Morphological Disambiguation and Text Normalization for Southern Quechua Varieties

Annette Rios
Institute of Computational Linguistics
University of Zurich
arios@ifi.uzh.ch

Richard Castro Mamani
Computer Science Department
Universidad Nacional de San Antonio Abad del Cuzco
rcastro@hinantin.com

Abstract

We built a pipeline to normalize Quechua texts through morphological analysis and disambiguation. Word forms are analyzed by a set of cascaded finite state transducers which split the words and rewrite the morphemes to a normalized form. However, some of these morphemes, or rather morpheme combinations, are ambiguous, which may affect the normalization. For this reason, we disambiguate the morpheme sequences with conditional random fields. Once we know the individual morphemes of a word, we can generate the normalized word form from the disambiguated morphemes.[1]

1 Introduction

As part of our research project we have developed several tools and resources for Cuzco Quechua. This includes a hybrid machine translation system Spanish-Quechua. The core system is a classical rule-based transfer engine, that we aim to improve with the addition of statistical modules.

An issue that is generally difficult to deal with in a rule-based approach is the lexical choice of translation options: writing context rules for every possible translation of a given input word is not feasible. Another solution is to include a language model, trained on Quechua texts, that can handle the lexical disambiguation. The total number of available Quechua texts is relatively small, and to complicate matters even further, these texts are written in a wide range of different orthographies. Therefore, the first step in order to obtain a language model is the normalization of the different spellings into a standardized orthography. Not every morphological ambiguity needs to be disambiguated for the normalization alone, but we need fully disambiguated texts for other applications (e.g. parsing). Therefore, we chose to disambiguate not only the cases that are relevant for the normalization, but all types of morphological ambiguities.

2 Related Work

In general, almost every automatic processing of agglutinative languages relies on a correct morphological analysis. Extensive research on morphological disambiguation has been done on Turkish: Görgün et al. (2011) used the WEKA toolkit to train and test several classifiers. With over 50,000 disambiguated sentences for training, they achieved 95.6% accuracy with the J48 Tree algorithm.

Hakkani-Tür et al. (2002) trained an N-gram language model on Turkish roots and another model on so called inflectional groups (groups of morphemes), and used a combination of these two models to disambiguate the output of their finite state analyzer. With a training set of almost 700,000 tokens, they achieved 93.95% accuracy.

Sak et al. (2007) use the combined language models from Hakkani-Tür et al. (2002) to produce an n-best list of morphological parses for a given Turkish sentence. In a second step, they rank the candidates with the voted Perceptron algorithm, trained on 42,000 disambiguated tokens. With this additional step, they achieved an accuracy of 96.8%.

[1] The tool can be tested online at http://kitt.ifi.uzh.ch/kitt/quechua/normalizer.html.

Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pages 39–47,
Dublin, Ireland, August 23 2014.

While the morphological situation with Quechua is comparable to Turkish, the size of the available training data is not: we have less than 3000 manually disambiguated sentences (∼38,000 tokens) that we can use for training. An approach such as the one described by Görgün et al. (2011), where the classifier learns to assign a class for each possible combination of morphemes (without the root), is therefore not feasible: the number of classes that can be learned from such a small training set will not suffice to classify unseen data. Similarly, a language model, even if trained on units smaller than words, as done by Hakkani-Tür et al. (2002), will not overcome the data sparseness in the training set.

For this reason, the approach presented in this paper attempts to break down the disambiguation process into several smaller steps: we move from the root to the last suffix, disambiguating only one morpheme class at a time. With this approach, we achieve an accuracy that is comparable to the results for Turkish.

3 Quechua

Quechua is a language family spoken in the Andes by 8-10 million people in Peru, Bolivia, Ecuador, Southern Colombia and the North-West of Argentina. Although Quechua is often referred to as a language and its local varieties as dialects, Quechua is a language family, comparable in depth to the Romance or Slavic languages (Adelaar and Muysken, 2004, 168). Mutual intelligibility, especially between speakers of distant 'dialects', is not always given.

In this project, we work with Cuzco Quechua (Southern Quechua), and in the following sections, the name Quechua is meant to refer explicitely to this variety. The number of available texts in this particular dialect is limited, therefore we have to include texts from other (similar) varieties of the Southern Quechua dialect group, such as Ayacucho and Bolivian Quechua.

