QUECHUA-SPANISH AND SPANISH-QUECHUA ELECTRONIC DICTIONARIES OF VERBS FOR NLP

Duran Maximiliano

[1] Université de Franche-Comté Besançon, France
duran_maximiliano@yahoo.fr

Abstract. The automatic processing of the Quechua-Spanish pair of languages (APQUSP) needs an electronic dictionary of Quechua-Spanish verbs and Spanish-Quechua.

This article presents advancements in constructing an electronic dictionary of both Quechua-Spanish Spanish-Quechua verbs.

Among the few Spanish monolingual digitalized dictionaries, I have based my work on the Spanish dictionary in the Nooj Module of Spanish Argentine-Chile, initially built at the Universidad Autónoma of Barcelona. It contains 8122 verbs and is entirely compatible with Nooj's formalism developed by Silberztein [12]{13]. A significant difficulty in translating it into Quechua is that the lexicon of simple verbs of the latter contains around 1,600 entries. How to match 8122 Spanish verbs with only 1,600 Quechua verbs?

In this work, I show how I use the remarkable Quechua strategy of generating new verbs by suffix derivation to answer this question. As a first step, I have inventoried all the Quechua suffixes allowing verb-verb derivation and detailed their corresponding Spanish semantic values. This set of suffixes, which I call IPS_DRV, contains 27 elements. Each Quechua verb gives rise to at least 27 derived verbs. The resulting 43,200 atomic linguistic units (CALU) will allow us to obtain most of the Quechua transfers of the SP verbs. The resulting electronic dictionary includes relevant semantic tags for each Quechua verb and its corresponding inflection grammar.

To complement this work, I present, in the end, the obtained Spanish-Quechua dictionary containing more than 16,000 Spanish verbs and phrasal verbs.

.

Keywords: Quechua-Spanish electronic dictionary, Verbal suffixes, Formalization of Quechua.

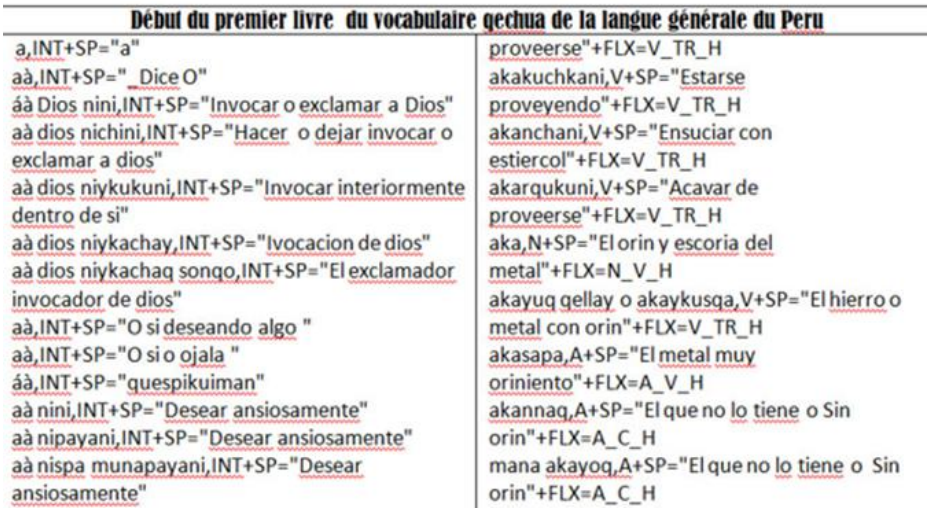
1 An electronic dictionary of verbs can help the NLP of Quechua

The automatic processing of the Quechua-Spanish pair of languages (APQUSP) needs an electronic dictionary of Quechua-Spanish Spanish-Quechua verbs.

Because QU remains mainly an oral language, the few written bilingual dictionaries (QU-SP) present several obstacles: lack of POS tagging for the entries; many nouns,

verbs, and adjectives are not presented in their lemmatized form but somewhat partially inflected; many lemmas are presented in incoherent orthography among the different dialects, etc.

One of the first and most comprehensive bilingual dictionaries is by Diego Gonzales Holguin [6], richer than Fray Thomas d'Aquino [5]. It was published in 1608 in Lima, Peru, under the title: "*Vocabulario de la Lengua General de todo el Peru llamada lengua Qquichua o del Inc*a." It contains 13,158 Quechua entries. I have, first, translitterated the San Marcos University 1962 edition Holguin [6]) into the Ayacucho-Chanka official orthography and put it in a digitalized format Fig. 1.



