Automatic retrieving of derived Quechua verbs

Abstract In this paper, I present a method that helps to automatically retrieve derived and inflected Quechua verbs embedded in a corpus. It consists of the formal construction of paradigms and grammars managing the grammatical combinations of verbal suffixes. I use these formalized grammars to extract all the verb forms in a corpus. Then, using the Electronic Dictionary of Quechua verbs that I have previously constructed, and with the help of some elementary statistical tools, I have developed a practical system that allows answering the following socio-linguistic questions: a. how frequent is the use of multi-suffixation of verbs (derived by a combination of two or more suffixes) in everyday life communication? b. Is there a notable difference in this phenomenon between written and oral communication?

Keywords: Quechua verb derivation, Quechua morphology, verbal suffixes

1 Introduction

Quechua is the Incas language, spoken by over 6 million people in Peru, Equator, Bolivia, and Argentina. This language has an important characteristic: its verbal morphology is rich and highly regular. The verbs contained in a corpus will generally appear as derived or inflected forms. If we want to retrieve and extract all the verbs in a corpus, we need to know how the language morphology generates derivations and inflections. In this paper, I present the formalization of such linguistic phenomena. I show how I constructed several formal grammars using NooJ[1] platform, developed by Silberztein (2010, 2016). With the help of these grammars and the French-Quechua Electronic Dictionary of verbs, we have written by translating into Quechua the dictionary of Dubois & Dubois-Charlier (1997, 2007). *Les verbes français* (LVF). NooJ's I have written several formal grammars which help us automatically recognize any occurrence of a verbal inflected or derived form and analyze it. Reciprocally, the same grammar may generate any inflected and derived form of a simple verb.

Foremost, I explain how I have formalized the conjugation of any Quechua verb. This formalization is based on the behavior of the Indefinite Tense endings (IT)[2] of Quechua and its morph-syntactic behavior. This IT set is used in the conjugation of any verb in the indefinite tense[3]. Then, using the results of Duran (2013, 2014, 2017), who studied the different transformations of IT, I constructed the paradigms allowing to obtain all the nine sets of transformed endings (TRITn, n=1..9)[4], which allowed Duran, to find more comprehensive grammars serving to obtain the conjugated forms of multi-suffixed derived verbs, in all tenses and moods. (eg. *rimani*/ I talk, *rimaiman*/ I would talk, …)

I have tried to apply those grammars to study a social science interrogation concerning the frequency of usage of multi-suffixation in everyday life communication among the Quechua population. For this, I used two kinds of corpora: one written and one oral. The written corpora consist of eight novels and some historical narrations in Quechua. The oral corpus

[1] NooJ is an open-source linguistic platform well adapted to NLP studies.

[2] The indefinite tense[2] suffixes are IT = (*ni, nki, n, nchik, niku, nkichik, nku*).

[2] The indefinite tense[2] suffixes are IT = (*ni, nki, n, nchik, niku, nkichik, nku*).

[3] We call « indefinite » tense (IT), because they serve at the same time, to conjugate the Present: *rima-ni / I talk*, the Past: *rima-ni* [I talked], and in the Habitual mood: *rima-ni* / *I am used to talk*.

[4] e.g. TRIT1 = (*i, nki, n, nchik, iku, nkichik, nku*) ; TRIT2 = (*i, iki, n, nchik, iku, ikichik, nku*), etc.

consists of the transcriptions of some registrations of radio broadcasting of Radio Quispillaqta in Cuzco, Peru. I utilize the constructed formal grammars, cited above, to retrieve the verb forms bearing concordances with them and extract those forms. Then, after applying some simple statistical calculations to these results, I was able to answer the posed interrogation.

2 Fundamental conjugation

Within the Quechua verb conjugation system, there is one which I consider a fundamental conjugation (FC). It has the following structure detailed in Table 1:

Vroot+ITending,

Table 1the indefinite-tense fundamental conjugation of a Quechua Verb.

QU Pronoun	*EN Pronoun*	*Verb root*	*ITending*
ñoqa	I	root	*ni*
qam	You	root	*nki*
pay	he, she	root	*n*
ñoqanchik	we (inc.)	root	*nchik*
ñoqaiku	we (excl.)	root	*niku*
qamkuna	you	root	*nkichik*
paykuna	they	root	*nku*

Where, ITending or ITE is the set of personal endings.

ITE= (*ni, nki, n, nchik, niku, nkichik, nku)*

Each one marks a grammatical person in the conjugation of a verb. For instance, if we see *miku**ni***, I can say: it's the verb *mikuy /to eat* conjugated in the first person, because of the presence of its corresponding mark: the suffix « *ni* ».