3.1 Dialectal and Orthographic Variation within the Southern Quechua dialect group (QIIC)

Apart from lexical differences, there is one major dialectal divergence between the Cuzco/Bolivian dialects on one side, and the Ayacucho/Argentina varieties on the other side: Cuzco/Bolivian Quechua has, like Aymara, a three way distinction of stops (plain, glottalized and aspirated), whereas Ayacucho and Argentina Quechua have only plain stops. Furthermore, some suffixes appear in different forms, e.g. the progressive in Ayacucho is marked by *-chka*, in Cuzco by *-sha*, and in Bolivia by *-sa* or *-sya*. Other suffixes are restricted to a particular variety: some dialects that are in close contact with Aymara, such as the Quechua spoken in Puno, have borrowed a number of Aymara suffixes, e.g. *-thapi*, *-t'a*, *-naqa*, that are unknown in other dialects (Adelaar, 1987).

Additionally, there are some morphotactic differences concerning the combination of suffixes: for instance, a number of Quechua suffixes change their vowel in combination with certain suffixes, but the exact contexts that induce this vowel change differ to some extent across dialects. Furthermore, the order of suffixes in combinations can vary.

Apart from the dialectal differences, there is also a wide range of orthographic variation within the Southern Quechua dialect group. Several standards have been proposed, most notably the standardized orthography as defined by Cerrón-Palomino (1994). This standard has been adopted by the Bolivian government (Villarroel, 2000), with one small adaption: in Bolivia, the glottal fricative [h] is written as /j/ instead of /h/. In Peru, the situation is slightly more complicated: Although the Ministry of Education has defined an official standard orthography[2],there is still some disagreement regarding the correct spelling of Quechua words. Also, many Quechua texts are written in a more or less Spanish orthography, where for instance /wa/ is written as /hua/, and /ki/ is written as /qui/. Table 1 illustrates the orthography of the Academia Mayor de la Lengua Quechua in Cuzco (first row), a typical 'Spanish' spelling (second row) and an old, non-standardized Bolivian spelling (last row), as opposed to the unified standard orthography as defined by Cerrón-Palomino (1994). This is the orthography that we use for normalization.

[2] As declared in the *Resolución Ministerial N° 1218-85-ED de 1985*

AMLQ	*mana qelqaq yachaq ñausa qelqa runasimipi kasqanku rayku...*
norm.	mana qillqaq yachaq ñawsa qillqa runasimipi kasqankurayku...
span.	*Cay teccsimuyuta, hanacc-pachatapas, Ccanmi tacyachinqui, Ccanmi ticrachinqui..*
norm.	Kay tiqsimuyuta, hanaq pachatapas, Qammi takyachinki, Qammi t'ikrachinki..
boliv.	*Chaywampis paykuna onqosqa kashajtinku, noqaqa llakiy qhashqa p'achasta churakorqani.*
norm.	Chaywanpas paykuna unqusqa kachkaptinku, ñuqaqa llakiy qhachqa p'achakunata churakurqani.

Abbreviations: AMLQ = Academia Mayor de la Lengua Quechua en Cusco, norm = normalized, span = Spanish orthography, boliv = (old) Bolivian orthography

Table 1: Different Orthographies with Corresponding Standardized Version

	variations	standard
progressive	*-chka, -sha, -sa, -sya*	*-chka*
genitive (after vowel)	*-p/-q/-h/-j*	*-p*
evidential (after vowel)	*-m/-n*	*-m*
additive	*-pis/-pas*	*-pas*
euphonic	*-ni/ñi*	*-ni*
1.&2. plural forms	*-chis/-chik/-chiq*	*-chik*
assistive	*-ysi/-schi/-scha*	*-ysi*
potential forms	*-swan/-chwan*	*-chwan*

Table 2: Suffix Variation and Normalization

4 Morphological Analysis

Quechua is an agglutinative, suffixing language. There are over 130 Quechua suffixes, the exact number, as well as the form of the suffixes exhibit substantial variation across dialects. There are five functional classes of Quechua suffixes: nominalizing (noun→verb) and verbalizing (verb→noun), nominal (noun→noun) and verbal (verb→verb) suffixes and so-called independent or ambiguous suffixes, that can be attached to both verbal or nominal forms, without altering the part of speech. The position of these suffixes is at the end of the suffix sequence, their relative order is more or less fixed, though dialects show minor variations. The functions of the independent suffixes include data source, polar question marking and topic or contrast, amongst others (Adelaar and Muysken, 2004, 208).