Début du premier livre du vocabulaire qechua de la langue générale du Peru

a,INT+SP="a"
aà,INT+SP="_Dice O"
áà Dios nini,INT+SP="Invocar o exclamar a Dios"
aà dios nichini,INT+SP="Hacer o dejar invocar o exclamar a dios"
aà dios niykukuni,INT+SP="Invocar interiormente dentro de si"
aà dios niykachay,INT+SP="Ivocacion de dios"
aà dios niykachaq sonqo,INT+SP="El exclamador invocador de dios"
aà,INT+SP="O si deseando algo "
aà,INT+SP="O si o ojala "
áà,INT+SP="quespikuiman"
aà nini,INT+SP="Desear ansiosamente"
aà nipayani,INT+SP="Desear ansiosamente"
aà nispa munapayani,INT+SP="Desear ansiosamente"
proveerse"+FLX=V_TR_H
akakuchkani,V+SP="Estarse proveyendo"+FLX=V_TR_H
akanchani,V+SP="Ensuciar con estiercol"+FLX=V_TR_H
akarqukuni,V+SP="Acavar de proveerse"+FLX=V_TR_H
aka,N+SP="El orin y escoria del metal"+FLX=N_V_H
akayuq qellay o akaykusqa,V+SP="El hierro o metal con orin"+FLX=V_TR_H
akasapa,A+SP="El metal muy oriniento"+FLX=A_V_H
akannaq,A+SP="El que no lo tiene o Sin orin"+FLX=A_C_H
mana akayoq,A+SP="El que no lo tiene o Sin orin"+FLX=A_C_H

Fig. 1. Digitalized Holguin's dictionary

Looking at verbs, many of them appear conjugated to the present first person, even though their Spanish transfers are put to the infinitive, i.e.
akanchani, ensuciar con estiércol (to make dirty)
*wayllu**ni**,Amar con muestras de affecto amoroso y tierno (to love)*
*rima**ni**, hablar (to talk)*
*puri**ni**, andar, caminar (to walk)*
*miku**ni**, comer, alimentarse(to eat)*
*taki**ni**, cantar (to sing)*

In order to extract the verb root from the inflected verb ending in the *–ni* suffix, I have applied to the digitalized dictionary the following filter using NooJ formalism,
NI_q_extr == Find/ Replace (PERL pattern, ni$|q$, extract lines)
Find / Replace (PERL pattern, ni$, filter-out lines)

In this way, I arrived to extract 3999 inflected verbal forms. This does not mean we have this quantity of simple verbs in the Holguin dictionary. This list contains many redundancies. For example, for the verb kay "to be," the form ka-ni appears 68 times. Therefore, we needed to program additional filters such as VOC-ni_q, NI_q_extr

(VOC-H_brut) to obtain a set of distinct verbal forms (636) that potentially contain verbal stems. However, they still contain different derivation suffixes for the same verbal stem. After applying new filters, we could isolate all the verbal roots in the Holguin dictionary. Surprisingly, there are barely 95 of them, and they give just as many simple verbs. On the other hand, the dictionary contains many other lexemes showing other types of inflected forms like *chakirqun* ("it has just dried"), which contains the verb stem *chaki-* inflected by a combination of the suffix *–rqu* and the ending *–n* of the third person to the present time. Notice that the corresponding simple verb *chakiy* "to dry", could not appear in the list of the 95 verbs because it does not contain the suffix *–ni*.

Furthermore, to unveil the verb radicals appearing as nouns, resulting from the derivation of verbs by the nominalizing suffixes -q, -na, -sqa or the combination -rqu-n , which Holguin uses as an aspectual mark for the third person, we programmed filters like the following:

Find/ Replace (PERL pattern, na$, extract lines)
Find/ Replace (PERL pattern, q$, extract lines)
Find/ Replace (PERL pattern, sqa$, extract lines)
Find/ Replace (PERL pattern, rqun$, extract lines)

This allowed us to get 531 nominalized verbal radicals, 327 participles in sqa, 33 agentive forms with q, and 181 forms of past tense 3+s.

To identify the verbs that appear as verbalizations of nouns, we use the verbalizing suffixes -chay, -yay, -y as variables and apply cascading NooJ queries to the Holguin dictionary such as:

Find/ Replace (PERL pattern, cha$, extract lines)
Find/ Replace (PERL pattern, ya$, extract lines)
Find/ Replace (PERL pattern, y$, extract lines)

At the end of these operations, we obtained less than 500 simple verbs, a modest harvest. To continue the search, I studied other sources, among them some contemporaneous ones like Guardia Mayorga's [7], Peroud's [9], and Middendorf's [8].

