MORPHOLOGICAL AND SYNTACTIC GRAMMARS FOR RECOGNITION OF VERBAL LEMMAS IN QUECHUA

Maximiliano Duran

Abstract

This article presents the process of using the inflectional and derivational structures of Quechua verbs to recognize verbal forms in a corpus. With the aid of morphological and syntactic NooJ grammars, we show how to retrieve and to extract the hidden verbs.

Introduction

Existing Quechua dictionaries contain less than 1500 verbs and yet ancient Quechua writings do contain many unknown verbs. They appear in inflected forms. My motivation was to isolate these verbal lemmas to enhance the verb lexicon. First, I describe briefly how I formalized the corpus which includes some ancient documents. Then, I present the set of morphological and syntactic grammars that I constructed with NooJ. These grammars will serve to analyze the corpus for searching verbal forms. Once these forms are identified I apply an algorithm of NooJ operators to extract and list the verbs. We have identified near three hundred unknown verbs.

Motivation for the project

The Quechua language was the official language of the Inca civilization. It originated in the central Andes of Peru around the first half of the first millennium of the present era. In 2009 UNESCO declared it a language in danger. We would like to contribute to its survival and its development.

Morphological and syntactic grammars for recognition of verbal lemmas in Quechua

Our long term project is to build a linguistic resources platform for automatic text processing of Quechua.

The first step is to build a French-Quechua electronic dictionary based on the 25 thousand French verbs of Dubois & Dubois[1].

The Corpus

The following documents from the begining of the XVI^{h} century contain a total of 67900 tokens:

- Gonçalez Holguin, Diego, 1608, *Vocabulario de la Lengua General de todo el Perú llamada Lengua Qquichua o del Inca*.
- Santo Thomas, Domingo de, 1560, *Lexicon, o vocabulario de la lengua general del Peru*.
- Francisco de AVILA's, 1598? *Dioses y hombres de Huarochiri*. A Quechua narrative gathered by Francisco de Avila

I have first standardized the orthography of these texts using the Ayacucho's Quechua alphabet.

Inflected verbal forms

A typical Quechua inflected verbal form has the following structure: V+ IPS + PR ENDING + PPS or <V><IPS><PR ENDING><PPS>

where: V: verb lemma

IPS[2] : Interposed suffix is a set of 31 suffixes, and

PPS[3] : Post-posed suffix containing 19 suffixes.

PR ENDING[4]: is the set of seven present tense endings (which behave as fixed points during the inflections).

[1] Dubois & Dubois 2007

[2] IPS =(chi, chka, ikacha, ikachi, ikamu, ikapu, ikari, iku, isi, kacha, kamu, kapu, ku, lla, mpu, mu, na, naya, pa, paya, pti, pu, ra, raya, ri, rpari, rqa, rqu, ru, spa, sqa, tamu, wa)

[3] PPS=(ch, chaa, chiki, chui, chun, chusina, maa, man, mm, mmi, ña, pas, puni, qa, raq, ssi, sis, taq, yaa)

We remark that Quechua is a polysynthetic language. For example the English sentence:

"We have to do the work leaving aside everything else" becomes *llamkananchikraqmi*

A whole sentence in English represents a single verbal form in Quechua.

Let us see the behavior of some of these suffixes and their combinations in the following inflections of the verb *qallariy* "to begin":

verb lemma : qallari-

*qallari-**nchik*** we begin

-nchik is the ending of PR p +1

*qallari-chka-**nchik*** we are beginning

*qallari-isi-chkak-**nchik*** we are helping someone to begin

*qallari-isi-chka-**nchik**-ña* we are already helping someone to begin

*qallari-isi-chka-**nchik**-ña- taq* and yet we are already helping him to begin

The personal ending ***nchik*** remains fixed at the end of the IPS combinations or before the PPS combinations.

The morphology of Quechua is very much dominated by this kind of agglutination of suffixes placed after a verbal, nom, adverb or adjective lemma.

Matrix approach to verbal suffix combinatorial

What we call the present tense is in fact an indefinite present. On the one hand it places the statement at the moment in which this statement takes place, but on the other hand it may also place it in a moment in which the statement has just taken place and is still not completed.

