Formalization of the Quechua Morphology

Maximiliano Duran
Université de Franche-Comté, CRIT, France
duran_maximiliano@yahoo.fr

Abstract

In this article, I present a method to formalize Quechua's morphology of nouns, verbs, and other Part Of Speech (POS) categories to develop Natural Language Processing (NLP) applications. I constructed an electronic corpus consisting of several digitalized texts and electronic dictionaries. After a detailed inventory of all Quechua suffixes, I classified them into specific sets corresponding to their POS category. Next, I formalized their grammatical behavior separately, using elementary matrices. The resulting tables describe valid combinations of two, three, and four suffixes. Finally, I converted these matrices into paradigms in NooJ's formalism, thus formalizing Quechua's inflection and derivation of each POS category.

Key Words: *Formalizing Quechua, electronic dictionary, suffix agglutination, Natural Language Processing, Computational Linguistics.*

1 Introduction

Until a few years ago, Quechua was a poorly endowed language. Now, because many internet sites referring to Quechua appear, we may think that the situation for this language has changed. However, if we intend to find open-source electronic lexical resources ready to use for an NLP project, the situation has only lightly evolved. The central question is does it exist an open-source tagged text corpus of Quechua.

1. The existing written corpus published up to the middle of the 20th century did not contain more than half a million Quechua tokens. It includes, in the majority, religious publications like the translation of the Christian Bible (New Testament) (492 859 tokens, where only less than 50 000 are distinct tokens), Divers Cuzco stories (12 223) cited by Monson et al. (2006), Stories of Urubamba of Lira (31 986), prayers (around 3 000), legends and stories of Huarochiri's manuscript (around 11 000), stories of Sta Cruz Pachakuti (around 2 500), Guaman Poma's chronicle Nueva Cronica y buen Gobierno (1496 different tokens) and some grammars and dictionaries such as Sto Thomas' (around 4 000), Holguín's (around 15 000). Blas Valera's (around 1000), Betanzos' (around 1000), and P. Meneses novels (around 5000). The number of distinct tokens in these texts does not exceed 80%.
2. Looking for electronic lexicons in Google, the answer to the query: 'Quechua electronic dictionary' shows an encouraging 7,680,000 results; however, using the same search engine, when I asked under the terms' Quechua dictionary' it gave 595,000 results, which is not coherent, and it let us suppose that the search engine has been misled by the word "electronic" which enlarges the result instead of reducing it. We then looked over the results.
3. For instance, I tried to enter a promising website under the title: Sketch Engine, which claims to have "tools to identify and analyze collocations, synonyms, and antonyms, examples of use in context, keywords or terms. Frequency word lists of Quechua single-word or multi-word expressions can be generated". When I typed the address https://www.sketchengine.eu/quechua-text-corpora", unfortunately, I got the answer "List of available Quechua corpora: No corpora available at the moment." Many of the found sites are only commercial publicity; they are presented because they contain one of the terms of the query.
4. Many listed printed dictionaries or PDF versions, such as the following two, don't allow downloading or visualizing any dictionary marked with the corresponding POS. https://issuu.com/idiomaquechua/docs/diccionarioquechua?pageNumber=1

https:// www.crisol.com.pe/libro-nuevo-diccionario-espanol-quechua-quechua-espanol-9789972607448.

Many of the listed dictionaries also lack POS tagging.

5. There are many question-answer sites to obtain translations of simple words or simple canonical phrases like: **http://www.sankaypillo.com/2014/07/diccionarios-electronicos-aymara.html** The dictionaries used in these applications are not visible.
6. Prestigious universities that include this language in their curricula and listed in the query results have inactive their Quechua websites. For instance, UCLA's **quechua.ucla.edu** website in the USA seems not operational since we get the message: "Digital Resources for the Study of Quechua is being upgraded and moved to new hosting. We will post new information here when it is available".
7. Google's Quechua site has an extensive database and dictionaries but remains a black box. Their methods favor the stochastic approach instead of the rule approach, but the linguistic resources used, such as their tagged corpus, are not accessible to the user.
8. Google's counterpart, "Español Quechua translator," does not show its databases, their corresponding tagged corpus, or other linguistic resources.
9. The website of one of the prominent universities of Peru, Universidad de San Marcos, https://www.facebook.com/CatedradeLenguaQuechuaUNMSM, contains specific information about the general courses of Quechua. It does not show documentation on Quechua linguistic resources ready to download.
   That one of the Universidad Católica in Peru https://idiomas.pucp.edu.pe/programas/quechua/curso-de-quechua/#nopresencial proposes Quechua basic lessons online also. It does not present any electronic dictionary or open-source corpora.
10. Some NLP projects have seen the light in the last decades. The Sequoia project of the University of Zurich, for instance, has developed several linguistic resources aiming to obtain a Quechua Spanish translation system. Their corpus includes several bilingual dictionaries and monolingual texts of the Cuzco variant.
    In her works, A. Rios (2011, 2016), one of the leading animators, proposes in her thesis and further articles some formalization for grammar, a morphological analyzer, a syntax analyzer, and an initial Spanish-Quechua translator.
11. The group of R.Cardenas, R. Zevallos, R. Baquerizo, and L. Camacho (2018), within the SIMINCHIK project, has been working on obtaining "a speech corpus suitable for training and evaluating speech recognition systems." They remark: "Peruvian native languages, amongst which Quechua is included, present scarce written footprint and are predominantly orally transmitted even today. Even worse, the amount of digital content in Peruvian languages is extremely low", and all of them are considered "under-resourced." In contrast, they announce good news: "we introduce the first speech corpus of Southern Quechua, Siminchik, suitable for training and evaluating speech recognition systems."[1] Concerning tokenization of their transcriptions using an algorithm used for specific agglutinative languages, called BPE, they make the following statement: "BPE represents text at the character level and then merges the most frequent pairs iteratively until a pre-determined number of merge operations has been reached," they remark "we note that the accuracy scores of our results are somewhat lower than the state-of-the-art for high-resource languages on the named-entity recognition (NER)." Which may come from the choice using statistics rather than grammar rules and the introduction of a class of sui-generis "prefix" and "postfix" particles in the marge of Quechua morphology which recognizes suffixes and their combinations.