(1) Guardia Mayorga presents their verbs as follows:

wankuy; vendar, envolver, amancornar.
waqaychay, guardar.
wachay, parir.
waqay,llorar. Fig. dar frutos.
wanquy, palanquear.
waykay, pegar a una persona entre varios.
without any POS tags.

(2) Perroud's verbs look like these:

chutay. 1.Estirar, jalar, extender, tirar de, arrastrar, apretar
2.desencoger, estirarse, desperesarse 3.largarse

It does not contain any tags either, and translates the infinive *chutay* also as if it were in pro nominalized form, which should appear as the derivation *chutakuy*. This presentation suggests the need to add semantic tags like the domain of use, the context of the verb, the sense, etc., for disambiguation operations.

(3) Middendorf's looks like this:

yuyay==pensar
churay==poner, colocar
iñiy== consentir, creer

It may contain several meanings and without any tags

Gathering all the verbs from these documents and my own introspection, I have obtained 1631 simple verbs with single meanings, a sample of which is shown in Fig 2.

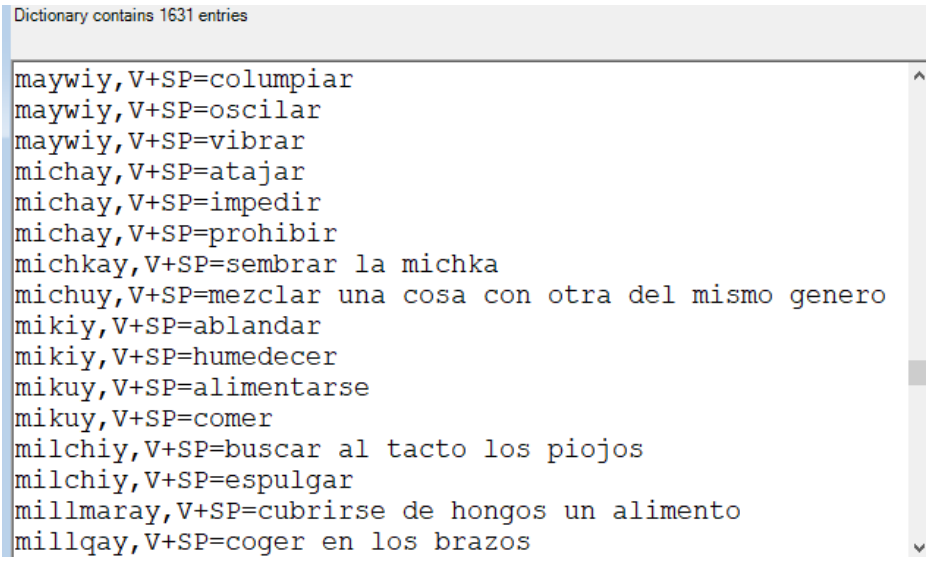


Fig. 2. Raw QU_SP simple verb dictionary

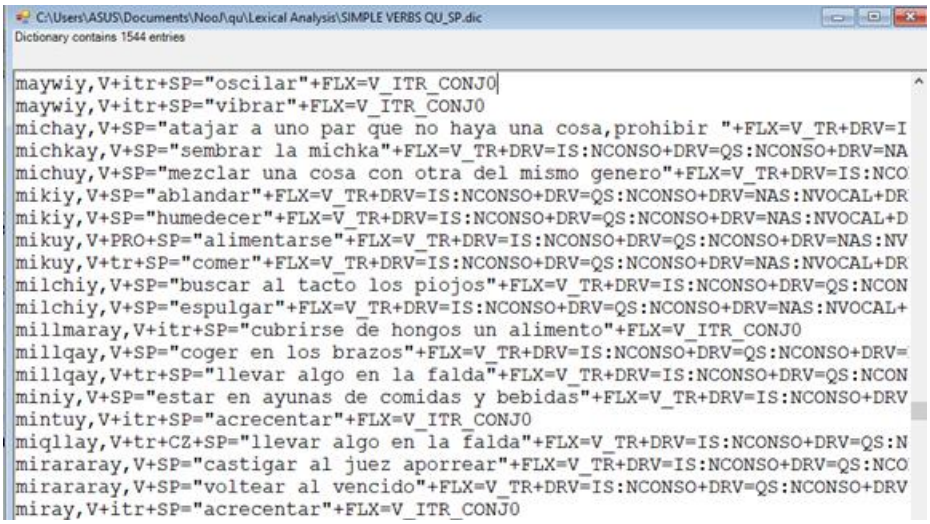


Fig. 3. The electronic version of QU_SP simple verb dictionary

In Fig. 3, we show the same dictionary but include semantic and morpho-syntactic tags and each entry bearing its corresponding inflection and derivation grammar.

