singular
pay	3rd Singular
ñuqanchik	1st Plural Inclusive
ñuqayku	1st Plural Exclusive
qamkuna	2nd Plural
paykuna	3rd Plural

Table 3: Cuzco and Ayacucho Quechua Pronouns

2.2.2 Nominal Morphology

Nominal morphology in Quechua can easily become complex, especially when combinations of suffixes can have non-compositional pragmatic content. For this reason, only a quick exposition of nominal morphology is given in this report.

Consider the following nominal roots and suffixes in Tables 4 and 5.

root	gloss
miku	food
qullqi	money
carru	car
ñawi	eye

Table 4: Sample nominal roots

By combining these roots and suffixes, new words can be obtained, as seen in Table 6.

Adjectives appear before the nouns they modify. Consider the adjectives *hatun* (big) and *miski* (sweet, tasty), and how they modify words in Table 7:

suffix	gloss
-y	1st P. possessive
-kuna	plural
-wan	instrumental

Table 5: Sample nominal suffixes

word	translation
miku-y	my food
carru-y	my car
carru-kuna	cars
ñawi-y-kuna	my eyes
qullqi-y-wan	with my money
ñawi-y-kuna-wan	with my eyes

Table 6: Examples of nominal suffixation

compound	translation
miski miku	tasty food
hatun carru-y	my big car
hatun ñawi-y-kuna-wan	with my big eyes

Table 7: Examples of adjectives modifying nouns

2.2.3 Verbal Morphology

The verbal morphology of Quechuan languages is extremely regular. Cuzco Quechua and Ayacucho Quechua do not have irregular verbs. Consider the following infinitive verbs in Table 8, where the infinitive suffix is *-y*

Verbs can be nominalized via the agentive suffix *-q*, as in Table 9.

For verbal inflection, Cuzco Quechua and Ayacucho Quechua have two moods: indicative and imperative. For the indicative, there are four possible conjugations: Present, Experienced Past, non-experienced Past, and Future. For the imperative, there are two possible conjugations: Positive and Negative. Full conjugation tables can be found in the Appendix.

In the present tense, there are different suffixes for every person, as in Table 10. Note that this is one of the few cases where there are different suffixes for Ayacucho Quechua and Cuzco Quechua. For the second person plural present suffix, Ayacucho Quechua uses *-niku*, while Cuzco Quechua uses *-yku*.

Experienced past in Quechua is used when the speaker has first-hand evidence of a past event. It is formed by using the suffix *-rqa* before the conjugations for the Indicative Present:

- rima-rqa-nki : you spoke (I know/saw)
- miku-rqa-nchik: we (inclusive) ate (I now/saw)

On the other hand, non-experienced past is used when the speaker has obtained information from a different source (rather than first-hand). It is formed by using the suffix *-sqa* before the conjugations for the Indicative Present:

- rima-sqa-nki : you spoke (I hear. People say)
- miku-sqa-nchik: we (inclusive) ate (I hear. People say)

verb	gloss
miku-y	to eat
rima-y	to speak
muna-y	to want

Table 8: Sample verbs in Quechua

verb	gloss
miku-q	eater, one who eats
rima-q	speaker, one who speaks
muna-q	one who wants

Table 9: Some agentives in Quechua

Suffix	POS	example	translation
-ni	1st P. Sing	rima-ni	I speak
-nki	2nd P. Sing	munan-nki	You (singular) want
-n	3rd P. Sing	muna-n	he/she/it wants
-nchik	1st P. Pl. Incl.	rima-nchik	We (inclusive) speak
-yku *(Cuzco)*	1st P. Pl. Excl.	miku-yku	We (exclusive) eat
-niku *(Ayacucho)*	1st P. Pl. Excl.	rima-niku	We (exclusive) speak
-nkichik	2nd P. Pl.	muna-nkichik	You (plural) want
-nku	3rd P. Pl.	miku-nku	They eat

Table 10: Indicative Present suffixes

The future tense, the positive and negative imperative each have their own suffixes which can be found in the Appendix. Note that the negative imperative uses the prohibition particle *ama* which is always placed immediately before the verbal form. For the purposes of this work, the negative imperative is analysed as using circumfixes whose left component is always *ama-*, as shown in Table 19.

lemma	inflected forms	tag set
rimay	rimachunku	V;IMP;POS;3;PL
rimay	rimanku	V;PRS;3;PL

Table 11: Sample entries for Cuzco Quechua from SIGMORPHON 2017

2.3 Out of scope

Quechuan syntax as well as other parts of its morphology are not considered to lie in the scope of this paper, the curious reader is referred to Cerrón-Palomino (2003) (in Spanish) and Adelaar (2004) (in English).