Quechua roots are, apart from a small number of particles, either verbal or nominal. Adjectives do not constitute a word class on their own on a morphological level, as they behave exactly the same as nominal roots. There may be some syntactic restrictions on true adjectives (Adelaar and Muysken, 2004, 208), but these can be ignored for a morphological analysis. Many roots are indeed ambiguous and can be used either as noun or verb without any derivational suffixes:[3]

(1) *taki* *-y*
song/sing -1S.ps
'My song'

(2) *taki* *-ni.*
song/sing -1S
'I sing'

Furthermore, nominalizing (NS) and verbalizing (VS) suffixes are very productive and can occur more than once in a word.

We obtain the morphological analysis from a finite state analyzer that splits the word forms into morphemes, and also normalizes the surface form of the morphemes. Roots are mapped to their standardized form according to Cerrón-Palomino (1994), e.g. the word for brain, *ñutq'u* in the standard, may appear as *nushqun, ñusqhun, ñusq'un, ñusqun or ñutqun*, depending on the dialect. The normalizer rewrites all these variants to *ñutq'u*. The normalizer also rewrites the form of certain suffixes, see Table 2.

Some of these suffixes are ambiguous in their non-standardized forms, e.g. the direct evidential suffix, written as *-n*, could also be a third person singular marker (verbal or nominal). In order to generate the

[3] Abbreviations used in glosses: Acc: accusative, Add: additive, Dim: diminutive, DirE: direct evidential, Fact: factitive, Fut: future tense, IndE: indirect evidential, Inf: infinitive, Imp: imperative, Loc: locative, NS: nominalizing, P: plural, Perf: perfect, ps: possessive, Rflx: reflexive, S: singular, Top: topic, VS: verbalizing

	Joven	Gregorio	Cancionero
normalizer	97.86%	73.00%	42.56%
Spanish strict	0.64%	21.87%	15.86%
normalizer relax	-	-	34.88%
Spanish relax	-	0.30%	1.48%
guesser	1.02%	2.36%	3.65%
total coverage	99.52%	97.64%	98.43%
unknown words	0.48%	2.46%	1.58%

Table 3: Morphological Analysis Coverage

normalized form of a word with a suffix *-n*, we need to know whether this particular *-n* is a person marker or an evidential suffix. Only in the latter case, *-n* needs to be rewritten as *-m* during normalization.

We have two normalizers in our pipeline: the first one handles text in 'regular' orthographies that show some minor dialectal variations. The second normalizer allows for more 'extreme' orthographies: For instance, both [k] and [q] (velar and postvelar stops) are pronounced as fricatives in certain positions ([x] and [χ]). In many texts both are written as /j/ (or sometimes /h/) if pronounced as fricatives. This introduces new ambiguities, for instance, a root written as *sajsa* could be *saqsa* - 'certain variety of corn' or *saksa* - 'satisfied,full'. In order to avoid additional ambiguities resulting from an analysis with relaxed orthographic rules, the transducer with the additional orthographic rules handles only word forms that were not recognized by the standard normalizer.

As most Quechua texts contain Spanish words, we included two additional finite state transducers that recognize Quechua words with Spanish roots.[4]. The first one recognizes only word forms with correctly written Spanish roots, whereas the second transducer includes several rules that allow for an alternative spelling of the Spanish words (e.g. /c/ might be written as /k/ in a Quechua text). Furthermore, we implemented a guesser that attempts to split word forms into morphemes if the root is unknown. In order to prevent highly unlikely analyses, we restrict the guessing to roots of at least two syllables and with at least one Quechua suffix attached.