The conjugation for the three singular persons has the following structure:

ñoqa (I) lemma +***NI***

[4] PR ENDING= (-ni , -nki , -n , -nchik , -niku , -nkichik, nku)

qam (you) lemma + ***NKI***

pay (he, she) lemma +***N***

For the future we have the scheme:

ñoqa (I) lemma +***SAQ***

qam (you) lemma + ***NKI***

pay, (he, she) root +***NQA***

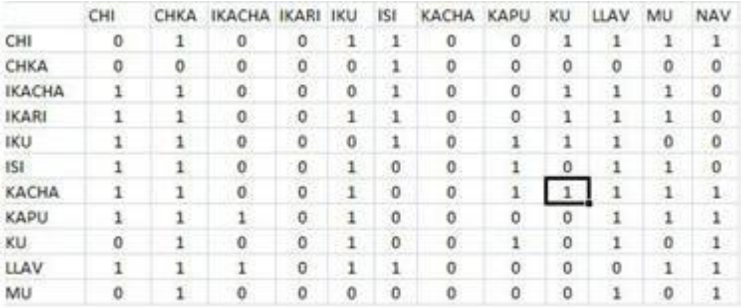
The present tense form plays a crucial role in the conjugated Quechua verbal form of the other tenses. It is a kind of fixed point around which all the inflectional topology based on the combinatorial of the suffixes is constructed (tenses, modes, aspects, etc.)

For instance, the past preterit is obtained by taking this present structure and interposing the IPS suffix -*rqa*-, between the verbal lemma and the ending of the person. We have then:

Present		Past preterit	
taki-ni	I sing	*taki-**rqa**-ni*	I sang
taki-nki	you sing	*taki- **rqa**-nki*	you sang
taki-n	he sings	*taki- **rqa**-n*	he sang

According to the Quechua verb morphology, we can build combinations of 2, 3, or 4 of IPS and PPS suffixes which are very productive inflection wise.

To obtain the complete set of these combinations that are syntactically correct, we first constructed manually a two-entry matrix having as the first row and the first column all the 33 inter-positional suffixes IPS in one case and the set of the PPS ones in the other case. We then filled the 1089 cells with 0 or 1 for the IPS's and 289 cells for the PPS's. The value "1" means "grammatically valid combination" and "0" means "not valid". For instance the cell corresponding to the point (*chi*, *chka*) as coordinates in this matrix bears "1", because the combination -*chichka* is compatible and may be agglutinated to the root *taki* of the verb "to sing" *takiy* to get the verbal form *taki-chichka-ni* "I am making him sing", or for the cell (*kacha, ku*), which bears "1" also, we'll have the combination -*kachaku*, *taki-kachaku-ni* "I keep singing once and over again". We have found for the time being 295 "1"'s for the IPS's case.



	CHI	CHKA	IKACHA	IKARI	IKU	ISI	KACHA	KAPU	KU	LLAV	MU	NAV
CHI	0	1	0	0	1	1	0	0	1	1	1	1
CHKA	0	0	0	0	0	1	0	0	0	0	0	0
IKACHA	1	1	0	0	0	1	0	0	1	1	1	0
IKARI	1	1	0	0	1	1	0	0	1	1	1	0
IKU	1	1	0	0	0	1	0	1	1	1	0	0
ISI	1	1	0	0	1	0	0	1	0	1	1	0
KACHA	1	1	0	0	1	0	0	1	1	1	1	1
KAPU	1	1	1	0	1	0	0	0	0	1	1	1
KU	0	1	0	0	1	0	0	1	0	1	0	1
LLAV	1	1	1	0	1	1	0	0	0	0	1	1
MU	0	1	0	0	0	0	0	0	0	1	0	1

Fig. 1 Matrix of bi-dimensional combinations of interposition suffixes

We then obtained the valid combinations of three IPS's. The corresponding matrix is one that has as the first row the set of 33 IPS's and as the first column the 295 valid binary combinations that we have just obtained. We found 57 valid or attested three-fold agglutinations. Here are some examples:

-ñachusinam -ñapaschá -ñapaschik -ñapaschu -ñapasmi -ñataqsi -punichusinam -puniñach -puniñachá - puniñachik -puniñachu? -puniñachu -puniñachusina -puniñamá -puniñam -puniñapas -puniñas

Postposition suffixes PPS

They are placed after the verbal ending, as in the following examples:

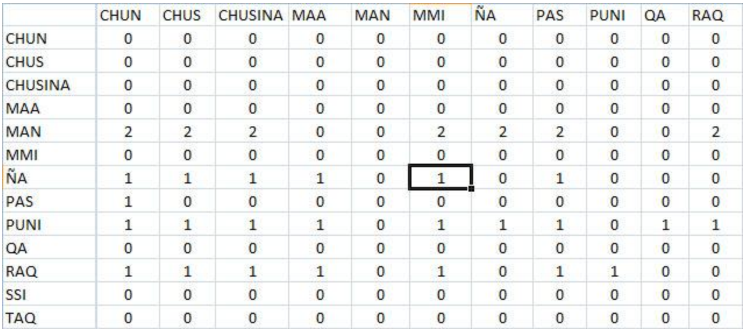
*rima-**nki** -man* you should talk

*rima-**nki** -man-pas* besides, you should talk

*rima-**nki** -man-pas-cha* you should perhaps talk

*rima-**n** - man-ña-taq* I fear that he speaks up

The binary PPS combinations matrix contains 56 compatible agglutinations as shown in Fig. 2.