[1] Called QuBERT, it appeared in 2022 claiming to be a "large combined corpus for deep learning of an indigenous South American low-resource language … created from text gathered from the southern region of Peru … the entire data set consists of 4,408,953 tokens and 384,184 sentences, including what are known as Chanka and Collao variants" indigenous South American low-resource language … created from text gathered from the southern region of Peru … the entire data set consists of 4,408,953 tokens and 384,184 sentences, including what are known as Chanka and Collao variants"

Before all these difficulties to obtain the needed resources and tools for building a rule-based system for MT of texts from French and Spanish into Quechua, I needed to start at the base: to build our own linguistic resources. We began thirty years ago with the construction of several electronic dictionaries (simple words, composed units, MWE, technical and scientific terms, etc.) in bilingual and bidirectional versions (SP-QU, QU-SP and FR-QU, QU-FR).

Concerning grammar, the obstacles were similar: a lack of coherent adequate grammar. Thus, I had to sketch an e-grammar (electronic grammar). There are many classical printed grammar texts of Quechua. Even though most respect the canonical rules of the language's morphology and syntax, almost all follow a pattern as if Quechua were an Indo-European language. They follow, in particular, the scheme traced by the Spanish linguist Nebrija of s. XV (searching prepositions, determinant adjectives, gender concordances, etc.). And yet, Quechua is not a Roman language. Its morphology and syntax are different; therefore, these grammar schemes are not necessarily pertinent to studying the Incas' agglutinative and SOV-type language. Quechua is a logical and poly-synthetic morpho-syntactic language, and we must approach its study as such. In this work, considering the indigenous point of view and, with the aid of computer tools, I present some details of our formalized Quechua e-grammar.

If we have to use statistical, stochastic, or neural network algorithms to process the language, we should possess a consequently written corpus and big data sets of aligned bilingual forms or expressions (FR-QU or SP-QU in our case). These resources are not available nowadays[2], but I hope they will be in the near future. To make things more complicated, there's still scarce written documentation in QU (as I have seen before, less than one million tokens if we include the recently added documentation). Let us present our advancements in building some fundamental resources: electronic dictionaries and formalized grammars.

2 Constructing electronic dictionaries

Because of its agglutinative attribute of the language, we should, in priority, formalize the rules of the ways how the suffixes agglutinate to the roots and combine with each other. First, I have inventoried all the Ayacucho Quechua suffixes taken from different authors like Perroud (1970), Pino (1980), Soto (1976), and my own introspection. We can remark that each POS possesses its set of suffixes. There exist some allomorphs amongst them. We present, for each POS, some details of their agglutinations and the corresponding paradigms allowing getting formal expressions for the inflection (FLX) or derivation (DRV), which will be attached to each of the entries in the dictionary. Thus, we'll present the corresponding suffixes for the Nouns, Adjectives, Verbs, pronouns, and adverbs separately.

2.1 Formalization of Quechua noun inflections

In 2012, I presented my first electronic dictionary of nouns in the article "Formalizing Quechua Noun Inflection" (Duran 2012, 2014), containing hundreds of simple nouns.

Duran (2021) presented an enhanced e-dictionary containing actualized noun inflection grammars FLX = NVOCAL or FLX=NCONSO, as shown in Figure 1.

Analyzing the nominal morphology for the Ayacucho-Chanka Quechua variant, I arrived to gather all the nominal suffixes shown in the set Suf_N.

SUF_N = *{-ch , -cha, -cha ,-chik, -chiki, -chu, -chu(?)*[3]*, -hina, -kama, -kuna, -lla, -má, -man, -manta, -m, -mi, -mpa, -nimpa, naq, -nta, -ninta*[4]*, -nintin, -ntin, -niraq, -niyuq, -niq, -ña, -p, -pa, -paq, -pas, -pi,*

[2] Recently, Cardenas et al. (op. cit.) say in their article that they have obtained a mono lingual data set of around 1 200 000 words, from the transcription of 97 hours of speech in Ayacucho and Cuzco dialects. This data set is not open yet. And for monolingual corpus see footnote 1.

[3] The interrogative and exclamation signs are used to indicate the presence ascendant intonation followed by a pause pour first and descendant followed by a pause pour exclamation sign.

[4] *ninta* is in fact a composition of the phonic support particle "*-ni"* and the suffix "*-nta",* applicable to nouns ending in a consonant. This same "ni" intervenes in the compositions *ninka*, *ninta*, *nintin*, etc.

-poss(7v+7c), -puni, -pura, -qa, -rayku, -raq, -ri, -s, -si, -sapa , -su, -ta, -taq, -wan, -y(!), -ya(!), -yá, -yupa , -yuq}[5] (50+7v+7c)

Where (7v+7c) represents the set of possessive suffixes (–i, -iki, -n, -nchik, -iku, -ikichik, -inku; –nii, -niiki, -nin, -ninchik, -niiku, -niikichik, -niinku). The first seven correspond to nouns ending with a vowel, and the remaining seven to nouns ending in a consonant.

A detailed description of the semantics induced to a noun by each nominal suffix can be found in chapter 2 of Duran (2017).

Besides Suf_N, there also exists a set of nominal suffixes which derive nouns into verbs:

S_N_V= {*y, yay, chay*}.

Some nouns accept only one, some two, and some all three of these suffixes, i.e.

taki/ song → takiy/ to sing
wasi/ house → wasiyay/ to become a shelter
wasi/ house → wasichay/ to cover a house
rumi/ stone → rumichay/ to cobble

Figure 1 presents an excerpt of both QU-SP and QU-FR noun dictionaries, in which each noun's inflectional (FLX) and derivational (DRV) grammars are specified.