Several suffix categories have not been mentioned in this work:

- Nominal case suffixes
- The zero suffix
- Evidential suffixes
- Reflexive verbal suffixes
- etc...

Quechua syntax shows a preference for a subject, object, verb word order. However, it is not uncommon to find almost free word order. Quechua is pragmatically very rich, especially when it comes to subordination and coordination.

3 Unimorph and SIGMORPHON

The Unimorph Schema (Sylak-Glassman, 2016) proposes a set of features (tags) for morphological analysis of any language with a cross-linguistic theoretical basis, allowing one to develop and compare multilingual morphological tools.

In the 2017 edition of the SIGMORPHON shared task, Cuzco Quechua data was scraped from Wiktionary and used to train morphological analyzers (Cotterell et al., 2017). This data is available in their Github[1] and contains 144 verbal inflections.

For Quechua verbal morphology, the SIGMORPHON dataset uses the features in Table 12. This dataset will be used to evaluate a finite state transducer which will be explained in the next section. The tag sets for all verbal inflections can be found in the Appendix.

4 PyFOMA for Quechua Morphology

Finite state methods have already previously been used to deal with Quechua grammar, including a method described in Rios (2010). However, this implementation varies from the one described in this report in several ways.

The work in Rios (2010):

- analyzes all morphology (nominal, verbal, etc.) for several Quechua dialects, while this paper only examines the verbal morphology in Cuzco and Ayacucho Quechua.

[1] https://github.com/sigmorphon/conll2017/tree/master/all/task1

Feature	Explanation
1	First person
1+EXCL	Exclusive first person
1+INCL	Inclusive first person
2	Second person
3	Third person
AGT	Agentive
FH	Firsthand
NFH	Non firsthand
SG	Singular
PL	Plural
PRS	Present
PST	Past
FUT	Future
IMP	Imperative
NEG	Negative
POS	Positive
NFIN	Non-finite

Table 12: Quechua verbal features in SIGMORPHON 2017

- relies on several software dependencies and an intricate installation process, while this project makes use of PyFOMA, which has a simpler installation and syntax.
- uses language-specific tags, while this project uses cross-linguistic tags.

Next is a short description of the creation and implementation of the PyFOMA grammar used in this project.

1 After downloading the data for Cuzco Quechua in a CSV file from GitHub, the data is first lemmatized.

2 For each tag set, a dictionary of suffixes and circumfixes is defined.

```
lemmas # {mikuy, rimay, munay, ...}
quechua_suffixes = {
...
'V;NFIN;AGT':       'q',
'V;PRS;1;SG':       'ni',
'V;PRS;2;SG':       'nki',
'V;PRS;3;SG':       'n',
...
}
quechua_circumfixes = {
'V;IMP;NEG;2;PL': ('ama', 'ychikchu'),
'V;IMP;NEG;2;SG': ('ama', 'ychu'),
'V;IMP;NEG;3;PL': ('ama', 'chunkuchu'),
'V;IMP;NEG;3;SG': ('ama', 'chunchu'),
}
tags # ['V;NFIN;AGT', 'V;PRS;1;SG', ...]
```

3 The next step is constructing the lexicon for the FST

4 To deal with circumfixes, a special symbol `[Prefix]` is added, which will be rewritten for the tag sets with circumfixes, and deleted for the rest.

5 All tag sets are added at the end of each word.

```
fsts = {}
fsts["S"] = [ ("'[Prefix]'", "TaggedVerb") ]
fsts["TaggedVerb"] = [ (lemma, "Tag") for lemma in lemmas]
fsts['Tag'] = [(tag, "#") for tag in tags]
fsts['lexicon'] = FST.rlg(fsts, "S")
print(Paradigm(fsts['lexicon'], ".*"))
```

Which outputs:

```
achalachiy      [Prefix][V;FUT;2;PL]      [Prefix]achalachiy[V;FUT;2;PL]
achalachiy      [Prefix][V;FUT;3;PL]      [Prefix]achalachiy[V;FUT;3;PL]
achalachiy      [Prefix][V;FUT;1;SG]      [Prefix]achalachiy[V;FUT;1;SG]
....
```