The five transducers are joined in a cascade: If the normalizer fails to analyze a word, the Spanish transducer is invoked. If this fails as well, the word is passed on to the second normalizer with relaxed orthography. If the word form has still no analysis, the second Spanish transducer with relaxed orthography attempts to find an analysis. Finally, if all transducers failed, the word is handed to the guesser. One of the texts used for evaluation, a story called *El joven que se subió al cielo* (Lira, 1990) contains relatively few words with Spanish roots, but in the other text, the biography of Quechua native speaker Gregorio Condori Mamani, almost every sentence contains at least one word with a Spanish root. In this case, the Spanish transducer makes a considerable difference: coverage increases by ∼22%, see Table 3. Furthermore, we tested the morphological analyzers on a third text, *Cancionero*, with an even more inconsistent spelling of Quechua words. The *Cancionero* contains religious (catholic) songs written in a 'Spanish' orthography, see the 'Spanish' example in Table 1. The restrictive Quechua and Spanish analyzers recognize only half of the word forms in this text, but the transducers with broader orthographic rules ('relax') increase the number of analyzed tokens to 96%, see Table 3.

5 Disambiguation

Given the fact that a Quechua word form can contain more than one morphological ambiguity, the disambiguation has to be done in several steps. The simplest approach is to disambiguate each word form from 'left to right':

- disambiguate the root (nominal vs. verbal)
- disambiguate nominalizing and verbalizing suffixes
- disambiguate verbal suffixes[5]

[4]The lexicon contains all the Spanish lemmas, except function words, from FreeLing (Padró and Stanilovsky, 2012)

[5]There are no ambiguous sequences within the nominal suffixes, therefore the third step involves only verbal suffixes.

```
suwa  suwa[NRoot][=ladrón]

papanchikta  papa[NRoot][=patata][--]nchik[NPers][+1.Pl.Incl.Poss][--]ta[Cas][+Acc]

tukunqa  tuku[NRoot][=lechuza][--]n[NPers][+3.Sg.Poss][--]qa[Amb][+Top]
tukunqa  tuku[VRoot][=acabar]][--]n[VPers][+3.Sg.Subj][--]qa[Amb][+Top]
tukunqa  tuku[VRoot][=acabar][--]nqa[VPers][+3.Sg.Subj.Fut]
```

Figure 1: Ambiguous Morphological Analysis for Example 3

possible lemmas	case	possible root tags	possible morph tags
suwa	lc	NRoot	-
papa	lc	NRoot	+1.Pl.Incl.Poss, +Acc
tuku	lc	NRoot, VRoot	+3.Sg.Poss, +Top, +3.Sg.Subj, +3.Sg.Subj.Fut

Table 4: Features for Disambiguation with Wapiti, Example 3

- disambiguate independent suffixes

We use Wapiti (Lavergne et al., 2010), a toolkit for sequence labelling that includes an implementation of conditional random fields, in order to train 4 crf models (one model for each step). We decided to use conditional random fields, as the task of morphology disambiguation is in many ways similar to PoS tagging. There is an inter-dependency between the labels: The decision which label a given instance should receive depends to certain extent on the labels of the previous *n* instances.

The training material consists of two Quechua texts that were analyzed with the xfst tools (see section 4) and then manually disambiguated: the biography of Quechua native speaker Gregorio Condori Mamani (Valderrama Fernandez and Escalante Gutierrez, 1977), that contains about 2500 sentences, and some stories from a collection (Lira, 1990), that amount to about 300 sentences.

5.1 Model 1: Disambiguation of Ambiguous Roots

Some Quechua roots can be used nominally or verbally without derivation, see Example 1 and 2. The disambiguation of roots can be regarded as PoS tagging with a very small tagset. Consider the following example (taken from a story in (Lira, 1990)):

(3) *..suwa papa -nchik -ta tuku -nqa..*
thief potato -1P.ps -Acc end -3S.Fut
'[..] the thief will take all our potatoes [..] (lit. 'the thief will end our potatoes')

The root *tuku-* 'to end' is ambiguous: *tuku-* can also be a nominal root with the meaning 'owl'. Furthermore, the sequence *-nqa* is ambiguous, apart from the 3rd singular future form, it could be a combination of *-n*, '3rd singular subject' or '3rd singular possessive', and *-qa*, 'topic', see Fig. 1 with the output of the xfst analyzer for this example. In a first step, the type of the root has to be determined, the ambiguity of *-nqa* is only relevant if the root is verbal and will be postponed for later. In order to disambiguate the root with Wapiti, every token needs to be converted into a set of features (an instance) extracted from the xfst output, see Table 4. The words *suwa* and *papanchikta* are not ambiguous and therefore have only one possible root tag, whereas *tukunqa* has two possible root tags: VRoot and NRoot. Model 1 will assign one of them as class label, considering the features and the context of the given token. Wapiti allows pre-labeled input data, therefore, we can already set the label of the unambiguous words *suwa* and *papanchikta*. Note that the instances do not contain the full word form; due to the small size of our training corpus, using full word forms leads to increased data sparseness and impairs the results.