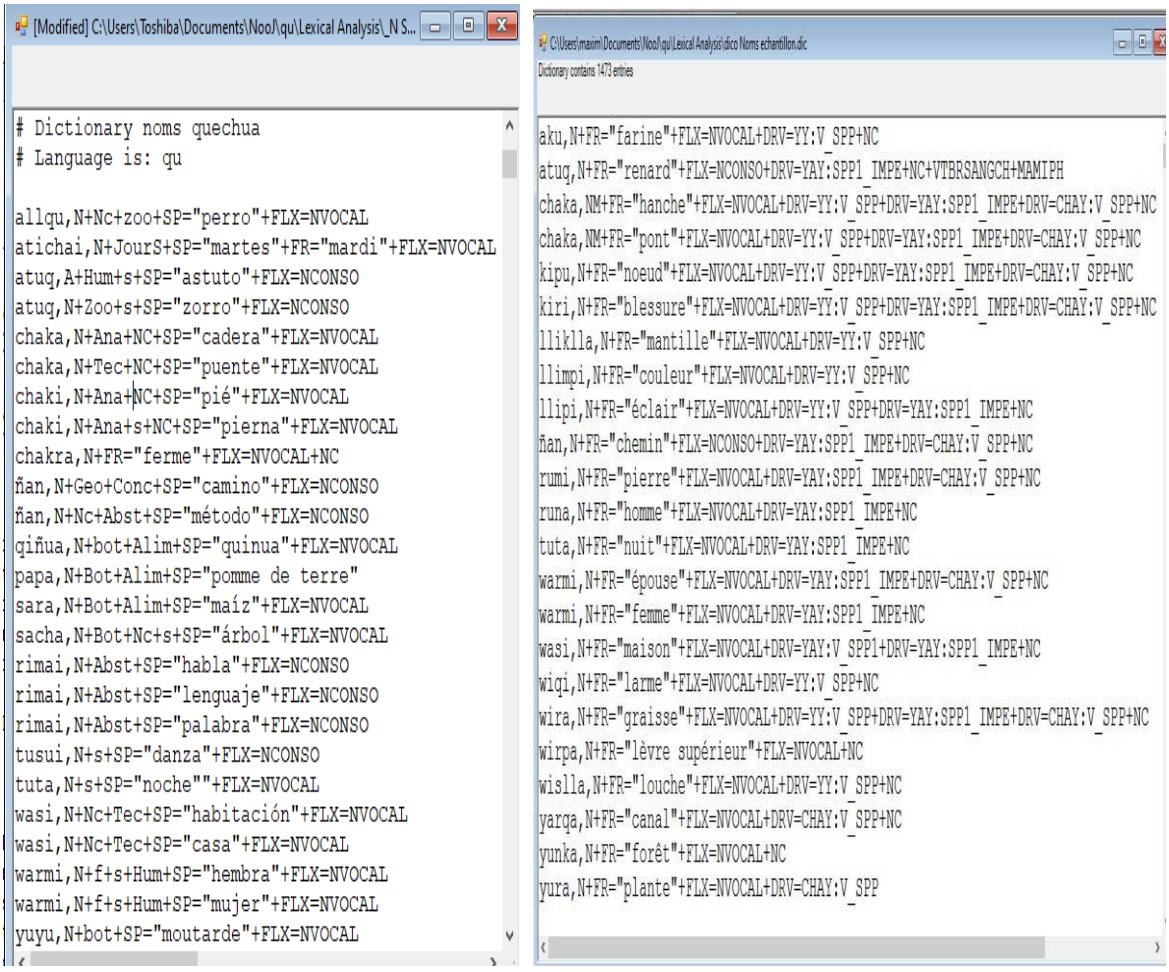

Figure 1. Sample of the QU-SP and QU-FR dictionaries containing (17000 entries for first).

Depending on whether the ending of the noun is in a vowel or a consonant, the inflection FLX grammar that will generate the corresponding inflections, maybe one of the following grammar rules:

[5] Found in (Guardia Mayorga 1973), (Perroud 1970), (Pino 1980), (Soto 1976)

NVOCAL= :NOM | :N_V_1 | :N_V_2 | :N_3_GEN ;
NCONSO= :NOM | :N_C_1 | :N_C_2 | :N_3_GEN ;

Where we find NOM symbolizing a sub-grammar involving 64 paradigms, it contains a single nominal suffix.
N_V_2 involves more than 600 paradigms containing combinations of two suffixes, whereas N_3_Gen involves several thousand paradigms containing combinations of three suffixes.

These combinations are given by the morpho-syntactic properties of the language:
wasikunaman/ towards the houses.

Parsing these grammars within the NooJ platform will generate thousands of inflected forms for each noun, as shown in the following extract for the noun *wasi*:

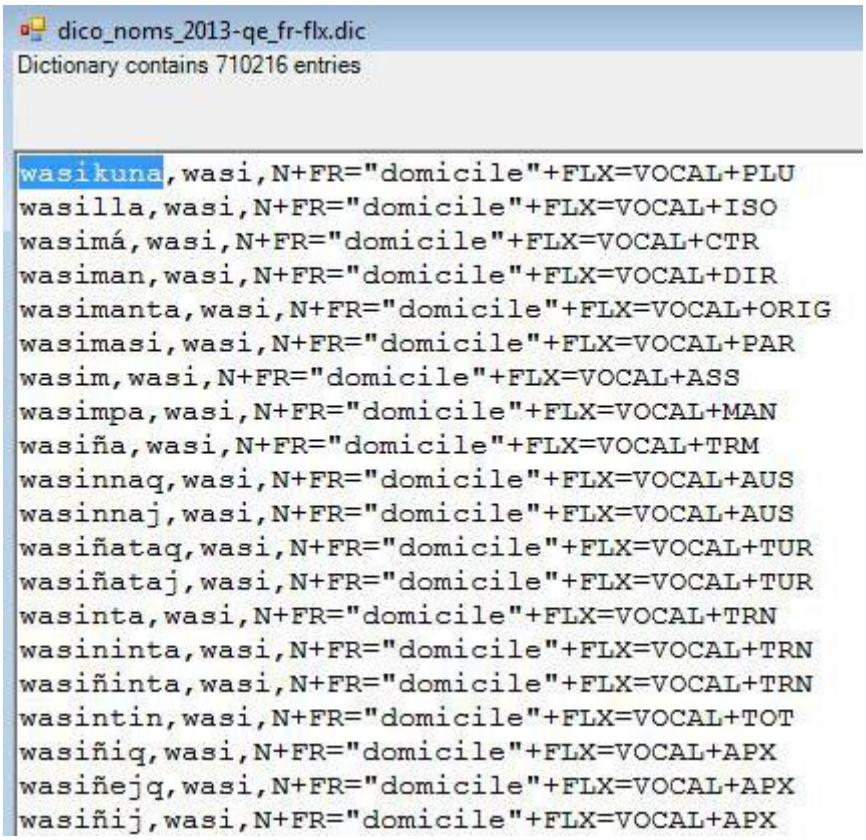

Figure 3 One-layer inflection of *wasi* /house

2.1.1 Tags.

In Figure 1, for the QU-SP case, we see the presence of some morpho-syntactic tags like N (representing nouns), Nc (common nouns), Nhum (human nouns), and (animal), FLX=NVOCAL (the inflectional paradigm for nouns ending in a vowel), and FLX =NCONSO (the inflectional paradigm for nouns ending in a consonant), Mamiph (mammal), zoo (zoology), Alim (food), Ana (anatomy), Tec (technic), etc. The full tag set contains around 50 non-exhaustive codes.

2.2 Formalizing Quechua verb morphology

Quechua's dictionary of verbs contains less than 1,200 simple verbs. It is a very modest size. Any translation project from SP FR, which contains more than 9,000verbs, should be handicapped by the difficulty of finding the translation of these even sizes. Fortunately, this language possesses an

interesting strategy based on a set of derivation suffixes and agglutinative properties, allowing considerable enhancement of the initial lexicon. Let us see how.

After the inventory of the verbal suffixes and, taking into account their morpho-syntactic behavior, I propose a basic classification of these verbal suffixes, as follows: Interposition suffixes (IPS)[6], Postpositional suffixes (PPS)[7], and the verb nominalizing suffix N_S[8]:

- A subset of the IPS's, which I symbolize as IPS_DV,[9] helps to generate new derived verbs,
- The N_S suffixes agglutinated, under certain linguistic conditions, to verbs will generate derived nominalized forms.
- The SPP will generate all kinds of agglutinated inflections, both with simple and derived verbs.

This classification has proved to be a very useful and pragmatic tool in the formalization of verb morphology [10] and helps to clearly recognize and analyze verb forms and avoid the difficulties and uncertainties of stochastic treatment as those Zevallos et al. (2022) signaled: "In order to statistically determine which branch of morphemes a verb phrase falls under can be difficult with Quechua since there are so few resources". Here are some examples of derivation V-V of simple verbs; in this case, I take the verb *maskay*/to search and obtain 27 new verbs agglutinating a single IPS:

maska-ku-y to search for oneself
maska-yku-y to search with determination
maska-pa-y to search zealously
maska-ri-y to search superficially
maska-chi-y to make someone search

Following similar steps as I did for the nouns, I constructed tables containing the matrices of grammatical agglutinations of two or more suffixes.

For instance, Figure 2 presents the Boolean matrix that represents the bi-suffix combinations of IPS suffixes.

	CHI	CHKA	IKACHA	IKACHI	IKAMU	IKAPU	IKARI	IKU	ISI	KACHA	KAMU	KAPU	KU	LLAV	MU	NAV
CHI	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	4
CHKA	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
IKACHA	1	1	0	0	1	1	0	0	1	0	0	0	1	1	1	0
IKARI	1	1	0	0	1	1	0	1	1	0	0	0	1	1	1	0
IKU	1	1	0	0	1	0	0	0	1	0	0	1	1	1	0	0
ISI	1	1	0	1	0	0	0	1	0	0	1	1	1	1	1	0
KACHA	1	1	0	1	1	1	0	1	0	0	0	1	1	1	1	1
KAPU	1	1	1	1	1	0	0	1	0	0	0	0	0	1	1	1
KU	0	1	0	0	1	1	0	1	0	0	0	1	0	1	0	1
LLAV	1	1	1	1	1	0	0	1	1	0	0	0	0	0	1	1
MU	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
NAV	0	0	0	0	0	0	0	0	0	0	4	0	0	2	0	0
KAMU	1	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0
NAKU	0	1	1	0	0	0	1	1	0	0	0	0	0	1	1	0
NAYA	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
PAYA	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
PU	0	1	0	0	1	1	0	1	0	0	0	0	1	0	0	0
RA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
RAYA	1	1	0	0	0	0	0	1	1	0	0	0	1	1	1	1

Figure 3 Part of the bi-dimensional combinations Matrix of IPS

[6] IPS= (*chaku, chi, chka, ykacha, ykachi, ykamu, ykapu, ykari, yku, ysi, kacha, kamu, kapu, ku, lla, mpu, mu, na, naya, pa, paya, pti, pu, ra, raya, ri, rpari, rqa, rqu, ru, spa, sqa, stin, tamu, wa).*
[7] SPP= (*ch, chá, chik, chiki, chu(?), chu, chusina, má, man, m, mi, ña, pas, puni, qa, raq, s, si, taq, yá(!)).*
[8] N_S= (*y,-na,-q,-sqa)*
[9] IPS_DV= (*chaku, chi, chka, ykacha, ykachi, ykamu, ykapu, ykari, yku, ysi, kacha, kamu, kapu, ku, lla, mpu, mu, naya, pa, paya, pu, raya, ri, rpari, rqu, ru, tamu)*
[10] Zevallos: "A short example sentence of how complex morpheme determination can be depicted in Table 1. In some cases, there are hundreds of options to choose from when choosing which suffix to use for a given Quechua word"

The corresponding Boolean matrix for the grammatical combinations of three IPS has, as its first row, the set of 27 IPSs' and, as its first column, the 295 valid binary combinations that I have obtained in Figure 3. Here are some resulting combinations:

> V_SIP3: rikuchka, rikuyku, rikuysi, rikukapu,rikulla, rikupa, rikupu, rikura, rikurqa, rikurqu, rikuru, rillaykacha, rillaykachi, rillaykari, rillaysi, rillara, rillaraya, rillarqa, rimuchka, rimuykari, rimulla, rimurqa, rimurqu,

There also exist combinations of four IPSs' like: *cha-ku-na-lla , cha-ku-lla-wa, cha-mu-chka-pti*

2.2.1 The electronic dictionary of Quechua verbs (FR-QU)

The electronic QU_FR dictionary of simple verbs, shown in Figure 4, contains **1,181 entries**.