6 All the tag sets for suffixes are replaced. Taking advantage of the fact that all verbal lemmas end in *-y*, instances of `-y[SYMBOL]` are replaced by their corresponding suffix.

```
rewrite_suffix_rules = ... # definition by list comprehension
rewrite_suffix_rules
#    (y'[V;FUT;2;PL]'):(nkichik)
# |  (y'[V;FUT;2;SG]'):(nki)
# |  (y'[V;FUT;3;PL]'):(nqaku)
# |  (y'[V;FUT;3;SG]'):(nqa)
# ...

fsts['suffix_rules'] = FST.re(
    "$^rewrite("+rewrite_suffix_rules+")"
    )
fsts['with_suffixes'] = FST.re(
    "$lexicon @ $suffix_rules",
    fsts
    )
print(Paradigm(fsts['with_suffixes'], ".*"))
```

Which has the output below. Note that tag sets corresponding to circumfixes have not been replaced.

```
achalachiy      [Prefix][V;FUT;2;PL]    [Prefix]achalachinkichik
achalachiy      [Prefix][V;FUT;2;SG]    [Prefix]achalachinki
achalachiy      [Prefix][V;FUT;3;PL]    [Prefix]achalachinqaku
...
achalachiy      [Prefix][V;IMP;NEG;2;PL]  [Prefix]achalachiy[V;IMP;NEG;2;PL]
...
```

7 When there is a tag set for circumfixes at the end of the word, the prefix symbol is replaced by *ama*.

```
# quechua alphabet
fsts["A"] = FST.re("[ñ\'a-z]")

circumfix_tags # "('[V;IMP;NEG;2;PL]'
#|'[V;IMP;NEG;2;SG]'
#|'[V;IMP;NEG;3;PL]'
#|'[V;IMP;NEG;3;SG]'
#)"

fsts['rewrite_prefix_symbol'] = FST.re(
"$^rewrite( ('[Prefix]'):(ama) / _ ( $A+ "+ circumfix_tags+ " ) )",
fsts
)
```

8 Now, as previously done with the prefixes, the tag sets for circumfixes are replaced with their corresponding endings.

9 Then, all instances of the prefix symbol are erased. Note that the prefix symbol was only left on words that do **not** come from circumfix tag sets.

10 The final result is a grammar for all Quechua verbs.

```
rewrite_circumfix_rules
# " (y'[V;IMP;NEG;2;PL]'):(ychikchu)
# | (y'[V;IMP;NEG;2;SG]'):(ychu)
# | (y'[V;IMP;NEG;3;PL]'):(chunkuchu)
# | (y'[V;IMP;NEG;3;SG]'):(chunchu)"
fsts['circumfix_rules'] = FST.re(
    "$^rewrite("+rewrite_circumfix_rules+")"
)

fsts['delete_prefix_symbol'] = FST.re(
    "$^rewrite('[Prefix]':'')"
    )

fsts['quechua_grammar'] = FST.re(
    """
    $with_suffixes
    @ $rewrite_prefix_symbol
    @ $circumfix_rules
    @ $delete_prefix_symbol
    """,
    fsts
)

print(Paradigm(fsts['quechua_grammar'], ".*"))
```

Which outputs:

```
achalachiy      [Prefix][V;FUT;2;SG]          achalachinki
achalachiy      [Prefix][V;FUT;3;PL]          achalachinqaku
achalachiy      [Prefix][V;FUT;3;SG]          achalachinqa
achalachiy      [Prefix][V;IMP;NEG;2;PL]      amaachalachiychikchu
...
```

5 Evaluation and Conclusion

After writing the grammar for Quechua verb Morphology, the dataset from Unimorph is used for evaluation.

The evaluation method used is relatively straightforward: feeding each lemma and tag set from the dataset into the grammar and comparing the generated form to the correct inflection.

- The results are extremely positive: 99.3% of the inflections are correctly generated by the grammar.
- There is only one inflection wrongly generated. The correct inflection for the verb *wichq'ay* in the 3rd person present singular is *wich'qan*, while the generated form is *wichq'an*.
- One possible explanation may be that this is the result of a phonological rule involving the distribution of ejective stop constants.
- In the literature referenced, the exact formulation of this phonological rule was not found. Since *wichq'ay* is the only wrongly generated inflection, it was left as-is.

Overall, the results demonstrate the efficacy of the FST grammar created for the purposes of this project.

6 Appendix

This appendix includes all verbal suffixes for Cuzco Quechua and Ayacucho Quechua (Zariquiey and Córdova, 2008).