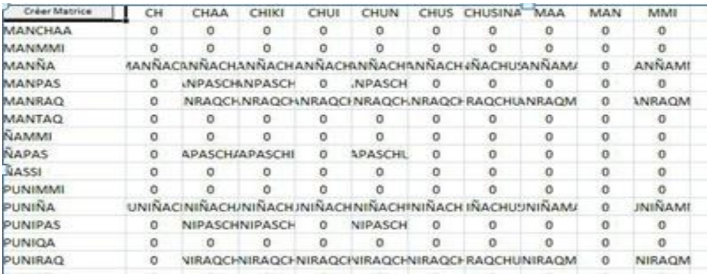


	CHUN	CHUS	CHUSINA	MAA	MAN	MMI	ÑA	PAS	PUNI	QA	RAQ
CHUN	0	0	0	0	0	0	0	0	0	0	0
CHUS	0	0	0	0	0	0	0	0	0	0	0
CHUSINA	0	0	0	0	0	0	0	0	0	0	0
MAA	0	0	0	0	0	0	0	0	0	0	0
MAN	2	2	2	0	0	2	2	2	0	0	2
MMI	0	0	0	0	0	0	0	0	0	0	0
ÑA	1	1	1	1	0	1	0	1	0	0	0
PAS	1	0	0	0	0	0	0	0	0	0	0
PUNI	1	1	1	1	0	1	1	1	0	1	1
QA	0	0	0	0	0	0	0	0	0	0	0
RAQ	1	1	1	1	0	1	0	1	1	0	0
SSI	0	0	0	0	0	0	0	0	0	0	0
TAQ	0	0	0	0	0	0	0	0	0	0	0

Fig. 2 Compatibility binary matrix of post positioned suffixes

Here too, the “1” corresponding to the point (*ña, mmi*) indicates that the combination –*ñammi* is grammatically valid, thus we have :

rima-nchik-ña-m, which we have already mentioned (*m* alone if it follows a vowel). (The “2” stands for the modified PRM2 of the PR structure).



Créer Matrice	CH	CHAA	CHIKI	CHUI	CHUN	CHUS	CHUSINA	MAA	MAN	MMI
MANCHAA	0	0	0	0	0	0	0	0	0	0
MANMMI	0	0	0	0	0	0	0	0	0	0
MANÑA	MANÑAC	ANÑACH	ANÑACH	ANÑACH	ANÑACH	ANÑACH	ÑACHUS	ANÑAMA	0	ANÑAMI
MANPAS	0	NPASCH	NPASCH	0	NPASCH	0	0	0	0	0
MANRAQ	0	NRAQCH	NRAQCH	NRAQC	NRAQCH	NRAQC	RAQCHU	NRAQM	0	NRAQM
MANTAQ	0	0	0	0	0	0	0	0	0	0
ÑAMMI	0	0	0	0	0	0	0	0	0	0
ÑAPAS	0	APASCH	APASCHI	0	APASCHI	0	0	0	0	0
ÑASSI	0	0	0	0	0	0	0	0	0	0
PUNIMMI	0	0	0	0	0	0	0	0	0	0
PUNIÑA	UNIÑAC	NIÑACH	NIÑACH	NIÑACH	NIÑACH	NIÑACH	IÑACHU	NIÑAMA	0	NIÑAMI
PUNIPAS	0	NIPASCH	NIPASCH	0	NIPASCH	0	0	0	0	0
PUNIQA	0	0	0	0	0	0	0	0	0	0
PUNIRAQ	0	NIRAQCH	NIRAQC	NIRAQC	NIRAQCH	NIRAQC	RAQCHU	NIRAQM	0	NIRAQM

Fig. 3. Partial view of the matrix of compatible tertiary combinations of post-positioned suffixes.

Similarly, we have obtained the matrix of tertiary combinations having the 56 compatible binary combinations as the first column and the vector PPS as the first row. The result contains only 80 non null elements, as shown in Fig. 3.

