Towards Universal Dependencies For Ancash Quechua

Johanna Cordova
ERTIM / CERLOM, Inalco
2 rue de Lille, 75007 Paris, France
johanna.cordova@inalco.fr

Abstract

This paper presents a brief description of some morphosyntactic features of Ancash Quechua, the majority variety of the Central Quechua language family (QI), for the purpose of building a corpus annotated according to the Universal Dependencies (UD) schema. The creation of such a corpus has two objectives: for Quechua linguistics, it opens up the possibility of more systematic linguistic studies and comparisons with other languages. It also enables the development of the first syntactic parser for a Quechua language of this family. For the UD project, adding Quechua, an agglutinative language with a rich morphology, makes it possible to point out some possible shortcomings of the universal annotation schema, and to fuel the discussion to adapt this schema to the specific features of the languages with a similar typology. The first step towards this work was first to gather and digitise the available linguistic resources, thus creating the first bilingual and sentence-aligned digital corpus in Ancash Quechua and Spanish. After identifying some linguistic features not fully described in the UD schema, we proposed annotation solutions, and built an initial corpus of around fifteen sentences, which we are making freely available.

Keywords: low-resourced language, Universal Dependencies treebank, morphosyntactic annotation

1. Introduction

Quechua is the Amerindian language family with the largest number of native speakers, with almost 7 million speakers in 6 Latin American countries. Despite this large geographic coverage, Quechua languages remain minority languages, and their extreme diversity has received little attention in the field of NLP. The Universal Dependencies (UD) project proposes a generic morpho-syntactic annotation scheme, which is intended to be applicable to any language and as robust as possible in respect of the typological particularities of each language. For an under-resourced language, the constitution of such a corpus is a step forward for its expansion in the field of NLP by enabling the training and development of parsing tools. The universal nature of the annotations aims to enable systematic and large-scale comparisons between typologically close languages. In this paper we present the preliminary work for the building of a treebank for Ancash Quechua, which currently has almost no digital resources.

1.1. Ancash Quechua

The Quechua language family is divided into two branches, designated as Quechua I and Quechua II. Most NLP research initiatives have focused on varieties of the QII family, which is the largest in both number of speakers and geographical spread. We focus here on one of the QI varieties, spoken in the Central Andes in Peru, which display some more conservative features with respect to Proto-Quechua. A large scale comparison of corpora in several varieties of the two language families would make it possible to refine the classification of the different varieties of Quechua (following the example of (Heggarty, 2005)), and to find out new elements for the reconstruction of Proto and Pre-Proto Quechua (Emlen and Dellert, 2020). Ancash Quechua is the majority language of the Quechua I family, and is itself made up of several varieties. We focused here on the Huaylas Quechua (ISO-639-3: `qwh`), spoken by around 400 000 speakers in the Ancash region in Peru.

2. Previous work

A treebank for Southern Quechua (majority language of the QII branch) has previously been developed (Rios et al., 2008) and was included to the first version of Universal Dependencies project. However, it has not been ported to the second version of the annotation schema and is no longer in the official release of the project. Another similar work is that of (Pankratz, 2021), who developed a coreference corpus from oral data in Conchucos Quechua (a variety of Ancash Quechua close to the Huaylas variety). This corpus of 1.400 words is segmented by morphemes and labeled into parts of speech, but does not provide syntactic analysis. As far as we know, no digitised corpus of larger size for NLP tasks is available yet for Ancash Quechua, and no extensive linguistic description has been published, apart from the (sometimes incomplete) first grammars (Escribens and Proulx, 1970), (Swisshelm, 1972), (Parker, 1976). With this paper, we hope to lay some foundations for developing extensive and good quality linguistic re-

LREC-COLING 2024, pages 16680–16685
20-25 May, 2024.

sources for this variety of Quechua.

3. Treebank composition

The corpus is composed of sentences from the following sources:

1. *Cuentos y relatos en el quechua de Huaraz* (1975), two volumes of autobiographical chronicles and tales from the oral tradition by Santiago Pantoja Ramos, a Quechua native speaker from the province of Huaraz. The texts are bilingual Quechua-Spanish. This work is the only written production of such a scale in a subvariety of Ancash Quechua. We digitised it almost completely from a paper version and we aligned the Quechua sentences to their Spanish translation (4427 sentences, CRH).

2. *Apu Kolki Hirka* [Mountain's God of silver], by Macedonio Villafán Broncano (1997), the first short novel (or prose poetry) in Ancash Quechua. The text comes with Spanish and English translations. It has been provided to us by the author (434 sentences, AKH).