Nonetheless, having only 1544 verbs translated, we need to search deeper into the language morphology to answer the initial problem of how to get the translation of the rest of the thousands of Spanish verbs remaining without their Quechua transfers.

2 Quechua morphology helps to generate new verbs

We have already met some verbs obtained by derivation of a simple verb, i.e. *akakuchkani* (estarse proveyendo), or *chakirqun* (it has just dried). They come from Holguin's, the verb *akay* (defecar) derived using the *–chka* suffix and the second using the suffix *–rqu*. As we have remarked, that dictionary contains 3999 derived verbs, among which only 91 are simple, and the remaining majority are derived. The following examples show some other verb suffixes that produce new verbs out of the verb *mas-kay*/to search:

maska-ku-y to search for oneself (rebuscarselo)
maska-yku-y to search with determination (buscar cuidadosamente)
maska-pa-y to glean (rebuscar)
maska-ri-y to search superficially (buscar superficialmente)
maska-chi-y to make someone search (ordenar la búqueda)
maska-paya-y to do research (investigar)

We may obtain up to 27 new verbs, which suggest the existence of a set of specialized suffixes allowing us to get V-V derivation, which may help us considerably enhance the initial verb lexicon.

I have inventoried all the verbal suffixes intervening in the verb inflection. I have proposed in Duran [3] [4] four classes[1] as follows: the conjugation endings (CE), Interposition suffixes (IPS), Postpositional suffixes (PPS), and the set of verb nominalizers (VN_S).

During the conjugation of a verb, there is a specific class of suffixes that are interposed between the verbal root and the conjugation ending, this Interposed suffixes (IPS) are: IPS = {-chaku, -chi, -chka, -kacha, -kamu, -kapu, -ku, -lla, -mpu, -mu, -naya, -pa, -paya, -pu, -rqa, -raya, -ri, -rpari, -rqu, -ru, spa, -sqa , -tamu, -wa, -ykacha, -ykachi, -ykamu, ykapu, -ykari, -yku, -ysi} (27)

We may express this transformation by the formula:

Dips(V1root) = V1root+ips+y=V2

This means that the function Dips applied to the verb V1 will generate the verb V2

Examples :

asiy « laugh »

Dri(asi) = asi+ri+y = asiriy

asiriy « to smile »

samay « to rest »

Drqu(ripu) = rsama+rqu+y = samarquy

samarquy « to bivouac »

Or for the verb maskay/to search we will have:

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

More explicitly, these IPS, are agglutinated to a verb radical, according to the formalized paradigm:

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) (y/INF) and, as we have mentioned, will generate 27 new verbs containing 1 IPS suffix as shown in the following sample:

mikurquy,mikuy,V+PRO+SP="alimentarse"+FLX=V_TR_INF+PAPT+INF
mikurpariy,mikuy,V+PRO+SP="alimentarse"+FLX=V_TR_INF+ASUR+INF
mikuriy,mikuy,V+PRO+SP="alimentarse"+FLX=V_TR_INF+DYN+INF

---

[1] The Interposed suffixes (IPS): appearing between the verbal root and the conjugation ending.

There is also suffixes that appear after the ending of a conjugation, I call them the Postposed suffixes (PPS): PPS ={- ch, -chá, -chik, -chiki, -chu, -chu(?), -chusina, -m, -mi, -má, -man, -ña, -pas, -puni, -qa, -raq, -s, -si, -taq, -yá} (20)

We find also a set of suffixes which nominalize a verb, we call it the Nominalizing suffixes (S_VN): S_VN = {-i, -na, -q, -sqa } (4)

The set of conjugation endings (CE) include the ones used in the undefined tense and those used for conjugation to the future:

S_DN = {i, nki, n, nchik, niku, nkichik, nku, saq, nki, nqa, sunchik, saqku, nkichik, nqaku} (14)

More generally, when we apply the grammar SIP1_IT_CONJ to the set of one thousand six hundred simple verbal roots, we obtain 36,622 new derived verbs, as shown in Fig 4, among which we will find many transfers of the Spanish verbs we seek.