NooJ - [C:\Users\ASUS\Documents\NooJ\qu\Lexical Analysis_V_SIMPLES QU_FR_CON_FLX.dic]

File Edit Lab Project Windows Info DICTIONARY

Dictionary contains 1181 entries

```
rikchay,V+tr+FR="ressembler(à)"+FLX=V_TR
rimay,V+tr+FR="parler"+FLX=V_TR
ripuy,V+itr+FR="partir"+FLX=V_ITR
ripuy,V+itr+FR="s'éloigner"+FLX=V_ITR
ripuy,V+itr+FR="s'enaller"+FLX=V_ITR
ripuy,V+itr+FR="voyager"+FLX=V_ITR
riqsiy,V+tr+FR="connaître"+FLX=V_TR
rirpuy,V+itr+FR="refléter"+FLX=V_TR
rirpuy,V+tr+FR="mirroiter"+FLX=V_TR
rirpuy,V+tr+FR="refleter"+FLX=V_TR
ritiy,V+imp+FR="grêler"+FLX=V_CONJ_IMP
ritiy,V+imp+FR="neiger"+FLX=V_CONJ_IMP
riy,V+tr+mov+FR="aller"+FLX=V_TR
ruiruy,V+tr+mov+FR="roder"+FLX=V_TR
rukriy,V+tr+FR="creuser"+FLX=V_TR
rukriy,V+tr+FR="miner"+FLX=V_TR
rupay,V+imp+FR="brûler"+FLX=V_CONJ_IMP
rupay,V+imp+FR="ensoleiller"+FLX=V_CONJ_IMP
rupay,V+imp+FR="faire chaud"+FLX=V_CONJ_IMP
rupay,V+tr+FR="brûler"+FLX=V_TR
ruray,V+tr+FR="fabriquer"+FLX=V_TR
ruray,V+tr+FR="faire"+FLX=V_TR
ruray,V+tr+FR="former"+FLX=V_TR
ruruy,V+itr+FR="fructifier"+FLX=V_ITR
rutuy,V+tr+FR="couper"+FLX=V_TR
rutuy,V+tr+FR="faucher"+FLX=V_TR
```

Figure 4 Extract of the electronic dictionary of Simple QU_FR verbs

In this extract:

- There are no compound verbs nor phrasal verbs, which are represented in another dictionary.
- Each verb has an inflectional paradigm FLX= V_TR for transitive verbs or FLX= V_ITR for intransitive ones. For instance, the entry unit *rimay* / to talk inflects according to the paradigm V_TR; thus, the entry becomes *rimay*,V+tr+FR= "parler"+FLX=V_TR.

- It also contains some syntactic and semantic information, like the two main classes of verbs: Transitive (tr): *rimay* / to talk, Intransitive (it): *mikuy* / to eat.
- The intransitive class is relatively small. It contains less than one hundred verbs. The class of impersonal verbs (imp) includes mainly those relating to the weather: *paray* to rain; *lastay* to snow.

The inflection formula V_TR:

V_TR =:V_SPP | :V_TR_SIP | :V_CONJ_TR | :I_TR;

Where we can find the following formulae embedded:

V_TR_SIP = :SIP1_G | :SIP2_G | :SIP3_G;

SIP1_G = <E>/INF | :SIP1 | :V_SIP1_INF | :SIP1_N;

V_SIP1_INF = <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)y/INF;

It gathers, among other paradigms, those of the conjugation of present, past, and future tenses

The detailed algebraic expression of PR and FUT, for instance, are:

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

Example: *taki-ni* I sing, *taki-nki* you sing *taki-n* he sings

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

Example: *taki-saq* I will sing, *taki-nki* you will sing *taki-n* he will sing

Parsing our dictionary of 1180 simple verbs with the V_TR grammar, we obtained 31,860 new verbs, an extract of them is shown in Figure 3. Which extends the initial verb lexicon of verbs considerably.

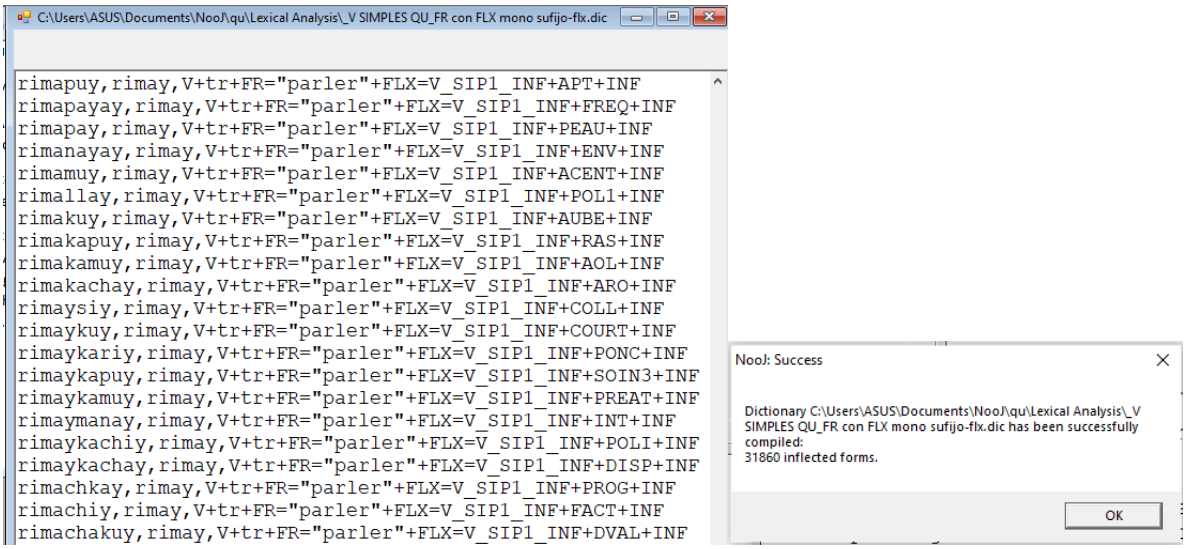

Figure 4 Derived verbs with one layer of IPS suffixes

Do all of these derived "new verbs" have an actual meaning? It was answered in the affirmative by Duran (2017), who translated 9,000 entries of the Dubois & Dubois-Charlier French verbs dictionary.