3. The Ancash Quechua dictionary issued by the Peruvian Ministry of Education for bilingual intercultural education (Menacho López, 2005) (879 entries, DIC-MINEDU-2005).

4. A set of tales published in low-circulation during the 1990s, from the *Cuentos Pintados del Perú* project (1994-2015) which aimed to support bilingual intercultural education by bringing together traditional tales from all over the world translated into Quechua (300 sentences, CP).

The spelling of the corpora was standardised according to the rules proposed by the Ministry of Education's working group for Intercultural Bilingual Education (MINEDU, 2020), and using the Quechua official alphabet.

4. Annotation guidelines

Quechua languages are agglutinative, with a suffixing strategy. They are mainly SOV and head-final languages, though the word order is highly flexible (Payne, 1993); moreover, Ancash Quechua tends to use SVO pattern more than other varieties. In this section we present some morphological and syntactic specificities of Ancash Quechua and our annotation proposals.

4.1. Lexical categories

For Ancash Quechua, we use all of the UD's 6 open class words: nouns, proper nouns, verbs, adjectives, adverbs and interjections. As in many agglutinative languages, the effective category of a root is mainly determined by the suffixes attached to it. Some roots can indeed be used as a nominal or verbal base without a derivational marker. Moreover, Quechua essentially uses non-lexical strategies to add adverbial information. For example, the adverbial form *waqallapa* 'all in tears' is based on the verbal root *waqa-* 'cry' modified by the emphatic suffixe *-lla* 'only' and derived to an adverb with the suffixe *-pa*.

4.2. Word segmentation

Quechua has three types of suffixes: nominal suffixes, verbal suffixes, and enclitics. The most productive roots are the verbal roots, which can be enriched with numerous suffixes: directional, associated movement, transitions, and TAM markers. Nominal suffixes are mainly case markers and flexional suffixes. The enclitics are the topic and focus/evidentials markers and the discursive markers; they can be placed on any root. The UD guidelines recommend performing segmentation on syntactic words, with morphological features being encoded as properties of words; however, this strategy is not the most appropriate for agglutinative languages (Han et al., 2020). In Quechua, unlike what has been done for other agglutinative languages such as Turkish, we systematically segment case markers as well as independent suffixes. These two types of morphemes will receive the morphosyntactic label ADP and PART respectively. Only verbal and derivational morphemes will be treated as word features. To keep track of the sometimes complex morphological decomposition of verbal inflected forms, we introduced a new notation in the MISC column which specifies each suffix used in an inflected form together with its morphological gloss in Leipzig format (Comrie et al., 2008). We hope that this strategy will highlight some syntactic phenomena and make it easier to establish linguistic comparisons with languages where case markers are segmented (Japanese, Korean, etc). Let us emphasise that this choice of segmentation is guided more by the objective of developing automatic processing tools from annotated data than by a purely linguistic motivation: we adopted the annotation strategy which allowed us to keep as much morphological information as possible, without straying too far from the basic rules of UD. In addition, we noted that the most recent annotation projects tend to produce richer annotations (in particular whose following the SUD schema (Kim Gerdes et al., 2021))

and to globally segment into morphemes (mSUD schema).

4.3. Morphology

Nominal suffixes Quechua uses suffixes for the following cases among those listed in UD's guidelines: genitive, dative, ablative, locative, terminative, benefactive / purposive, comitative. Nominative is non-marked. The same suffix can indicate both the direct or indirect object (objective case `Obj`). We use additional tags for prolative (`Prl`) and causal (`Csl`) cases.

Quechua also features a set of nominal derivationnal suffixes that derive nouns into nouns or adjectives and whose meaning relates to endowment: ornative *-yuq*, augmentative *-sapa*, privative *-nnaq*, sociative *-ntin*. We didn't segment these morphemes since they are neither case markers nor enclitics.

Verbal suffixes Aktionsart (lexical aspect) can be marked by suffixes, and can co-occur with another aspect. In example 1, at least two aspects are expressed: lexical aspect (durative) and grammatical aspect (progressive). Only grammatical aspect is reported in the morphological features.

(1) kicha-ra-ykaa-naq
open-DUR-PROG-RPST
'was lying open '

Pronominal core arguments Pronominal subject (S) and object (O) in the transitive constructions are coded by verbal suffixes. Their paradigm is more fusional than agglutinative:

- /-q/ encodes S=1 and O=2. Example: *niq* 'I say to you'.
- /-maA/ encodes O=1 when combined with S=2 or S=3. Example: *niman* 'he says to me'. It's combination with the 1.Pl.Incl marker encodes S=3 and O=1.Pl.Incl.
- /-shu/ only combines with /-nki/ and encodes S=3 and O=2. Example: *nishunki* 'he says to you'. *O=3 is unmarked.*

Western Sierra Puebla Nahuatl having similar morphological features, we use the same tags as those used in the corresponding treebank (Pugh et al., 2022) : Person[subj], Person[obj], Number[subj], Number[obj].