We formalize this conjugation structure, symbolized as IT, by the following NooJ grammar:

IT = <B> (*ni*/IT+1+s | *nki*/IT+2+s | *n*/IT+3+s | *nchik*/IT+1+pin | *niku*/IT+1+pex | nkichik/IT+2+p | *nku*/IT+3+p);

Where we have: 1+s: first-person singular; 2+s: second-person singular; 3+s: third-person singular; 1+pin: inclusive first-person plural; 1+pex: exclusive first-person plural; 2+p: second-person plural; 3+p: third person plural and the annotation of the indefinite tense IT. Parsing IT on the verb to *work/ llamkay* generates table 2.

Table 2 the indefinite-tense conjugation of the verb llamkay/ to work.

QU Conjugation	*EN Conjugation*
ñoqa llamkani	I work
qam llamkanki	you work
pay llamkan	he, she works
ñoqanchik llamkachik	we (inc.) work
ñoqaiku llamkaniku	we (excl.) work
qamkuna llamkankichik	you work
paykuna llamkanku	they work

For the future, the corresponding set of endings is FE= (*saq, nki, nqa, sunchik, saqku, nkichik, nqaku*). Here also, we remark that each person has a specific mark. We symbolize by FUT the corresponding formalized grammar of the future tense:

FUT = <B> (saq/F+1+s | *nki*/F+2+s | *nqa*/F+3+s | *saqku*/F+1+pex | *sunchik*/F+1+pin | *nkichik*/F+2+p | *nqaku*/F+3+p) ;

Table 3 the future-tense conjugation of the verb llamkay/ to work.

QU F Conjugation	*EN F Conjugation*
ñoqa llamkasaq	I will work
qam llamkanki	you will work
pay llamkanqa	he, she will works
ñoqanchik llamkasunchik	we (inc.) will work
ñoqaiku llamkasaqku	we (excl.) will work
qamkuna llamkankichik	you will work
paykuna llamkanqaku	they will work

A more general typical Quechua conjugated verb form[5] has the following structure:
Vroot + IPS + ITE or
Vroot + IPS + FE
Where:
V: stands for the root of the verb
IPS[6]: Interposed suffixes, is the set of 27 derivation suffixes
ITE: is the set of seven indefinite tense endings cited above
FE: is the set of seven future tense endings cited above.

3 Verb to verb derivations

According to Quechua verb morphology, the 27 IPS suffixes, agglutinated to the root of a verb, can generate a similar number of new verbs. For instance, let's take the verb *tikray/ to turn* and agglutinate to it the IPS suffixes (***chi, ku, mu, pa, yku***), we'll obtain the following derivations:

*tikra**yku**y* / to revolve
*tikra**pa**y* / to mix up
*tikra**mu**y* / to turn towards the speaker
*tikra**ku**y* / to turn against
*tikra**chi**y* / to handle

If we want to get all the mono-suffixed verb-to-verb derivations, we may apply the graphical grammar of Fig 1:

[5] Actually, a more general conjugated form may include the post-position suffixes (placed after the ITE endings or F endings), I present it in Section 5 of this paper.
[6] IPS =(chaku, chi, chka, ykacha, ykachi, ykamu, ymana, ykapu, ykari, yku, ysi, kacha, kamu, kapu, ku, lla, mu, naya, pa, paya, pu, raya, ri, rpari, rqu, ru, tamu)

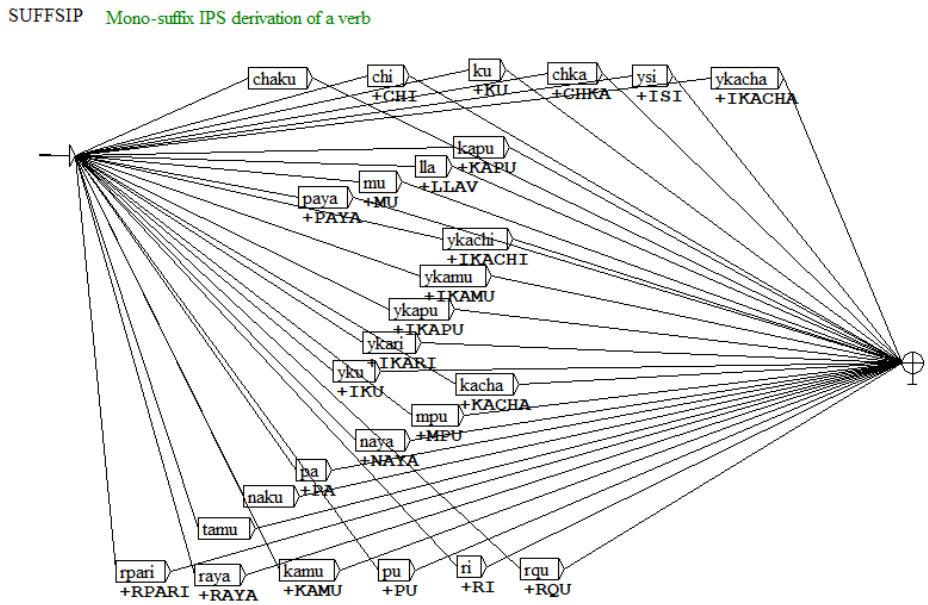

Figure 1 Mono-suffix verb to verb derivation grammar.