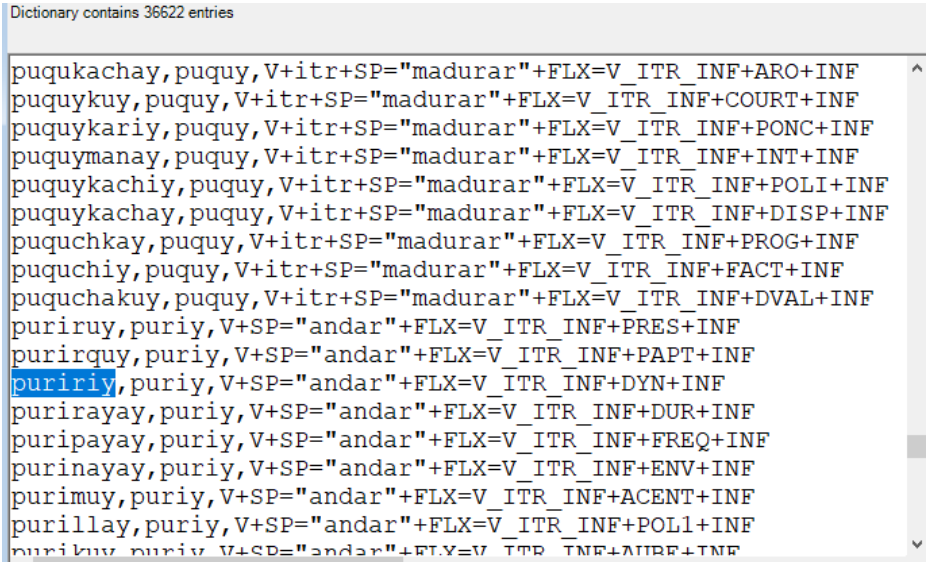


Fig. 4. Dictionary of mono-IPS derivation applied to the set of simple verbs (36 222)

Another source of new verbs is obtained from a much larger set of derived verbs obtained by the agglutination of combinations of two IPS. The IPS can combine with each other, following the Boolean matrix proposed by Duran [4]. When agglutinated to the original 1600 simple verbal roots, the grammatical combination generates 268,827 new derived verbs, among which we may still find many transfers for the composed Spanish verbs. See a sample in Fig 5.

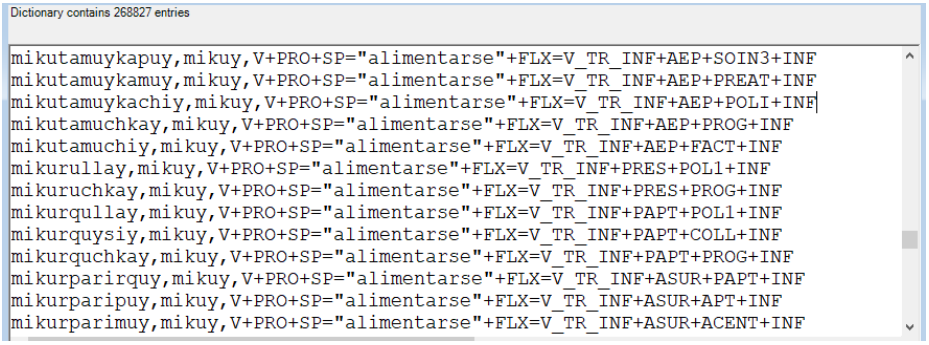


Fig. 5. QU-SP derived verbs bearing two IPS (268,827)