2.2.2 Mixed verbal phrases.

A remarkable property during the inflection of any verb is the behavior of present tense PR-ENDINGS; they act as fixed points around which IPS or PPS suffixes may be agglutinated to obtain a more complex word as a mixed verbal form (which may represent a long phrase in an Indo-Europeans language). The following examples illustrate this property for the *ni* ending:

Miku-**ni** (I eat)
Miku-chka-**ni** (I am eating)
Miku-chka-**ni**-raq (I am eating before anything else)
Miku-chi-chka-***ni***-raq-mi (I am eating before anything else indeed)
Miku-chi-yku-chka-***ni***-lla-raq-mi (I am carefully helping him to eat before anything else, indeed)

To formalize this important property, I propose the following expression:

<V><IPS><PR ENDING><PPS>, where:

- V: Verb stem *rima* (comes from *rimay*/ to talk)
- IPS[11]: Inter-posed suffixes (placed between the verb stem and the ending),
- ENDING[12]: PR endings or their transformations
- PPS[13]: Postposed suffix (placed after the ending)

The PR- ENDING is the set of seven present tense personal endings (which will behave as fixed points during the inflections). This set may be topologically transformed into nine sets as detailed in Duran 2017 (to obtain the gerunds or other aspect forms).

Examples:
*rima-**nki***, 'you talk', present tense 1+s
*rima-ri-**nki***, 'you start talking', the IPS suffix *-ri* is interposed
*rima-**nki**_man*, 'you should talk', the PPS suffix *-man* is postposed
rima-ri-***nki***-man, 'you should perhaps start talking', the IPS suffix *-ri* and the PPS suffix *-man* are mixed.

In these examples, each class of suffix intervenes in the inflection only once at both sides of the ending. But, the Quechua grammar allows having several layers of combinations of IPS and PPS. Using the NooJ format, we have been able to program the following paradigms of mixed agglutinations:

```
V_MIX1=(:SIP1_PR_V) (:SPP1_V) | (:SIP1_PR_C)(:SPP1_C)
        | (:SIP1_PRM_V ) (:SPP1_V) | (:SIP1_PRM_C)(:SPP1_C);
```

Ex. *Miku-chi-**ni**-raq*

```
V_MIX12=:SIP1_PR_V)(:SPP2_V)|(:SIP1_PR_C)(:SPP2_C) |(:SIP1_PRM_V )
(:SPP2_V)|(:SIP1_PRM_C)(:SPP2_C);
```

Ex. *Miku-chka-**ni**-raq-mi*

```
V_MIX21= (:SIP2_PR_V)(:SPP1_V)|(:SIP2_PR_C)(:SPP1_C);
```

Ex. *Miku-chi-chka-**ni**-raq*

```
V_MIX22= (:SIP2_PR_V)(:SPP2_V)|(:SIP2_PR_C)(:SPP2_C);
```

Ex. *Miku-chi-chka-**ni**-raq-mi*

Where V_MIX12 stands for mixed verbal agglutination of 1 IPS and 2 SPP
SIP1_PR_V: Derivation using 1 IPS and conjugation following the PR scheme, etc.

Applying these grammar rules for the verb *rimay*/to talk generates 289 413 mixed verbal forms.

[11] IPS={*chaku, chi, chka, ykacha, ykachi, ykamu, ykapu, ykari, yku, ysi, kacha, kamu, kapu, ku, lla, mpu, mu, na, naya, pa, paya, pti, pu, ra, raya, ri, rpari, rqa, rqu, ru, spa, sqa, stin, tamu, wa}*
[12] ENDINGS = {ni, nki, n, nchik, niku, nkichik, nku, … }
[13] PPS={*ch, chaa, chik, chiki, chu(?), chu, chusina, má, man, m, mi, ña, pas, puni, qa, raq,ri, si, s, taq, yá*}

2.2.3 The LVF_QU dictionary.

LVF_QU is a bilingual electronic dictionary containing around 8 600 French verbs, which I translated to Quechua from the Dubois-Dubois Charlier LVF dictionary (2007). Consequently, it served me to build the QU_LVFQ (Quechua-French) dictionary.

2.2.4 Automatic translation of derived verbs.

As we said in 2.2.1, the basic simple-verb dictionary contains around 1200 QU-Fr verbs, and applying the V_TR_INF grammar, we have generated around 31 800 derived verbs. Some of the resulting new verbs are already lexicalized in some printed dictionaries, but most of them do not have written translations. I have been working on the automatic formal translation of them into French and Spanish. I have translated around 7 000 derived verbs as part of this project. Example: the simple verb *asiy*/ to laugh, when derived by the "*ri*" IPS suffix, gives the new verb *asiriy*/ to smile, which appears in some written dictionaries, whereas the derived verb *asichakuy*/, derived by the "*chaku*" IPS suffix, is not present in any of those dictionaries, it means *to laugh ridiculously*. Other examples of derived verbs which do not appear in printed dictionaries: ***ripuy/*** to leave > *ripu-**ku**-y*/ to move; ***rakiy/*** to split> *raki-**naya**-y/* to divorce; ***samay/*** to rest > *sama-**rqu**-y* / to bivouac.

2.3 Formalizing adjective inflection

The formalization of the adjective inflection and derivation also consists in describing the paradigms of grammatical agglutinations of the adjectival suffixes Suf_A:

Suf_A = {-ch , chá, -cha ,-chik, -chiki, -chu, -chu?, - hina, -kama, -kuna, -lla, -má, -man, -manta, -masi, -m, -mi, -naq, -nka, -ninka, -nta , -ninta , -nintin, -ntin, -niraq, -niq, -ña, -p, -pa, -paq, -pas, -pi, -puni, pura, -qa, -rayku, -raq , -ri, -s, -si, -su, -ta, -taq, -wan, -yá, -yupa }

I present a portion of the table that includes the Boolean matrix of its corresponding bi-suffix combinations in Figure 5.