Following the same method, we obtain the matrix MPPS_{3x1} of compatible combinations of four post- positioned suffixes. Some of the resulting valid combinations are listed below:

-manñapaschá, -puniraqpaschá, -puniraqpaschiki, puniraqpaschu -raqpuniñachu, -raqpuniñachus

-raqpuniñachusina, -manñapaschiki, -manñapaschun

We are working on the matrix of combinations of five and six PPS.

The Quechua morpho-syntax rules allow the mixing of both cases to obtain a large number of inflectional forms, of verbal forms with mixed agglutinations of inter and post-positioned suffixes as in the examples:

*rima-ri-**nki**-man* you should perhaps talk

*rima-ri-lla-**nki**-man-raq* I think you should before, etc.

Here again we see that the endings behave as stable fixed points.

Programming the corresponding NooJ grammars

We applied these results to program the corresponding paradigms Using similar criteria as K. Bogacki[5] we choose NooJ inflection descriptions rather than the graphical approach to describe the grammars. . Here are some examples:

conjugVERBES = <E>/INF|:CHU |:progCHU |:pasCHU |:futCHU |:impeCHU |:iptiiCHU |:imanCHU |:nominITA|:GSTA;

VERBEAY = <E>/INF|:CHU |:PRESENT |:FUT |:RQA |:PREASS |:CHKAASS |:IMP |:COND |:PPL |:PTIC |:presenCHU |:progCHU|:impeCHU |:FUTCHU |:iptiiCHU |:imanCHU |:sqaCHU |:ptiiqaCHU |:GDYN |:STINCHU |:TA |:SQAIKI |:TRTS1 |:WANCHU |:TRDE1 |:DE1PCHU |:TRDE3 |:DE3PCHU |:TRTA2 |:accustfTA |:WANKICHU |:TRDE1CHU |:TRTS1CHU |:PIDF2 |:PIDF2CHU |:PIDF2 |:PICTR |:PICTRAC |:SPA |:GSPA |:IPI;

Using the verb dictionary, NooJ will generate the corresponding verbal forms. Below we have a small list of entries for the verb *takiy* “to sing”.

takinichu,takiy,V+FR="chanter"+FLX=conjugVERBES+s+1+NEG

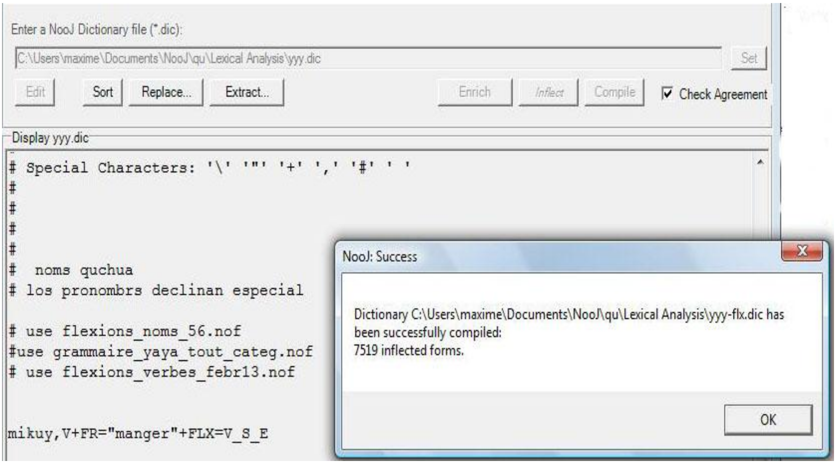
takichka**ni**kuchu,takiy,V+FR="chanter"+FLX=conjugVERBES+pex+1 +NEG

[5] K. Bogacki, 2008

takichka**n**kuchu,takiy,V+FR="chanter"+FLX=conjugVERBES+p+3+NEG

We have programmed more than 200 paradigms up to now. In the future, we will complete the study for the cases of combinations of more than three suffixes.

When we apply the program to a transitive verb like *mikuy* to eat, we obtain more than 7500 inflected forms as shown in Fig. 4.