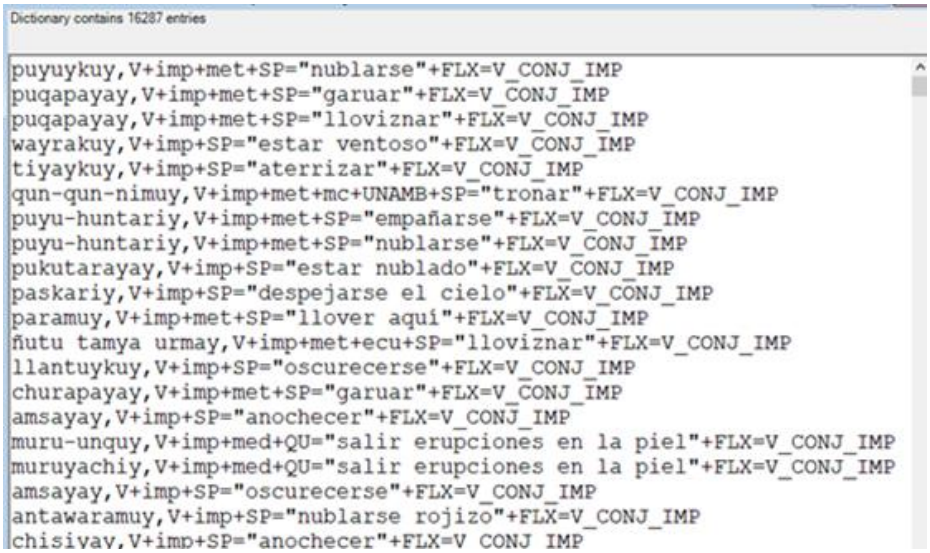


Fig. 6. The QU-SP electronic verb dictionary contains 16,287 entries.

3 From the electronic French-Quechua *verb* dictionary *to* the SP-QU dictionary

The automatic processing of the Quechua-Spanish pair of languages (APQUSP) needs an electronic dictionary of Spanish-Quechua-Spanish verbs.

The QU_SP dictionary shown in Fig 2 and those appearing in the references helped us build the first Quechua-Spanish dictionary containing 5600 simple and phrasal verbs.

To enhance it, I have taken profit from a previous work Duran [2],[3] [4], where I had manually obtained the translation of 8677 French verbs to Quechua, coming from de Dubois and Dubois French dictionary |1]. Using a specific algorithm, I used the existing internet data for translating the FR verbs to Spanish and then put them online with the corresponding Quechua transfers.

After this step, to continue with the translation of more QU (contained in the derived 36,622 new derived verbs), I followed roughly the following scheme:

1. If I start with a simple verb like *asiy*/ to laugh, one of its derivations is obtained using the IPS suffix « ri », it gives the new verb *asiriy*/ to smile, which appears in some written dictionaries. In contrast, 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 that do not appear in printed dictionaries but I had to deduce from the semantics of the employed suffix like: *ripuy*/ to leave derives into *ripu-ku-y*/ to move; *rakiy*/ to split > *raki-naya-y*/ to divorce; *samay*/ to rest > *sama-rqu-y* / to bivouac

It has been a long manual work: In over a decade,to finally obtain the Quechua-Spanish electronic verb dictionary containing 16,287 verbs shown in Fig 5.

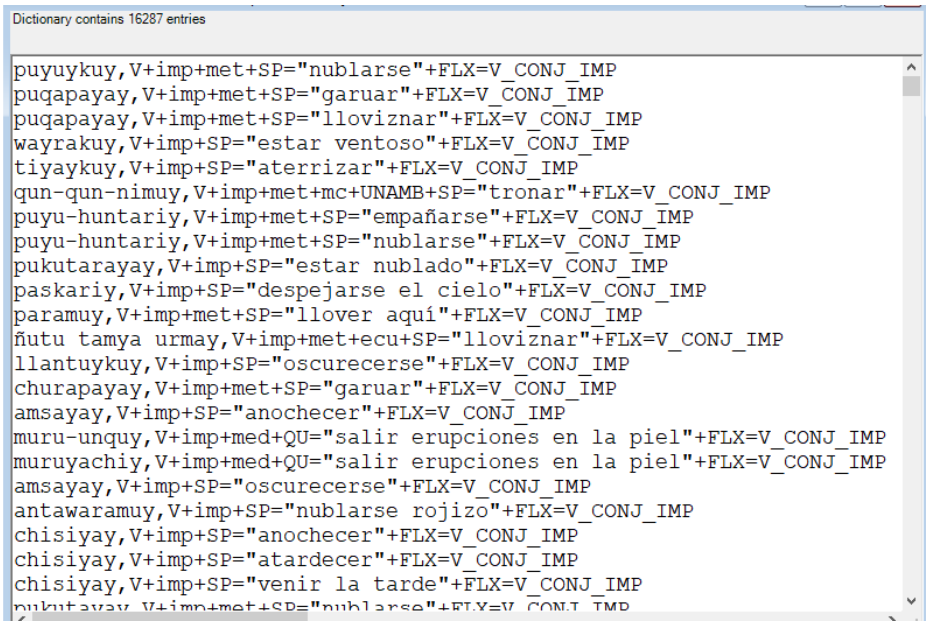


Fig. 7. The Quechua- Spanish electronic verb dictionary containing 16 287 verbs

4 Electronic Spanish verbs

In order to enlarge our linguistic resources aiming for our MT project, I have tried to build the corresponding Spanish-Quechua electronic dictionary of verbs. For this, I needed to choose a Spanish monolingual electronic dictionary.