	Suff_A	CH	CHAA	CHIK	CHIKI	CHUI	CHUN	DCHA	GEP	GEPA	HINA	KAMA	KUNA	LLA	MAA	MAN	MANTA	MASI	MI	MM	MP
2	CH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	CHAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	CHIK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	CHIKI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	CHUI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	CHUN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	DCHA	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	
9	GEP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	GEPA	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	0	1	
11	HINA	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	1	0	0	1	
12	KAMA	1	1	1	1	1	1	0	0	1	0	0	0	1	1	0	0	0	0	1	
13	KUNA	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	0	1	
14	LLA	1	1	1	1	1	1	0	0	1	1	1	0	0	1	1	1	0	0	1	
15	MAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	MAN	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	
17	MANTA	1	1	1	1	1	1	0	0	0	1	1	0	0	1	0	0	0	0	1	
18	MASI	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	0	0	1	
19	MI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	MM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	MPA	1	1	1	1	1	1	0	0	1	0	0	1	1	0	0	0	0	0	0	

Figure 5. Matrix of bi-suffix combination of Suf_A

The general grammar to inflect or derive an adjective ending in a vowel is AVOCAL = :A_V_1 | :A_V_2 | :A_V_3; The first component is the paradigm involving only one suffix, the second one two suffixes, and the last three suffixes. The grammar A_V_2 comes from the matrix of Figure 5 and looks like the following:

```
A_V_2 = :DCHACH| :DCHACHAA| :DCHACHIK| :DCHACHIKI| :DCHACHUI | :HINACHIKI |
:HINACHUI | :HINACHUN | :HINAGEPA | :HINAKAMA | :HINALLA  | :HINAMAA |
:HINAMAN  | :KAMACHUI | :KAMACHUN | :KAMAGEPA | | :LLANIRAQ | :LLAñA  |
:LLAPAQ | :LLAPAS | :LLAPI | :LLA | :LLA| :LLAPUNI| :LLAPURA| :LLATA |
:NIRAQKAMA| :NIRAQLLA| :NIRAQMAA| :NIRAQMAN | :NIRAQMANTA | :NIRAQMI  |
:NIRAQWAN  | :NIRAQYAA| :ñAMAA | :ñAMM  | :ñARI | :ñASISV | :ñATAQ | :ñAYAA
| :PAQCHAA | :PAQCHIK | :PAQCHIKI| :PUNITA  | :PUNITAQ | :PUNIWAN  |
:PUNIYAA| :PURACH| :PURACHAA | :PURACHIK | :TACH | :TACHAA  | :TACHIK  |
:TACHIKI | :TACHUI | :TACHUN  | :TAñA | :TAPAS | :TAPUNI | :TAQA | :TARAQ
| : . . .  | :YUPAQA | :YUPARI | :YUPASISV  | :YUPAWAN | :YUPAYAA;
```

And for the adjectives ending in a consonant or in the particle *ai* is:

ACONSO = :A_C_1 |:A_C_2 |:A_C_3; the first is the paradigm involving only one suffix, and the last two components originate from similar matrices as of Figure 5 manually constructed.

2.4 Formalization of Adverb inflection

To formalize the adverb inflection and derivation, we also need to build the Boolean matrices of combinations of two or three of the adverbial suffixes Suf_ADV

> Suf_ADV= (ch , chaa, chik , chiki, chun, chui, kama, lla, maa , manta, m, mi, ña, paq, pas, pi, puni, qa, hina, raq , ri , sisc, siv, nta, ninta, taq, wan, yá),

And then put them in formal expressions the corresponding paradigms:

> ADV_V=<E>/ADV |:ADV_V_1 |:ADV_V_2 |:ADV_V_3; for adverbs ending in a vowel
> ADV_C=<E>/ADV |:ADV_C_1 |:ADV_C_2 |:ADV_V_3; for adverbs ending in a consonant.
> Their first component, for the vowel and the consonant cases, looks as below:

```
ADV_V_1 = :CH | :CHAA | :CHIK | :CHIKI | :CHUN | :CHUI | :KAMA | :LLA | :MAA
| :MAN | :MANTA | :MM  | :ñA  | :PAQ | :PAS | :PI | :PUNI | :QA | :HINA |
:RAQ | :RI | :SIV | :NTA | :TAQ | :WAN | :YAA;
```

```
ADV_C_1 = :CHAA  | :CHIK  | :CHIKI  | :CHUN  | :CHUI | :KAMA  | :LLA | :MAA
| :MAN | :MANTA | :MI  | :ñA  | :PAQ | :PAS | :PI  | :PUNI | :QA  | :HINA |
:RAQ | :RI | :SISC | :NINTA | :TAQ | :WAN | :YAA;
```

And the special inflection formula for the adverb of negation *mana/non*

> ADV_MANA = :CH |:CHAA |:CHIK |:CHIKI |:CHUI |:MAA |:MM |:ñA |:PAS |:PUNI |:RAQ |:RI |:SIV |:TAQ |:YAA;

A partial list of 733 bi-suffixed generated adverbial forms looks as follows for the adverb *paqarin*/tomorrow:

```
paqarinhinach,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+CMP+DINT
paqarinhinachá,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+CMP+DINT
paqarinhinachu?,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+CMP+ITG
paqarinhinakama,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+CMP+MET
paqarinhinalla,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+CMP+ISO
paqarinhinaman,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+CMP+DIR
paqarinhinam,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+CMP+ASS
paqarinhinaqa,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+CMP+THE
paqarinkamachu,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+MET+NEG
paqarinkamalla,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+MET+ISO
paqarinkamamá,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+MET+CTR
paqarinkamam,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+MET+ASS
paqarinkamaña,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+MET+TRM
```

```
paqarinpihina,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+LOC+CMP
paqarinpiwan,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+LOC+INS
paqarinpiya,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+LOC+IVOC
paqarintawan,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+ACC+INS
paqarinyupata,paqarin,ADV+EN=tomorrow+FLX=ADV_C_2+CPR+ACC
```

The electronic dictionary of QU_FR adverbs contains 565 indexed Quechua-French adverbs.