These newly generated verbs can be conjugated, in the indefinite tense IT[7], by the following formal grammar written for the NooJ platform:

SIP1_IT_CONJ = <B>(:CHAKU |:CHI | :CHKA |:YKACHA | :YKACHI | :YMANA| :YKAMU | :YKAPU | :YKARI | :YKU | :YSI | :KACHA | :KAMU | :KAPU | :KU | :LLAV | :MU | :NAYA | :PAV | :PAYA | :PU | :RAYAV | :RIV | :RPARI | :RQU | :RU | :TAMU) (ni/PR+1+s |nki/PR+2+s |n/PR+3+s |nchik/PR+1+pin |niku/PR+1+pex |nkichik/PR+2+p |nku/PR+3+p);

For instance, considering the verb *rimay/ to talk*, the grammar of Mono-suffix verb to verb derivation grammar. Figure 1 Mono-suffix verb to verb derivation grammar., generates the following derived forms:

```
rimatamuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+AEP+INF
rimaruy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+PRES+INF
rimarquy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+PAPT+INF
rimarpariy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+ASUR+INF
rimariy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+DYN+INF
rimarayay,rimay,V+FR="to talk"+FLX=V_SIP1_INF+DUR+INF
rimapuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+APT+INF
rimapayay,rimay,V+FR="to talk"+FLX=V_SIP1_INF+FREQ+INF
rimapay,rimay,V+FR="to talk"+FLX=V_SIP1_INF+PEAU+INF
rimanayay,rimay,V+FR="to talk"+FLX=V_SIP1_INF+ENV+INF
rimamuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+ACENT+INF
rimallay,rimay,V+FR="to talk"+FLX=V_SIP1_INF+POL1+INF
rimakuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+AUBE+INF
rimakapuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+RAS+INF
rimakamuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+AOL+INF
rimakachay,rimay,V+FR="to talk"+FLX=V_SIP1_INF+ARO+INF
rimaysiy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+COLL+INF
rimaykuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+COURT+INF
rimaykariy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+PONC+INF
rimaykapuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+SOIN3+INF
rimaykamuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+PREAT+INF
```

[7] To get the gerund forms we should add the corresponding paradigms: spa/GER |:SPA_GER1 |:PTI_GR2 | stin/GER3).

```
rimaymanay,rimay,V+FR="to talk"+FLX=V_SIP1_INF+INT+INF
rimaykachiy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+POLI+INF
rimaykachay,rimay,V+FR="to talk"+FLX=V_SIP1_INF+DISP+INF
rimachkay,rimay,V+FR="to talk"+FLX=V_SIP1_INF+PROG+INF
rimachiy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+FACT+INF
rimachakuy,rimay,V+FR="to talk"+FLX=V_SIP1_INF+DVAL+INF
```

Then, we pick up any of these newly derived verbs like *rimaykuy/ to greet* and we can conjugate it automatically applying the above cited grammar SIP1_IT_CONJ and obtain 189 conjugated forms looking like in the following list:

```
rimanayani,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ENV+PR+1+s
rimanayanki,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ENV+PR+2+s
rimanayan,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ENV+PR+3+s
rimanayanchik,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ENV+PR+1+pin
rimanayaniku,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ENV+PR+1+pex
rimanayankichik,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ENV+PR+2+p
rimanayanku,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ENV+PR+3+p
rimamuni,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ACENT+PR+1+s
rimamunki,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ACENT+PR+2+s
rimamun,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ACENT+PR+3+s
rimamunchik,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ACENT+PR+1+pin
rimamuniku,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ACENT+PR+1+pex
rimamunkichik,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ACENT+PR+2+p
rimamunku,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+ACENT+PR+3+p
rimallani,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+POL1+PR+1+s
rimallanki,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+POL1+PR+2+s
rimallan,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+POL1+PR+3+s
rimallanchik,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+POL1+PR+1+pin
rimallaniku,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+POL1+PR+1+pex
rimallankichik,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+POL1+PR+2+p
rimallanku,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+POL1+PR+3+p
rimakuni,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+AUBE+PR+1+s
rimakunki,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+AUBE+PR+2+s
rimakun,rimay,V+FR="to greet"+FLX=SIP1_IT_CONJ+AUBE+PR+3+s
```

The corresponding formalized grammar for the conjugation in the future tense of the derived verbs has the following formula[8]:

SIP_F_CONJ = <B>(:CHAKU |:CHI | :CHKA |:YKACHA | :YKACHI | :YMANA| :YKAMU | :YKAPU | :YKARI | :YKU | :YSI | :KACHA | :KAMU | :KAPU | :KU | :LLAV | :MU | :NAYA | :PAV | :PAYA | :PU | :RAYAV | :RIV | :RPARI | :RQU | :RU | :TAMU)(saq/F+1+s| nki/F+2+s | nqa/F+3+s | saqku/F+1+pex | nqaku/F+3+p | sunchik/F+1+pin | nkichik/F+2+p).