4.4. Discourse markers and evidentials

The enclitics can be divided into two groups: discourse markers and evidentials, which are also focus markers (Muysken, 1995). These suffixes are flexible in their distribution and have many semantic nuances: they can be glossed differently according to the discourse context, and it is therefore difficult to link them to predefined morphosyntactic functions. Further studies should be carried out on these particles in order to refine their annotation.

Discourse markers Some have emphatic functions : constrastive *-taq* and limitative *-llaa* (often glossed as 'only') can get the synctactic tag `advmod:emph`. The inceptive *-na* ('already') and continuative *-raq* ('still') can be considered as aspectual perfective and imperfective when linked to a verb or a converb. The discourse marker *-pis* can have many syntactic roles : additivity (`advmod`), coordination (`cc`), topic shift marker, or concessive subordination marker; in the latter cases we generally annotate it with the `discourse:conn` tag, in the same logic as in the Hittite treebank (Andersen and Rozonoyer, 2020).

Evidentiality Quechua has three evidential enclitics, listed in Table 1[1]. (Faller, 2003) showed that they encode evidentiality at an illocutory level. They can only appear once per clause. For *-chi*, we introduced UniMorph's tag `INFER`. For synctactic annotation, UD schema lacks a specific category for evidential particles. In some treebanks, focus-marking copulae have the syntactic relation `AUX` when modifying a verb (for example in Wolof (Dione, 2019)), but Quechua evidentials can occur on any root type and have no morphological features. (Thomas, 2019) annotes the illocutory particles of Mbyá Guarani with `amod` when they occur on noun roots. We choose to use `advmod:emph` tag, used for rhematisers and to modify noun phrases, which best corresponds to the properties of Quechua evidentials. Ancash Quechua has a narrative past marker *-naq*[2] that also denotes reportative evidentiality. Following the Estonian annotation schema, we encode the latter feature as a mood, and use the tag `Mood=Qot`. The distinction with the *hearsay* evidentiality is that they operate on different discursive levels: evidentiality applies beyond discourse, while that of *-naq* remains internal to the proposition in which it appears (Faller, 2003). Reportative *-naq* and *-shI* can be used concurrently on a verbal root.

-mI	assertive	`Fh`
-shI	reportative-quotative	`Hrs`
-chi	conjectural-inferred	`Infer`

Table 1: Ancash Quechua evidentials

[1] /-mI/ and /-shI/ have two allomorphs: *-m* and *-sh* after a short vowel, and *-mi* and *-shi* in other cases.

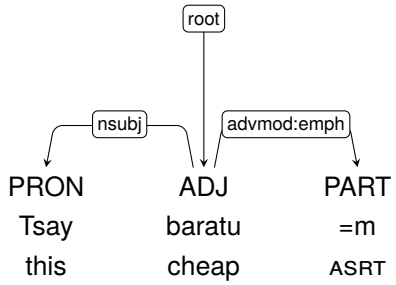
[2] Very similar to Turkish reportative past *-muş*.

5. Some morphosyntactic features

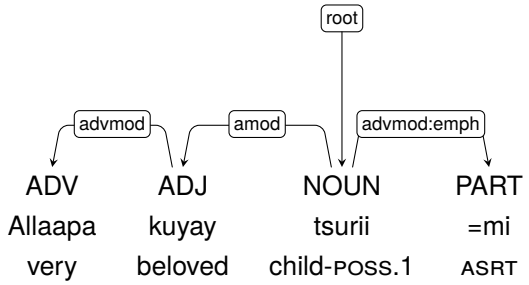
5.1. Non-verbal predication

Attributive predication When the subject is in the third person singular and present tense, no overt linking word is needed for attributive predication: first nominal will be the subject and second nominal the predicate. In independent clauses, focus must be added with one of the evidential, which plays the role of a copula (example 2). When the subject is zero, the evidential is always required and marks the head of the clause (example 3). For other persons and tenses, the copula *ka-* is used with needed inflections.

(2) Tsay baratum. [CRH-03-282]
This is cheap.