Few open-source mono-lingual Spanish dictionaries possess an electronic configuration. In her works, A. Rios [11] makes use of an electronic Spanish-Quechua dictionary using, for the first, a set of PoS tags coming from the Royal Academy of Spanish and, for the se second, a set of tags corresponding to the Cuzco variant of Quechua. In our case, we work with the de Ayacucho Quechua variant.

Applications online, like Google or other recent translation programs, probably have constructed their own, but we cannot use them because they are not open source.

Fortunately, within the Nooj platform library, we find, imbedded in the Spanish module, the electronic dictionary RAE (Real Academia Española) [10]. It was initially built at the Universidad of Barcelona and then implemented with particular lexemes

from the Argentine-Rioplatense and Chile variants of Spanish. It contains 68,783 entries and is entirely compatible with Nooj's formalism. I thus decided to take it as my starting reference.

Table 1. of corresponding equivalent tags between SP an QU

Abreviaciones seguido de su desripcion en SP y QU			
Equivalencias simbólicas esenciales entre el QU y el SP			
SP	QU	**FLEXIÓN en Castellano**	**FLEXIÓN en Quechua**
inf	INF	inf infinitivo	INF Infinitivo
ppio	GER1	ppio participio indefinido	GER1 gerundio flexionado
ger	GER	ger gerundio	GER gérondif -spa simultanée
pres	PR	pres presente	PR presente
ppp	PPA	ppp habia comido	mikusjani SQA_PR_PLU
pps	PRES	pps pasado simple cantaste	PRES pasado simple RU takirunki
fut	F	fut futuro	F futuro
cond	C	cond condicional	C conditional
pi	PASS	pi pretérito perfecto indefinido cantabas	PASS RA passé simple accompli cantabas
	PASSA	habia rapidamente	PASSA RQUSQA_PRM2
subj	SUBI	subj subjuntivo fut+subj	SUBI subjuntivo inperfecto SPAJA
subj	SUBIPAS	subj subjuntivo fut+subj	SUBI subjuntivo inperfecto SQA_PR_SS subj
subj	SUBI	subj subjuntivo fut+subj	SUBI subjuntivo inperfecto SQA_PR_SS subj
imp	IP	imp imperativo	IP imperativo directo
ppcomp	ppi	pretérito perfecto rja	PPI pretérito perfecto indefinido rja
1a	1+s	1a primera persona singular	1+s primera persona singular

Having made this SP monolingual dictionary as the base of this part of my work. After applying several algorithms to obtain the reciprocal transfers, I verified and corrected the result. I have obtained, up to now, the Quechua transfers of 11,368 Spanish verbs, as shown in Fig 8.

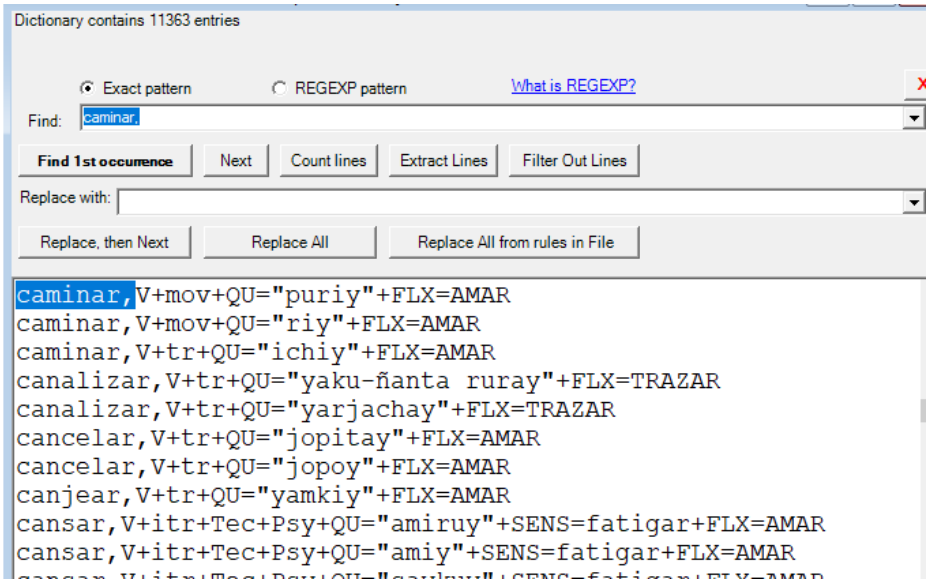


Fig. 8. Electronic Spanish-Quechua verb dictionary

We note that we have include in this dictionary simple verbs as well as phrasal Spanish verbs like:

andar a ciegas (taplaykachay a verb derived from taplay /) or

andar a paso lento (hapay).