If we apply a combination of these grammars to the verb *llamkay/ to work*, we'll obtain automatically, all the conjugated forms in the Present, Past, and Future tenses, as well as in the different moods and aspects. I show in Figure 2 an extract of the 329 generated conjugated forms.

[8] The details of the construction of this paradigm can be found in Duran, 2017.

```
# NooJ V7 # Dictionary #Language is: qu
llamkani,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+1+s
llamkanki,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+2+s
llamkanku,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+3+p
llamkan,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+3+s
llamkankichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+2+p
llamkanki,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+2+s+F
llamkanqa,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+F+3+s
llamkanqaku,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+F+3+p
llamkasaq,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+F+1+s
llamkankichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+F+2+p
llamkai,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+IP+2+s
llamkachun,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+IP+3+s
llamkasun,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+IP+p+1
llamkaichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+IP+2+p
llamkarqun,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PAPT+3+s
llamkarqunkichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PAPT+2+p
llamkaspa,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER
llamkaspaiki,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+2+s
llamkaspanku,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+3+p
llamkaspai,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+1+s
llamkaspan,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+3+s
llamkaspaikichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+2+p
```

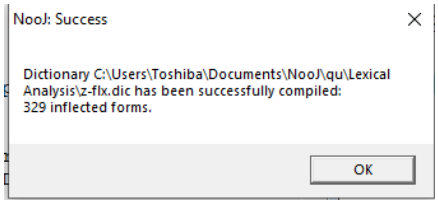

Figure 2 Automatically generated conjugated forms in different tenses, moods and aspects for llamkay[to work].

At this point, we may already use the reciprocal function in NooJ to retrieve and extract all the verbal forms containing one IPS suffix. I applied it to our corpus and found 773 derived or inflected forms, as shown in Figure 3.

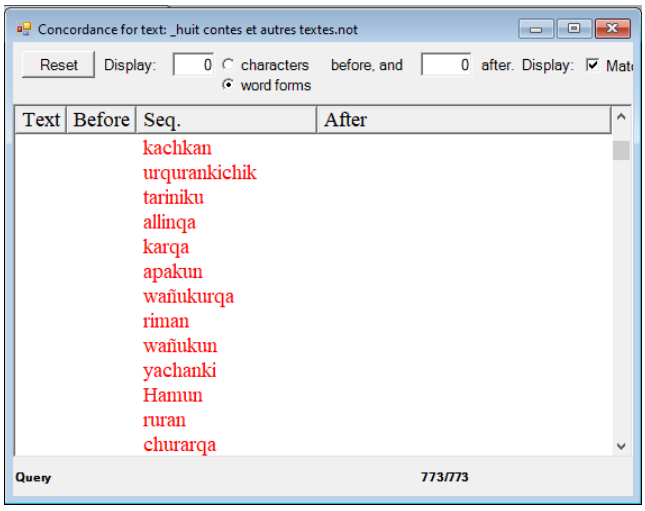

Figure 3 Retrieved 1-IPS verbal forms in the corpus.

4 Topological transformations of the indefinite tense IT

We remark that certain verbal forms are conjugations based on simple (topological) transformations of the set of ITE endings mentioned above. For instance, if we take the following transformation:

- The *–ni*, the first singular person mark loses '*n*' and is replaced by *'-i*' and the ending -*niku* marking the first plural person in its exclusive [9] form, loses also its '*n*' and is replaced by '*-iku'*, and we leave the rest of the endings unchanged, this produces the function T1 as a first transformation of ITE:

  T1(*ni, nki, n, nchik, niku, nkichik, nku*) = (*i, nki, n, nchik, iku, nkichik, nku*)

[9] Inclusive 'we' includes the interlocutor(s) to whom we speak to, whereas the exclusive 'we' excludes him (them).

Using this transformed set of endings we may obtain the conjugations in the optative, the conditional, and the potential moods if we associate it with the PPS[10] suffix '*-man'*.
If we take into consideration the verb *mikuy/ to eat*, we'll have for the conditional mood, the following conjugation (put in Spanish):

ñoqa	*miku-i-man*	yo comería
qam	*miku-nki-man*	tu comerías
pay	*miku-n-man*	él (ella) comería
ñoqanchik (inc.)	*miku-nchik-man* [11]	nosotros comeríamos(incluyendo a Uds.)
ñoqaiku (excl.)	*miku-iku-man*	nosotros comeríamos (sin incluir a Uds)
qamkuna	*miku-nkichik-man*	Uds. comerían
paykuna	*miku-nku-man*	ellos (ellas) comerían

Analyzing the Quechua morphology carefully, we can identify some other sets of endings as being the result of similar transformations of the ITE endings. For instance, a second type of transformation (T2) is the following: T2(*ni*, *nki*, *n*, *nchik*, *niku*, *nkichik*, *nku*) = (*i*, *iki*, *n*, *nchik*, *iku*, *ikichik*, *nku*)