5.2 Model 2: Disambiguation of Nominalizing and Verbalizing Suffixes

Even after the disambiguation of the root type, the final word form can still be either nominal or verbal, as certain nominalizing and verbalizing suffixes are homophonous with verbal or nominal morphemes.

Consider the following examples:

(4) *wasi -cha -y*
house -Fact(VS) -Inf(NS)/2.Imp
'to build a house' or 'build a house!'

wasi -cha -y
house -Dim -1S.ps
'my small house, cottage'

(5) *rikhu -sqa -yki*
see -Perf(NS) -2S.ps
'the one you saw, your seeing'

rikhu -sqayki
see -1S>2S.Fut
'I will see you'

The suffix *-cha* attached to a nominal root can be either a diminutive or a factitive suffix ('make'): With the diminutive, the resulting word form is still a noun, whereas the factitive suffix produces a verb. In total, model 2 handles eight different cases of ambiguous verbalizing/nominalizing vs. verbal/nominal suffixes. The features in models 2-4 are essentially the same as those in model 1 (see Table 4), but of course the root type is no longer ambiguous, consequently there is only one root tag. With models 2-4, we classify only words that exhibit a verbalizing/nominalizing vs. nominal/verbal ambiguity, whereas words that are unambiguous for the particular model receive a dummy label ('none').

5.3 Model 3: Disambiguation of Verbal Morphology

In the next step, we disambiguate six possible ambiguities in verb forms. One of the ambiguities in question is the sequence *-nqa* from example 3: After applying model 1, we know that the root *tuku* in *tukunqa* is verbal, but *-nqa* can still be either the 3rd singular future form or a combination of 3rd singular present and topic marker, see example 6. Other ambiguities of this type involve *-sun*, which can be either the imperative or future form of the first plural inclusive, as well as the sequence *-sqaykiku*, which can be either the indirect past or future form of the first plural exclusive acting on a 2nd singular person.

(6) *tuku -nqa*
end -3S.Fut
'he will end'

tuku -n -qa
end -3S -Top
'he ends'

(7) *llamk'a -sun*
work -1Pl.incl.Fut
'we will work'

llamk'a -sun
work -1Pl.incl.Imp
'let's work'

(8) *qhawa -sqaykiku*
look -1Pl.excl.>2S
'we (excl.) watch you'

qhawa -sqa -ykiku
look -IPst -1Pl.excl
'we (excl.) watched [they say]'

5.4 Model 4: Disambiguation of Independent Suffixes

Model 4 disambiguates ambiguities that concern independent suffixes. None of these potential ambiguities occur in all dialects and orthographies, but all of them concern the normalization and are therefore important. There are 3 types of ambiguities that relate to independent suffixes:

The most common case involves the suffix *-n*, when the word form is nominal and *-n* follows a vowel: in this case, *-n* can be the 3rd singular possessive, or it can be the allomorph of the evidential suffix *-mi*. The latter is written as *-m* in the standard orthography, as well as in texts written in Ayacucho Quechua, but occurs as *-n* in many texts written in Cuzco and Bolivian Quechua, see Example 9. A further ambiguity that occurs only in Cuzco and Bolivian Quechua concerns the sequence *-pis*: *-pis* can be the additive suffix (in Ayacucho Quechua always *-pas*) or a combination of the locative suffix *-pi* and the evidential suffix *-s*, see Example 10. The third ambiguity of this type concerns Spanish words that end in *-s*: In this case, *-s* can be an evidential suffix, but it can also be the Spanish plural[6], see Example 11.