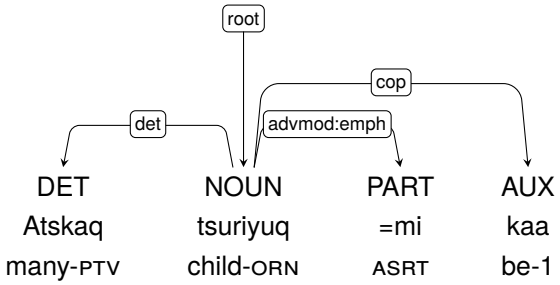


(3) Allaapa kuyay tsuriimi. [CRH-95-104]
She is my dearest daughter.



Ornative predicate The suffix *-yuq* can be used to form noun phrases or adjectival clauses indicating the possession of an object or characteristic expressed by the noun to which it is attached. When the possessor is in 1st or 2nd person, the *ka-* copula with the person mark must be added (example 4).

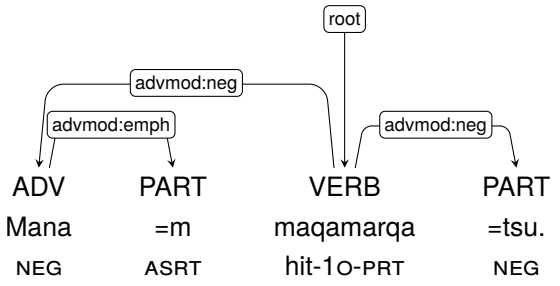
(4) Atskaq tsuriyuqmi kaa. [CRH-06-108]
I have many children.



5.2. Negation

Negation involves two elements: the propositional operator *mana*, and the enclitic *-tsu* that marks the focus of the negation (*-tsu* occupies the same morphological slot as evidentials). Either component of the negation may be omitted in particular contexts. The structure of negation is similar to that of the K'iche language (Yasavul, 2011). We use the same syntactic tag `advmod:neg` that was defined for the treebank of this language (Tyers and Henderson, 2021).

(5) *He didn't hit me*. [CRH-16-077]

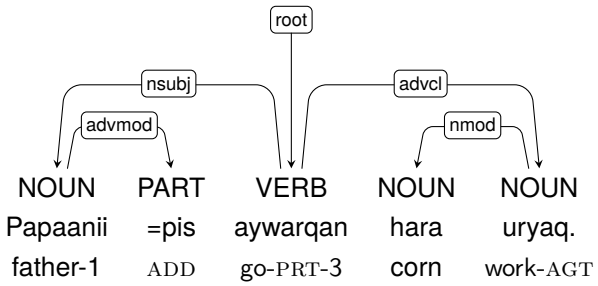


5.3. Complex clauses

Relativisation is performed through agentive nominalisation, marked with the suffixe *-q* (6). The agentive form can also be associated with a finite motion verb to express motion-cum-purpose (7).

(6) Tsaychawmi qatuyaq qichwapita
here-ASRT sell-PL-HAB valley-ABL
shamu**q**kuna.
come-AGT-PL
'Those who came from the valley use to sell there.' [CRH-02-056]

(7) *My father too went to work the corn.*
[CRH-02-018]



In Quechua, subordinators are a set of deverbaliser suffixes that derive non-finite verbs to converbs. A converb with its dependents forms an adverbial clause. This subordinating strategy is very similar to that of Turkish; (Zeyrek and Webber, 2008) describe these suffixes as simplex subordinators. Ancash Quechua has three converbial forms:

- `V` + ***-nqa*** + `Poss` + `Case` (example 8). The person of the possessive suffix corresponds to the person of the subject of the subordinate clause. The subordinator type is given by the case suffix. The suffix *-nqa* generally denotes an imperfective aspect.

- `V` + ***-r***[3] (+ `Poss.3`) (example 9). Without possessive marker, *-r* converbs are similar to English gerunds and are used when the subject of the non-finite verb is the same as that of the main verb. With the possessive marker, it forms a subordinate clause, often with an anaphoric reference to an already introduced subject.

- `V` + ***-pti*** + `Poss` (example 10). The subordinate *-pti* is used when the subject of the main clause differs from that of the subordinate clause.

(8) pahapis quchqu-**nqa**-n-yaq
straw-DIS.CON grind-SUB-POSS.3-TER
'until the straw is ground. ' [CRH-21-024]

(9) alli sapateru ka-**r**-ni-n, atska
good shoemaker be-SUB-EP-POSS.3 lot.of
obrayuq kanaq
work-POSS AUX-RPST
'Being a good shoemaker, he had a lot of work.' [CRH-107-006]

(10) mana qaraynin chura-ya-**pti**-yki-m
NEG his.gift put-PL-SUB-POSS.2-ASRT
Tayta Kolki Hirkaqa qillayninta
Father Silver Mountain-TOP his.metal-OBJ
apakushqa.
taken.back
'Because you did not put offerings to him, Father Kolki Hirka has taken back his metal.' [AKH-361]

UD guidelines recommend annotating converbs with the `VERB` POS-tag. Though Quechua converbial forms also have possessive flexion, which is a noun-type features, this flexion will still be encoded as a `Person` verbal feature.