Looking at the sample, we can observe:

1. Each entry includes several morphemes representing morpho-syntactic and semantic tags like V= verb, tr= transitive, itr= intransitive, mov= movement, …
2. The corresponding Quechua translation. In case of polysemy we put a line for each meaning, for example *caminar* (to walk) has three Quechua translations.

2. Each entry bears a flectional grammar FLX, allowing the entry to generate different inflections. A FLX grammar will generate all the verb conjugated and derived forms for the verb. For example, the FLX=AMAR appearing with the verb *caminar* has the following formula:

AMAR = (:TER | :TER1) | (<B2> (:TER2 | :TER3 | :TER4 | :TER5));

and each TER component looks like the following paradigm:

TER1 = (é/fut+ind+1a+sg | ás/fut+ind+2a+sg |á/fut+ind+3a+sg | emos/fut+ind+1a+pl |éis/fut+ind+2a+pl | án/fut+ind+3a+pl | ía/cond+ind+1a+sg | ías/cond+ind+2a+sg; etc.

Conclusion

We have arrived to build an electronic Quechua-Spanish dictionary each entry bears its specific inflection grammar and different semantic and morpho-syntactic tags.

In parallel, having the Spanish verb dictionary contained in the SP NooJ module as a base, I have enhanced it into an electronic Spanish-Quechua verb dictionary.

Considering our MT project, I am working on a system that helps us put the 1 467 000 Spanish conjugated verbal forms online with the corresponding Quechua transfers chosen from the 3 236 000 QU conjugated forms.

References

1. Dubois, J. & F. Dubois-Charlier. (a). Dictionnaire électronique des verbes français. Table Excel, 26.610 entrées. Téléchargeable sur http://www.modyco.fr/fr/Ressources/ldlvf.html. Version XML sur http://rali.iro.umontreal.ca/rali/?q=fr/lvf (1994).
2. Duran, Maximiliano, Diccionario Quechua-Castellano-Quechua, Editions HC. Paris, (2009).
3. Duran, Maximiliano, Formalizing Quechua verbs Inflexion. Proceedings of the NooJ 2013 International Conference, Saarbrücken, (2013).
4. Duran, M.: Dictionnaire électronique français-quechua des verbes pour le TAL. Thèse Doctorale. Université de Franche-Comté. Mars 2017 (2017).
5. Fray Domingo de Santo Tomas, Lexicom, o Vocabvlario de la Lengua General del Perv. Impresso en Valladolid par Francisco Fernandez de Cordova, (1560).
6. Gonçalez Holguin, Diego, Vocabulario de la Lengva General de todo el Perv llamada Lengua Qquichua o Lengua del Inca. Impresa en la Ciudad de los Reyes. Por Francisco del Canto, (1608).
7. Guardia Mayorga, César, Gramatica Kechwa, Ediciones Los Andes. Lima Peru, (1973).
8. Middendorf, Ernst W., Gramatica Keshua. Trad. d'allemand [Das Runa Simi oder die Keshua Sprache, wie sie gegenwärtig in der Provinz von Cusco gesprochen wird] par Ernesto More. Aguilar, Madrid, 1970.
9. Perroud, Pedro Clemente, Diccionario castellano kechwa, kechwa castellano. Dialecto de Ayacucho. Santa Clara, Peru. Seminario San Alfonso, (1970).
10. Real Academia Española: Asociación de Academias de la Lengua Española: Nueva Gtamática de la Lengua Española. Espasa-Calpe, Madrid, España (2009)
11. Rios (Annette Rios, Anne Gohring, and Martin Volk. 2008. A quechua-spanish parallel treebank. LOT Occasional Series, 12:53–64.
12. Silberztein Max, 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. (2010).
13. Silberztein Max, La formalisation des langues. ISTE Editions. London, (2015).