Enter a NooJ Dictionary file (*.dic):

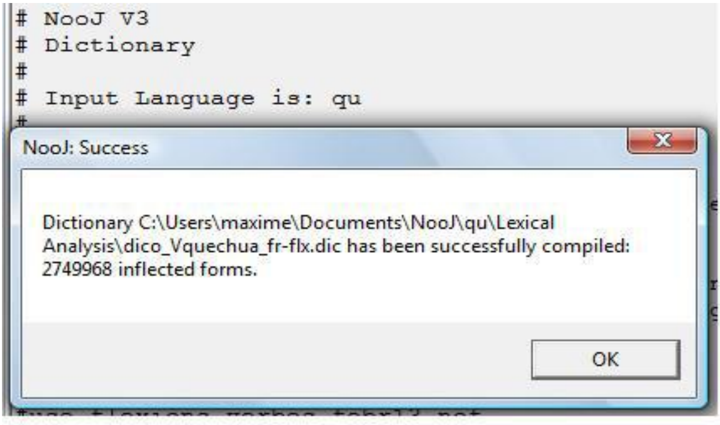
C:\Users\maxime\Documents\NooJ\qu\Lexical Analysis\yyy.dic

```
# Special Characters: '\' '"' '+' ',' '#' ' '
#
#
#
#
#  noms quchua
# los pronombrs declinan especial

# use flexions_noms_56.nof
#use grammaire_yaya_tout_categ.nof
# use flexions_verbes_febr13.nof


mikuy,V+FR="manger"+FLX=V_S_E
```

NooJ: Success

Dictionary C:\Users\maxime\Documents\NooJ\qu\Lexical Analysis\yyy-flx.dic has been successfully compiled:
7519 inflected forms.

Fig. 4 result of the flections of the verb *mikuy*


```
# NooJ V3
# Dictionary
#
# Input Language is: qu
#
```

NooJ: Success

Dictionary C:\Users\maxime\Documents\NooJ\qu\Lexical Analysis\dico_Vquechua_fr-flx.dic has been successfully compiled:
2749968 inflected forms.

Fig 5 Inflected forms for 400 transitive verbs

The same program, applied to a set of 400 transitive Quechua verbs generates 2,749,968 inflected forms as is shown I Fig. 5.

Recognition, extraction and recovery of lost verbs

We have applied several queries of concordances on our corpus using operators like

NI_q_extr == Find/ Replace (PERL pattern, ni$ | q$, extract lines)

VOC-ni_q == NI_q_extr (VOC-H_brut)

VOC-chay_rayay_nayay == *Find/ Replace (PERL pattern, chay$*|, *rayay$*| *nayay$*|, *extract lines)*

which gave us all the verbal forms containing potential verbal lemmas (5.541 forms).

From which we can extract the verbal lemmas hidden in this set by applying some algorithms using NooJ operators. We have got a list of 298 verbs considered “lost” verbs. The gathering of this catalog of new verbs is an important step for the lexical preservation for the language. We present a sample of the obtained unknown verbs:

qalluykuy	to cheat;
aknay	to exhibit;
rampay	to guide a blind;
tullpuy	to dye;
utiy	to become mad;
takuriy	to revolutionize;
tokapuy	to decorate.

Conclusion

We have conducted a comprehensive study of the way how inter- and post-positioned suffixes combine to generate thousands of inflections out of a single verb. Our corpus contains many complex inflected verbs. The grammar paradigms that we programmed in NooJ served us to identify among these inflections all the verbal forms in our corpus, which helped us to obtain 298 lost “new” verbs for our verb lexicon.

References

BOGACKI, Krzysztof. (2008) : *Polish module for NooJ*. In the Procedings of the 2007 Nooj Conference. Autonomus University of Barcelona. Cambridge Scholars Publishing. Newcastle.

DUBOIS, Jean et DUBOIS-Charlier, Françoise (D&D). (2007) : *Le verbes français (le «dictionnaire électronique des verbes français* (DEV), l992 Diffusé à partir de septembre 2007 par MoDyCo dans un format Excel.

GONÇALEZ Holguin, Diego. (1608): *Vocabulario de la Lengua General de todo el Perú llamada Lengua Qquichua o del Inca*. Edición y Prólogo de Raúl Porras Barrenechea. Lima, Universidad Nacional Mayor de San Marcos 1952.

SANTO THOMAS, Domingo de. (1560), *Lexicon, o vocabulario de la lengua general del Peru*. Valladolid: Francisco Fernandez de Cordova.

FRANCISCO DE AVILA, (1598?): *Dioses y hombres de Huarochiri. Narracion Quechua recogida por Francisco de Avila Traduccion J. M. Arguedas*. Lima. Peru 1966. Edicion bilingüe facsimilar 2012.

DURAN, M. (2009): *Diccionario Quechua-Castellano*. Éditions HC. Paris.

ITIER, (2001) : César. *Dictionnaire Quechua-Français*, Paris. L’Asiathèque. Paris.

PERROUD, Pedro Clemente. (1970): *Diccionario Castellano - Kechwa Dialecto de Ayacucho*. Lima . Edición.

SILBERZTEIN, M. (2003): *NooJ Manual*. htpp://www.nooj4nlp.net (220 pages updated regularly).