[6]In certain Bolivian dialects *-s* is also used on native roots as plural suffix, see the Bolivian word *p'achasta* (normalized *p'achakunata*) in Table 1.

			gregorio	joven
model 1	root tag	Wapiti	**95.35**	**85.71**
		baseline	65.12	72.62
model 2	NS/VS	Wapiti	**97.44**	**87.88**
		baseline	80.49	17.47
model 3	verbal s.	Wapiti	85.71	66.67
		baseline	**88.89**	**75.00**
model 4	independent s.	Wapiti	**85.37**	**86.11**
		baseline	64.10	50.00

Table 5: Evaluation: Precision of the Morphological Disambiguation Steps

(9) *wasi -n*
house -DirE
'house'

wasi -n
house -3S.ps
'his house'

(10) *chay -pis*
this -Add
'also this'

chay -pi -s
this -Loc -IndE
'there [they say]'

(11) *derechu -s*
right -IndE
'right [they say]'

derechus
rights
'rights'

6 Evaluation

We used the same test sets as for the evaluation of the morphological analysis in section 4: The last 72 sentences from the autobiography of Gregorio Condori Mamani (Valderrama Fernandez and Escalante Gutierrez, 1977), and the Andean story *El joven que se subió al cielo* from (Lira, 1990) with about 250 sentences. Both test texts were excluded from the training set.

Table 5 illustrates the percentage of correctly disambiguated words with the particular ambiguity for each step. Note that there were only a handful test cases for model 3 (verbal suffixes) in both texts, therefore, the results for this step might not be accurate. Furthermore, the number of instances extracted from the training material for model 3 is smaller than for the other models, as these types of ambiguities are relatively rare. For the normalization, errors in model 3 do not affect the outcome, as these ambiguities have no effect on the surface forms in the standard orthography. Considering for instance example 6, *-nqa* will be *-nqa* in the standard, irrespective of whether the analysis is *-n -qa* or *-nqa*.

Table 6 contains the evaluation of the whole texts. Although the percentage of tokens with a wrong morphological analysis is almost the same in both texts, the total number of correctly analyzed words is lower in the biography. This is due to the fact that this text contains many words with Spanish roots, sometimes with 'quechuized' spelling. Many of these words were not recognized by the xfst analyzer and were therefore not normalized.

The baseline for both Table 5 and 6 was calculated based on the frequencies of the forms in the training material: The baseline shows the results that we obtain if we disambiguate the test texts choosing always the most frequent class in every decision. The biggest difference as opposed to the Wapiti models is that with this approach, we do not consider any context information. As you can see in Table 5, Wapiti outperforms the baseline in every step except for model 3, where the training instances are too sparse. There is a considerable difference in the baseline for the two test texts (see Table 6): on the biography, the baseline is much higher. This is due to the fact that the largest part of the training material is part of the same book, therefore the probability distribution of the individual classes in this test text correlates better with the frequencies calculated from the training material. While the conditional random fields improve the disambiguation on the test set similar to training material only slightly compared to the baseline (+2%), the effect they have on the results for a test set from a different text is considerable: $>$10%. Table 6 also contains the results obtained with the RFTagger (Schmid and Laws, 2008) and Morfette (Grzegorz et al., 2008) for comparison. The main difference between our approach and the morphological taggers is that the latter analyze and label the complete word form at once, whereas with our approach, we disambiguate and normalize each word in several steps, proceeding from left to right. The tagset used by the morphological taggers is thus much more fine-grained, as each tag contains the

	El joven que subió al cielo		*Gregorio Condori Mamani*	
total sentences:	258		72	
total token	1865		1015	
punctuation marks:	567		171	
xfst failures:	9	0.48%	25	2.46%
total word forms	1298		844	
correct analysis:	1252	**96.46%**	789	**93.48%**
wrong analysis:	33	2.54%	17	2.01%
guessed, no analysis in gold:	4	0.31%	6	0.71%
ambiguous words:	282	21.73%	127	15.05%
still ambiguous:	0		7	5.51%
correct of ambig.:	249	88.30%	103	81.10%
wrong of ambig.:	33	11.70%	17	13.39%
morphological tagging (tag whole word form):				
RFTagger (bigrams):		65.49%		72.21%
Morfette:		65.1%		78.32%
baseline (most frequent morphemes):		85.98%		91%

Table 6: Evaluation: Disambiguated Texts

morphology of the whole word. The results show clearly that our training corpus is too small to achieve satisfactory results with morphological tagging. As mentioned before, not all ambiguities are relevant for the normalization. In fact, many morphological ambiguities are not relevant for the conversion to the standard orthography, therefore, the number of correctly normalized forms is higher than the proportion of correctly disambiguated words from Table 6. In the text *El joven que subió al cielo*, the percentage of correctly normalized words amounts to 99.61%, whereas for the biography of Gregorio Condori Mamani, we achieve only 98.93%.