It's known that, this new set of endings is associated with the following IPS suffixes: '*-na*, *-pti*, *-spa*, *-sqa'*. They participate in the conjugations of the gerunds, the subjunctive and in the obligation moods. Let's see how this happens: Taking again the verb *mikuy/ to eat*, we can observe how the transformed suffixes are present in the conjugation for the gerund 1, (translation put in Spanish):

ñoqa	miku-spa-i	risaq	comiendo iré
qam	miku-spa-iki	rinki	comiendo irás
pay	miku-spa-n	rinqa	comiendo irá (él, ella)
ñoqanchik	miku-spa-nchik	risunchik	comiendo iremos
ñoqaiku	miku-spa-iku	risaqku	comiendo iremos (sin tí o Uds.)
qamkuna	miku-spa-ikichik	rinkichik	comiendo irán (Uds.)
paykuna	miku-spa-nku	rinqaku	comiendo irán (ellos, ellas)

The other transformed endings that I have arrived to identify, which I symbolize by T3… T6 and applicable to transitive verbs appear below in table 2.
Taking into account all these transformed forms, I have enhanced my grammars to retrieve and extract out of the same corpus additional verbal forms, those containing these transformations. I found 320 more forms, as shown in Figure 4.

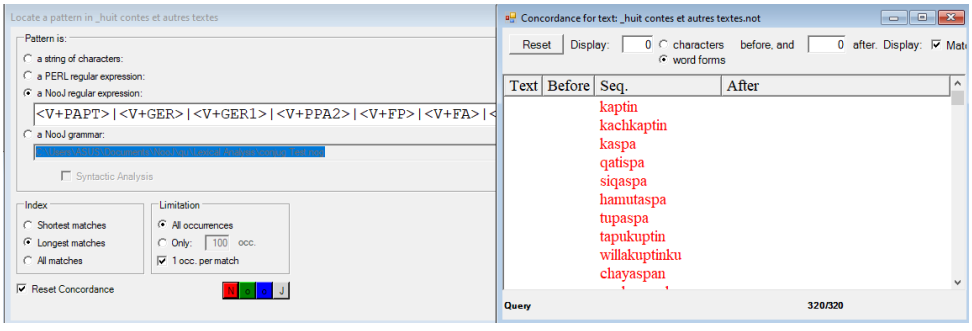

Figure 4 Recognized and extracted transformed forms of ITE.

[10] The PPS, post-position suffixes are placed after the ending ITE, or its transformations, of a conjugated form. They are: (*ch, chá, chik, chiki, chu ?, chu, chusina, má, man, m, mi, ña, pas, puni, qa, raq, ri, si, s, taq, yá*).
[11] The combined form *-chwan,* which gives *mikuchwan* is equivalent to the form –inchikman which gives *mikunchikman*.

Table 4 Endings structures obtained by transformations of ITE.

IT ending	T1	T2	T3	T4	T5	T6
ni	*i*	*i*	_	_	_	*iki*
nki	*nki*	*iki*	*nki*	_	*i*	_
n	*n*	*n*	*n*	_	*chun*	*sunki*
nchik	*nchik*	*nchik*	_	*nchik*	_	_
niku	*iku*	*iku*	_	*niku*	_	*ikiku*
nkichik	*nkichik*	*ikichik*	*nkichik*	*nkichik*	*ichik*	_
nku	*nku*	*nku*	*nku*	*nku*	*chunku*	*sunkiku*
For all the IPS suffixes	For SPP *man*	*For na, pti, spa, sqa*	*For -wa translative on object s+1 & subject 2+s & 3*	*exclusively applied to plural subjets*	imperative, translative on 1+s with *wa*	For translative on 2+s

5 Multi-suffix derivations

It's possible to combine two IPS suffixes. Furthermore, a remarkable property of the language is that many of these combinations give rise to valid bi-suffix verb-verb derivations. For instance, the combination *-chimu-*, of the suffixes *-chi* and *-mu*, applied to the transitive verb root *miku-*/ *to eat* produces the derivate form:

*miku-**chi-mu**-y* / make someone talk.

Its conjugation at the 2nd singular person is: *miku-**chi-mu**-nki*/ you go to make someone to eat.

I have found a total of 292 compatible two-suffix combinations like the following:

(*-chichi, -chichka, -chiykachi, -chiyku, -chiysi, -chykapu, -chiku, -chilla, -chimu, -chykamu, -chipu, -chira, -chirpari, -chirqa, -chisqa, -chitamu, -chiykamu, -chiykapu, -ykarisqa, -ykuchi, -ykuchka, -ykuku, -ykulla, -ykumu, -ykupu, -ykura,…*).

This list does not include the IPS suffixes (*-na,- pti, -ra, -rqa, -spa, -sqa, -stin, -wa)* because they are not compatible with the IT structure. They are associated with the transformed structures T1, T2, T3,…T6.