[3] *-shpa* in some varieties, especially in Huaraz Quechua.

6. Conclusion and future work

We have described a few features of Ancash Quechua for annotating a corpus according to the UD schema. Following these conventions, we have annotated a sample of twenty sentences, freely accessible[4]. This is a work in progress: other morpho-syntactic features are being studied, and we hope to refine the annotation guidelines later. The overall corpus available for annotation has more than 6,000 sentences. Part of this corpus is available in conll format on the Arborator Grew tool (Guibon et al., 2020) for participatory annotation[5]. We are preparing to implement a morphological analyser that will automatically segment the tokens and preannotate the part-of-speech categories.

7. Bibliographical References

Erik Andersen and Benjamin Rozonoyer. 2020. A small universal dependencies treebank for hittite. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 1–7.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. *Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig. Retrieved January*, 28:2010.

Cheikh M Bamba Dione. 2019. Developing universal dependencies for wolof. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 12–23.

Nicholas Q Emlen and Johannes Dellert. 2020. On the polymorphemic genesis of some proto-quechuan roots: Establishing and interpreting non-random form/meaning correspondences on the basis of a cross-linguistic polysemy network. *Diachronica*, 37(3):318–367.

Augusto Escribens and Paul Proulx. 1970. *Gramática del quechua de Huaylas*. Universidad Nacional Mayor de San Marcos.

Martina Faller. 2003. Propositional-and illocutionary-level evidentiality in cuzco quechua.

[4] https://github.com/rumiwarmi/UD_Quechua_qwh/tree/main

[5] https://arboratorgrew.elizia.net/?#/projects/UD_Ancash_Quechua

Semantics of Under-Represented Languages in the Americas, 2(1):3.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Ji Yoon Han, Tae Hwan Oh, Lee Jin, and Hansaem Kim. 2020. Annotation issues in universal dependencies for korean and japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 99–108.

Paul Heggarty. 2005. Enigmas en el origen de las lenguas andinas: aplicando nuevas técnicas a las incógnitas por resolver. *Revista Andina*, 40:9–57.

Bruno Guillaume Kim Gerdes, Sylvain Kahane, and Guy Perrier. 2021. Starting a new treebank? go sud. In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, pages 35–46.

Leonel Alexander Menacho López. 2005. *Yachakuqkunapa Shimi Qullqa, Anqash Qichwa Shimichaw*. Ministerio de Educación, Lima, Perú.

MINEDU. 2020. *Chawpi qichwata alli qillqanapaq maytu 2*.

Pieter Muysken. 1995. Focus in quechua. In Katalin É Kiss, editor, *Discourse configurational languages*, pages 375–393.

Elizabeth Pankratz. 2021. qxoref 1.0: A coreference corpus and mention-pair baseline for coreference resolution in conchucos quechua. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 1–9.

Gary J. Parker. 1976. *Gramática Quechua Ancash-Huailas*. Ministerio de Educación. Instituto de Estudios Peruanos.

Doris L Payne. 1993. Meaning and pragmatics of order in selected south american indian languages. *The role of theory in language description*, pages 281–314.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. Universal dependencies for western sierra puebla nahuatl. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020.

Annette Rios, Anne Göhring, and Martin Volk. 2008. A quechua-spanish parallel treebank.

G. Swisshelm. 1972. *Un diccionario del quechua de Huaraz: quechua-castellano, castellano-quechua*. Estudios culturales benedictinos.

Guillaume Thomas. 2019. Universal dependencies for mbyá guaraní. In *Proceedings of the third workshop on universal dependencies (udw, syntaxfest 2019)*, pages 70–77.

Francis Tyers and Robert Henderson. 2021. A corpus of k'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20, Online. Association for Computational Linguistics.

Murat Yasavul. 2011. Negation and focus in k'iche'. *Proceedings from the Fourteenth*.

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for turkish: Annotating discourse connectives in the metu corpus. In *Proceedings of the 6th workshop on Asian language resources*.

8. Glossing abbreviations

1 / 2 / 3	1st / 2nd / 3rd person
1O	1st person object
AGT	agentive
ASRT	assertive
DIS.CON	discursive connector
DUR	durative
EP	epenthetic morpheme
HAB	habitual past
LIM	limitative
MAN	manner
ORN	ornative
POSS	possessive
PROG	progressive aspect
PRT	preterite (perf. past)
RPST	reportative past
SUB	subordinator
TER	terminative