7 Conclusions

As standardized spelling is an indispensable prerequisite for any statistical processing, we built a pipeline to normalize Quechua texts through morphological analysis and disambiguation. The morphological analysis includes 5 cascaded transducers, two with Quechua root lexica and two with Spanish root lexica, as Spanish loan words occur very frequently in Quechua texts. In every pair of transducers, the first one follows a relatively strict orthography, whereas the second one has a set of phonological rules that allow for more variation in the spelling of word forms. Furthermore, the cascade includes a guesser that attempts to split word forms into morphemes if all the other transducers failed to do so. The transducers rewrite the individual morphemes according to the Unified Southern Quechua orthography (Cerrón-Palomino, 1994), but many words involve morphological ambiguities that might affect the normalized form. In order to choose the correct analysis, we conduct a morphological disambiguation with conditional random fields. We disambiguate the Quechua words in 4 steps, with four models trained to classify the different types of ambiguities. Finally, we generate the normalized word forms from the now disambiguated sequence of morphemes. Our initial results are comparable to morphological disambiguation on Turkish texts, despite the fact that we have a much smaller training corpus ($\sim$ 2800 sentences, compared to over 50,000 (Görgün and Yildiz, 2011) and 45,000 sentences (Sak et al., 2007)). A possible explanation is that Turkish morphology is more complex: Turkish has more productive suffixes than Quechua, and there are relatively complex morpho-phonological rules that determine word formation, such as two dimensional vowel harmony and context-sensitive realizations of consonants (Oflazer, 1994). Quechua on the other hand, is a very regular agglutinative language.

Certain parts of the disambiguation pipeline suffer from data sparseness, in fact, at least one possible ambiguous sequence never occurred in our training corpus and can therefore not be disambiguated, see section 5.4. As the annotation of our treebanks proceeds, we will have more manually disambiguated text, since the syntax trees are built on morphemes, not on whole words. With more training material, the accuracy of the disambiguation and normalization process should increase.

References

Willem F. H. Adelaar and Pieter Muysken. 2004. *The Languages of the Andes*. Cambridge Language Surveys. Cambridge University Press, Cambridge, UK.

Willem F. H. Adelaar. 1987. Aymarismos en el quechua de Puno. *Indiana*, 11:223–231.

Rodolfo Cerrón-Palomino. 1994. *Quechua sureño, diccionario unificado quechua-castellano, castellano-quechua*. Biblioteca Nacional del Perú, Lima.

Onur Görgün and Olcay Taner Yildiz. 2011. A Novel Approach to Morphological Disambiguation for Turkish. In Erol Gelenbe, Ricardo Lent, and Georgia Sakellari, editors, *Computer and Information Sciences II - 26th International Symposium on Computer and Information Sciences*, pages 77–83, London, UK. Springer.

Chrupala Grzegorz, Georgiana Dinu, and Josef van Genabith. 2008. Learning Morphology with Morfette. In Khalid Choukri Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2002. Statistical Morphological Disambiguation for Agglutinative Languages. *Computer and the Humanities*, 36:381–410.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical Very Large Scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513, Uppsala, Sweden. Association for Computational Linguistics.

Jorge Lira. 1990. *Cuentos del Alto Urubamba*. Centro de Estudios Regionales Andinos "Bartolomé de las Casas", Cuzco, Peru.

Kemal Oflazer. 1994. Two-level Description of Turkish Morphology. *Literary and Linguistic Computing*, 9(2).

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2007. Morphological disambiguation of Turkish text with perceptron algorithm. In A. Gelbukh, editor, *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 107–118, Mexico City, Mexico. Springer.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1 of *COLING '08*, pages 777–784. Association for Computational Linguistics.

Ricardo Valderrama Fernandez and Carmen Escalante Gutierrez. 1977. *Gregorio Condori Mamani - Autobiografía*. Biblioteca de la Tradición Oral Andina. Centro de Estudios Rurales Andinos 'Bartolomé de las Casas', Cuzco, Peru.

Alfredo Quiroz Villarroel. 2000. *Gramática Quechua*. Ministerio de Educación, Cultura y Deportes, Fondo de las Naciones Unidas para la Infancia (UNICEF), Bolivia.