6 Mixed conjugated forms of IPS and PPS.

In our corpus, as in any daily oral or written communication, many Quechua verbal forms are mixed IPS-PPS forms; that mean that a single «word» may contain one or more IPS and one or more PPS. I call them «mixed inflections», let's see some examples (with translations in French):

mikukurqanpaschik « on peut croire qu'il a mangé aussi » (2+2).

This form contains two IPS suffixes (*ku, rqa*) and two SPP suffixes (*pas, chik*);

mikuchakunallanpaq « pour qu'il ait la gentillesse de manger » (4+1).

In this form there are four IPS suffixes (cha, *ku,na, lla*) and one SPP suffix (*pas*);

qayachakullawaptiykipas « même dans le cas où tu m'appellerais gentiment » (5+1) (with T1).

In this form there are five IPS suffixes (*cha, ku, lla, wa, pti*) and one SPP suffix (*pas*);

qayachamuchkallaykimanraqpashina « c'est comme si j'étais gentiment en train de t'appeler » (4+4) (with T1) ;

This form includes four IPS suffixes (*cha, mu, chka, lla*) and four SPP suffix (*man, raq, pas, hina*);

qayachamuchkallanmanraqpashinas « on dit que c'est comme s'il étais en train d'appeler gentiment » (4+5).

In this form we have four IPS suffixes (*cha, mu, chka, lla*) and five SPP suffix (*man, raq, pas, hina, s*);

In order to generate all the mixed combination forms automatically, we need to formalize these mixed grammars. I have built some. I show the following grammar V_MIX1 which includes 2x2= 4 sub-grammars, 2 for the IPS and 2 for the PPS:

```
V_MIX1= (:SIP1_TI_V) (:SPP1_V) | (:SIP1_TI_C) (:SPP1_C)
 | (:SIP1_TIM_V) (:SPP1_V) | (:SIP1_TIM_C) (:SPP1_C);
```

Let's present the extensive formula of one of the included paradigms, that of SIP1_T1_V:

```
 SIP1_TI_V = :CHI_TI_V |:CHKA_TI_V |:YKACHA_TI_V| :YKACHI_TI_V
| :YKAMU_TI_V | :YKAPU_TI_V | :YKARI_TI_V | :YKU_TI_V |:YSI_TI_V
|:KACHA_TI_V | :KAMU_TI_V | :KAPU_TI_V | :KU_TI_V | :LLAV_TI_V
| :MPU_TI_V | :MU_TI_V | :NAV_TIM2_V |:NAYAV_TI_V |:PAV_TI_V
|:PAYA_TI_V |:PU_TI_V | :RAYAV_TI_V | :RI_TI_V | :RPARI_TI_V
| :RQU_TI_V | :RU_TI_V |:SQA_TI_V |:TAMU_TI_V;
```

The grammar V_MIX1 generates 4 495 mixed combination forms for the verb *kuyay/ to love.* They look like the following forms:

kuyachikunchikmi, *kuyay*,V+FR=« aimer »+ FLX=V_MIX21+FACT+AUBE+TI+pin+1
kuyachikunchikña, *kuyay*,V+FR=« aimer »+ FLX=V_MIX21+FACT+AUBE+TI+pin+1
kuyachikunchikchu?, *kuyay*,V+FR=« aimer »+ FLX=V_MIX21+FACT+AUBE+TI+pin+1
kuyachikunchá, *kuyay*,V+FR=« aimer »+ FLX=V_MIX21+FACT+AUBE+TI+s+3
kuyachillankichikchiki, *kuyay*,V+FR=« aimer »+ FLX=V_MIX21+FACT+POL1+TI+p+2

Following the same scheme, I have constructed some other grammars capable of generating other types of mixed verbal forms:

V_MIX12 (1 IPS et 2 PPS)
V_MIX21 (2 IPS et 1 PPS)
V_MIX22 (2 IPS et 2 PPS).

The first grammar concerns the verbal forms made of one IPS and two PPS suffixes, the second one concern the verb forms made of two IPS and one PPS suffixes, etc.

We may ask ourselves if all these automatically generated forms have actual meaning and are utilized in daily communication. We answer this question in the affirmative in a separate work in progress.

7 Retrieval of inflected and derived verbs

Having constructed the preceding grammars, we can use them to retrieve and extract the verbs that are present, in any conjugated form, within a given corpus. We start by searching the verb forms bearing one derivation suffix. To get them, we have utilized the paradigm V_TR_SIP1_CONJ, combined with two additional ones V_ITR_SIP1_CONJ and V_IMP_SIP1_CONJ[12] and found the forms containing intransitive and impersonal verb roots; for the verb-forms including two or more IPS or mixed combination forms, we apply the other grammars that I have presented before. As a result, I have found 2484 derived, inflected, and mixed-conjugated verb forms in the written corpus. I show some of them in Figure 5 (we can distinguish them in the column labeled as Seq.).