2.5 Formalization of Pronoun inflection

Pronouns also can be inflected and derived in Quechua, i.e., *Qampas*/ you also, *paypaq*/ for him, *qamlla*/ only you, etc.

To formalize pronoun inflection and derivation, I follow similar steps that I have applied for the precedent POS: Make an inventory of the set of pronominal suffixes Suf_PRO, construct tables and matrix calculations reflecting the grammatical combinations of two or three suffixes, deduce from these values the corresponding paradigms of inflection of pronouns ending in a vowel or a consonant, construct the formal grammars (samples presented below) for the NooJ platform, parse the dictionary of 565 simple and composed pronouns and so, obtain more than 117 000 inflected or derived pronominal forms as shown in Figure 6, automatically generated by the following grammars inflecting singular or plural pronouns (ending in a vowel).

```
PROVOCAL_S =<E>/PRO  | :PRO_V_S_1  | :PRO_V_2  | :PRO_V_3;
PROVOCAL_P =<E>/PRO  | :PRO_V_P_1  | :PRO_V_2  | :PRO_V_3;
where
PRO_V_S_1 = :CH | :CHAA | :CHIK | :CHIKI | :CHUN | :CHUI | :CHUSINA  | :GEP
| :GEPA | :HINA | :KAMA | :LLA | :MAA | :MAN | :MANTA | :MM   | :NAQ |
:NTIN | :NIRAQ | :ñA | :NIQ | :PAQ  | :PAS | :PI | :PUNI  | :QA | :RAIKU |
:RAQ | :RI | :SSIV | :TA | :TAQ | :WAN | :YAA | :YUPA;
```

In Figure 6, at left appears the electronic dictionary of QU_FR pronouns, and at the right, an extract of the inflected forms as pronominal phrases, and at the bottom, the total inflected pronominal forms automatically generated.

Figure 6 QU-FR pronouns and the corresponding 117 000 inflected forms

I have been working on the automatic formal translation (QU-SP, QU-FR) of the generated inflected POS: Nouns, Verbs, adjectives, pronouns, and adverbs. Example: the simple adjective *yuraj*/white

gives rise to the inflected form *yurajniraj*/ is automatically translated as “near to white (pale),” which is correct.

3 Perspectives: Automatic translation of the POS-inflected linguistic units.

As I have presented here, the dictionaries and the results obtained by parsing them using the numerous inflection paradigms and grammars represent several millions of new Quechua tokens. Most of them do not have their Spanish or French translations. For the moment, I work with different software to build, on the one hand, all the conjugated Quechua verbs with the corresponding French conjugated forms, and thereafter the corresponding Spanish conjugated verbs, and then to obtain extended bilingual QU-FR and FR-QU vocabularies. Up to now, I have gotten around 400 000 conjugated translated forms.

References

Cardenas R., Zevallos R., Baquerizo R. and Camacho L., 2018, *Siùinchik:A speech Corpus for Preservation of Southern Quechua. lrec-conf.org. http://lrec-conf.org › workshops › lrec2018.*

Dubois, J. et al., 2007 Dictionnaire Linguistique et Sciences du langage, Editions Larousse, Paris.

Duran, Maximiliano, 2012: *Formalizing Quechua verbs Inflexion*, Proceedings of the NooJ 2013 International Conference, Saarbrücken. Cambridge Scholars.

Duran, Maximiliano, 2014: *Morphological and syntactic Grammars for Recognition of Verbal Lemmas in Quechua*. Proceedings of the 2014 International Conference and Workshop. Sassari.

Duran, M.: *Dictionnaire électronique français-quechua des verbes pour le TAL*., 2017 : Thèse Doctorale.

Duran Maximiliano,2021: *Morfología y diccionario electrónico de nombres en Quechua March* DOI: 10.35305/an.vi1.4 Conference: Proceedings of the Linguistic Resources for Automatic Natural Language Generation .
Université de Franche-Comté. Mars 2017 (2017)

Guardia Mayorga, César, *Gramatica Kechwa*, Ediciones Los Andes. Lima Peru, 1973.

Monson, C., Llitj´os, A. F., Aranovich, R., Levin, L., Brown, R., Peterson, E., Carbonell, J., and Lavie, A. 2006. *Building NLP systems for two resource-scarce indigenous languages: Mapudungun and Quechua. Strategies for Developing machine translation for minority languages.* aclanthology.org. https://aclanthology.org › LREC-2006-Monson

Perroud, Pedro Clemente,1970*: Diccionario castellano kechwa, kechwa castellano. Dialecto de Ayacucho. Santa Clara, Peru.* Seminario San Alfonso.

Pino Duran, A. German,1980*: Uchuk Runasimi (Jechua - Quechua).* Conversación y vocabulario Castellano-Quechua Ocopa, Concepción Perú.

Rios, Annette., 2011: *Spell checking an agglutinative language Quechua*. The University of Zurich. Zurich Open Repository and Archive.

Rios, Annette. 2016. A basic language technology toolkit for Quechua. A Basic Language Technology Toolkit for Quechua. Thesis. DOI: 10.5167/uzh-119943.

Silberztein M, La formalisation du dictionnaire LVF avec NooJ et ses applications pour l’analyse automatique de corpus. Langages 3/2010 (n° 179-180), (2010), pp. 221-241.

Silberztein M, La formalisation des langues. ISTE Editions. London, (2015).

Silberztein M, NooJ Manual. http:/www.nooj4nlp.net (220 pages, updated regularly), (2003).

Soto Ruiz, Clodoaldo, 1976: *Gramática quechua: Ayacucho-Chanca*. Lima: Ministerio de Educación, Instituto de Estudios Peruanos. 184 p.

Zevallos, R., et al., 2022: *Introducing QuBERT: A Large Monolingual Corpus and BERT Model for Southern Quechua*. Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language. Association for Computational Linguistics