6.1 Non finite suffixes

Suffix	POS	UniMorph tag
-y	Infinitive	NFIN
-q	Agentive	NFIN;AGT

Table 13: Non-finite Quechua suffixes

6.2 Indicative Present suffixes

Suffix	POS	UniMorph tag
-ni	1st P. Sing	PRS;1;SG
-nki	2nd P. Sing	PRS;2;SG
-n	3rd P. Sing	PRS;3;SG
-nchik	1st P. Pl. Incl.	PRS;1+INCL;PL
-yku (*Cuzco*)	1st P. Pl. Excl.	PRS;1+EXCL;PL
-niku (*Ayacucho*)	1st P. Pl. Excl.	PRS;1+EXCL;PL
-nkichik	2nd P. Pl.	PRS;2;PL
-nku	3rd P. Pl.	PRS;3;PL

Table 14: Indicative Present suffixes

6.3 Indicative Experienced Past suffixes

Suffix	POS	UniMorph tag
-rqa-ni	1st P. Sing	PST;FH;1;SG
-rqa-nki	2nd P. Sing	PST;FH;2;SG
-rqa-n	3rd P. Sing	PST;FH;3;SG
-rqa-nchik	1st P. Pl. Incl.	PST;FH;1+INCL;PL
-rqa-niku	1st P. Pl. Excl.	PST;FH;1+EXCL;PL
-rqa-nkichik	2nd P. Pl.	PST;FH;2;PL
-rqa-nku	3rd P. Pl.	PST;FH;3;PL

Table 15: Indicative Experienced Past suffixes

6.4 Indicative Non-experienced Past suffixes

Suffix	POS	UniMorph tag
-sqa-ni	1st P. Sing	PST;NFH;1;SG
-sqa-nki	2nd P. Sing	PST;NFH;2;SG
-sqa-n	3rd P. Sing	PST;NFH;3;SG
-sqa-nchik	1st P. Pl. Incl.	PST;NFH;1+INCL;PL
-sqa-niku	1st P. Pl. Excl.	PST;NFH;1+EXCL;PL
-sqa-nkichik	2nd P. Pl.	PST;NFH;2;PL
-sqa-nku	3rd P. Pl.	PST;NFH;3;PL

Table 16: Indicative Non-experienced Past suffixes

6.5 Indicative Future suffixes

Suffix	POS	UniMorph tag
-saq	1st P. Sing	FUT;1;SG
-nki	2nd P. Sing	FUT;2;SG
-nqa	3rd P. Sing	FUT;3;SG
-sunchik	1st P. Pl. Incl.	FUT;1+INCL;PL
-saqku	1st P. Pl. Excl.	FUT;1+EXCL;PL
-nkichik	2nd P. Pl.	FUT;2;PL
-nqaku	3rd P. Pl.	FUT;3;PL

Table 17: Indicative Future Quechua suffixes

6.6 Positive Imperative suffixes

Suffix	POS	UniMorph tag
-y	2n P. Sing	IMP;POS;2;SG
-ychik	2nd P. Pl.	IMP;POS;2;PL
-chun	3rd P. Sing	IMP;POS;3;SG
-chunku	3rd P. Pl.	IMP;POS;3;PL

Table 18: Imperative Quechua suffixes

6.7 Negative Imperative circumfixes

Suffix	POS	UniMorph tag
ama- -y-chu	2n P. Sing	IMP;NEG;2;SG
ama- -ychik-chu	2nd P. Pl.	IMP;NEG;2;PL
ama- -chun-chu	3rd P. Sing	IMP;NEG;3;SG
ama- -chunku-chu	3rd P. Pl.	IMP;NEG;3;PL

Table 19: Negative Imperative Quechua circumfixes

References

Adelaar, Willem FH (2004). *The languages of the Andes*. Cambridge University Press.

Cerrón-Palomino, Rodolfo (2003). *Lingüística quechua 2da edición*. Centro de estudios rurales andinos Bartolomé de Las Casas.

Cotterell, Ryan et al. (2017). "CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages". In: *arXiv preprint arXiv:1706.09031*.

Rios, Annette (2010). "Applying finite-state techniques to a native American language: Quechua". In: *Institut für Computerlinguistik, Universität Zürich*.

Sylak-Glassman, John (2016). "The composition and use of the universal morphological feature schema (unimorph schema)". In: *Johns Hopkins University*.

Zariquiey, Roberto and Gavina Córdova (2008). *Qayna, kunan, paqarin. Una introducción práctica al quechua chanca*. Pontificia Universidad Católica del Perú. Estudios Generales Letras.