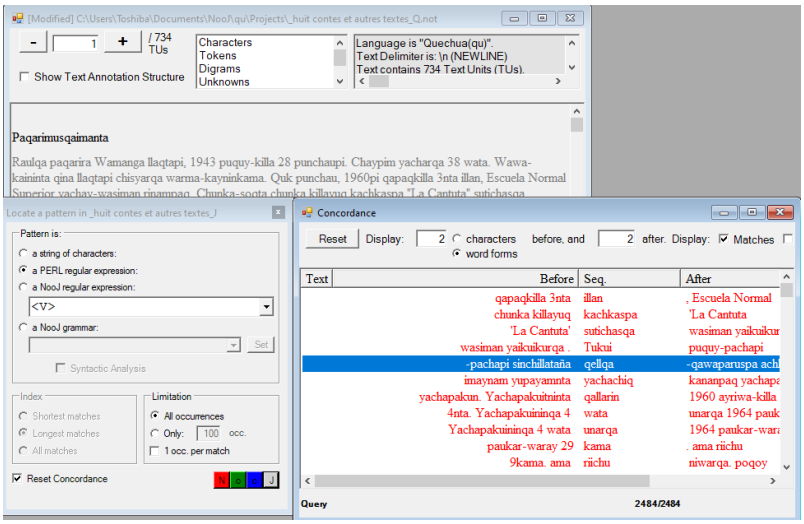

Figure 5 total conjugated verb forms recognized in the written corpus.

8 A socio-linguistic application

How frequent is the use of derived verbs in everyday communication?

a. Analyzing our first results, we remark that the mono-suffixed verb forms appearing in Figure 3 cover 27.38% of the entire verb forms embedded in the text.
b. For the **written** narrations, like our corpus of tales, a frequency analysis shown in Figure 6, seems to show the relevance of the use of the verb *niy*[to say] and its derivations: 114/680 (16,76%), followed by the verb *kay[*to be] and its derivations: 60/680 (8,82%).

[12] The detailed description of these paradigms, corresponding to intransitive and impersonal verbs, can be found in Duran, 2017.

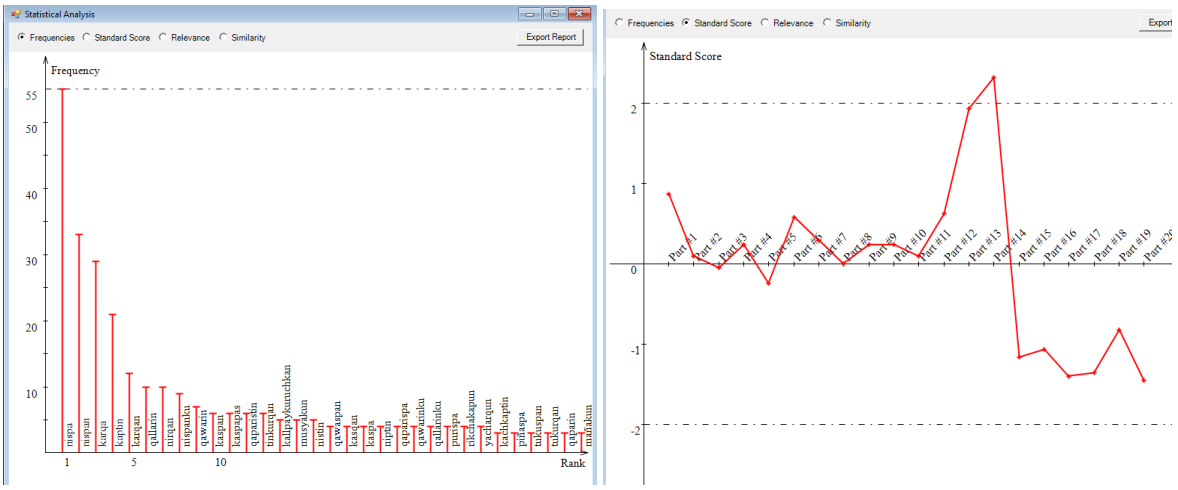

Figure 6 Frequency of the use of the verb niy/to say in written corpus and its standard score.

This fact contrasts with the values obtained for the oral discourse[13] corpus (see Figure 7), in which we found that the most frequently used verb is *kay/ to be* and its derivations (10/50) 20, 00%, whereas the second most used one is the verb *niy/ to say* and its derivations (7/50) 12, 72%.

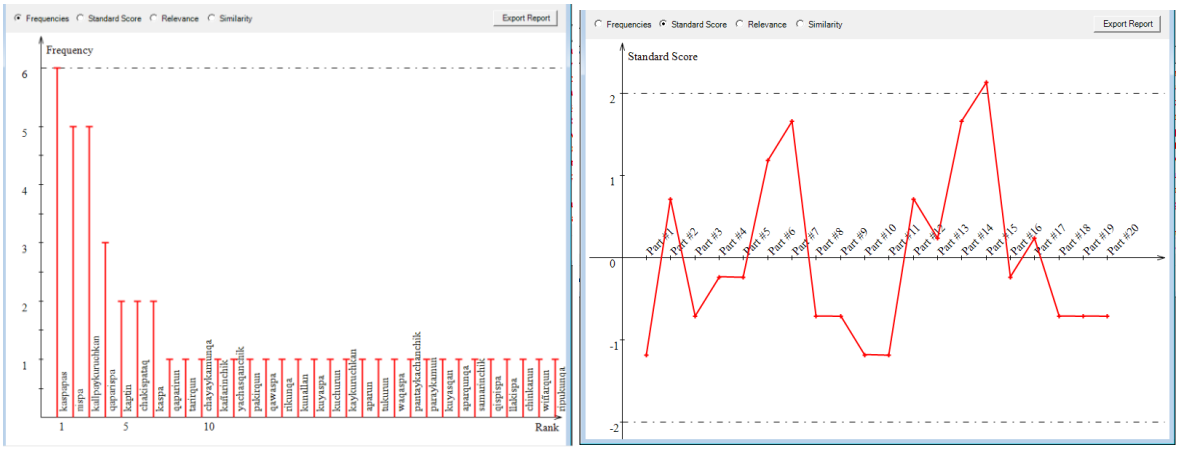

Figure 7 Frequency and standard score of the verb kay/ to be (in the oral corpus).

On the other hand, as we have seen the derivation of a Quechua verb may be obtained by combinations of two or more IPS suffixes. We would like to know, how frequently a speaker utilizes such complex verb form. Does the frequency of these multi-suffixed verb forms depend on the context of a written or an oral discourse? For an answer to the interrogation, let's analyze the following data.

In the case of **oral** discourse:

Total of verb forms with more than one derivation suffix in the oral corpus: 50

forms containing 2 or more derivation suffixes	Freq: 13	26%
forms containing 3 or more derivation suffixes	Freq: 6	12%

[13] A transcription of a transmission of the program « *Musuq kallpa* » in radio Kispillacta (*Mikuykunamanta* 19/03/2003) was realized by C. Itier, 2003.

In the case of the **written** discourse:

Total of verb forms with more than one derivation suffix in the written corpus: 680

forms containing 2 or more derivation suffixes	Freq: 60	8,8%
forms containing 3 or more derivation suffixes	Freq: 2	0,3%

These values suggest that verb forms containing two or more derivation suffixes are more frequent in an oral discourse than in a written one. In my opinion, the fact that Quechua remains, dominantly, an oral language demands more detailed phrasing to avoid ambiguities, and the abundance of suffixes furnishes more precisions.

Of course, this conclusion, taken from these limited corpora, may not represent the general case; I plan to verify it with larger corpora. We hope that the method and the paradigms we have presented in this work may contribute to it.

9 Conclusion and perspectives

I have presented a method allowing the automatic retrieval of derived and inflected Quechua verbs embedded in a corpus. It consists in the construction of paradigms and grammars, in the NooJ formalism, managing the grammatical combinations of verbal suffixes. Using these formalized grammars, I have found all the verb forms in written and oral corpora.

I have obtained a practical system that helps us to answer socio-linguistic questions. For the first question: How frequent is the use of multi-suffixation of verbs (derived by a combination of two or more suffixes) in everyday communication? We have found that more than one-fourth of the verb forms employed in oral communication are multi-derived verbs. About the second issue, is there a notable difference in this phenomenon between written and oral communication? We have found that verb forms containing two or more derivation suffixes are more frequent in an oral discourse (26%) than in a written one (8, 8%).

We plan to continue our tests on larger corpora to confirm or not our hypothesis on written context.

References

1. Dubois J. and Dubois-Charlier F. (2007). Dictionnaire Linguistique et Sciences du langage, Editions Larousse, Paris.
2. Dubois, J et Dubois-Charlier, F. (1997). *Les verbes français*, Larousse, Paris.
3. Duran M. (2013). Formalizing Quechua verbs Inflexion. Proceedings of the NooJ 2013 International Conference, Saarbrücken.
4. Duran M. (2014) *Morphological and Syntactic Grammars for the Recognition of Verbal Lemmas in Quechua*. Proceedings of the NooJ 2014 International Conference, Sassari.
5. Duran M. (2017). Dictionnaire électronique français-quechua des verbes pour le TAL. Université Franche-Comté, Besançon-Paris. Thèse doctoral soutenu en mars, 600 p.
6. Itier C. (2003). Transcription des émissions de Radio Quispillaqta: Llaqtanchikpa sunqun, 18-03-2003.
7. Silberztein M. (2010). La formalisation du dictionnaire LVF avec NooJ et ses applications pour l'analyse automatique de corpus. Langages 3/2010 (n° 179-180), pp. 221-241.
8. Silberztein M. (2016). Language formalization: the NooJ's Approach. Wiley Eds. London.