V-Unit Summer 2005
Ariadna Font Llitjós
aria@cs.cmu.edu
October 10, 2005

Developing a Quechua-Spanish Machine Translation system

Peru (June-August 2005)

Introduction and Motivation

Information and communication technology (ICT) is playing an increasingly influential role in reshaping trade, growth, employment and production in large parts of the world. ICT presents unprecedented opportunities to combat poverty by increasing income, opening markets and providing a channel through which the voices of the poor can be heard. One of the world's main challenges is to broaden the reach of ICT to those who risk being left behind.

According to the World Bank reports, there is increasing consensus among indigenous leaders that new information and communication technologies open up major opportunities for indigenous development. Pilot schemes are underway in Latin America and the Caribbean region, in which these new technologies are being effectively used to promote indigenous participation in governmental and international donor-sponsored development programs. At the same time, the ICTs can be used to enhance experience-sharing among indigenous leaders, promote online education, and facilitate improved dialogue with international donors and national governments. Nevertheless, there are huge disparities in access to these technologies. The so-called *digital divide* is reflected in lack of access to the technologies and, especially, in lack of training in how to use them.

Institutions like the World Bank have several projects in Latin America to close the digital divide, but there is no project that directly addresses the issue of developing Machine Translation systems for indigenous languages. The development of technologies that will allow indigenous people to communicate with the rest of the world in their own language (either for assimilation or dissemination purposes) do not lack interest or importance. However, there are no commercial systems available for these so-called minority languages. The main reason for this being that developing technology for such languages is non-profitable since such languages typically do not have an economy supporting them, which can afford the time and costs of hiring linguists and computational linguists.

The Quechua language is no exception. There is a lack of technology available in Quechua to bridge the communication gap between Andean indigenous communities and the cities and their official Spanish-speaking Peruvian government.

And what is worse, for a language like Quechua, it seems hardly feasible that there will be any private or governmental efforts to create and develop technology in Quechua for

Quechua speakers any time soon. Yet, there is a real need for bridging the communication gap between monolingual Quechua speakers who need to go to the city often to interact with monolingual Spanish speakers, such as doctors and lawyers.

About Quechua

Quechua or *runasimi*, which means language of the people, is the indigenous language of a large portion of the South American highlands (Andes), and there are about 10 million speakers today. However, we know of no electronic resources in Quechua, let alone any information and communication technologies in Quechua.

The term Quechua covers a variety of distinct languages and dialects. A search of the Ethnologue Data Base (http://www.sil.org/www/ethnologue/ethnologue.html) showed 46 dialect of Quechuan, 32 spoken in Peru. Quechua is also spoken in Bolivia, Ecuador, South of Colombia and North of Argentina. The most important dialect is that spoken in Cuzco, the seat of the former Inca Empire. Quechua spread by means of conquests realized before and during that empire. It displaced several earlier languages, only to find itself increasingly displaced today by Spanish. In spite of this intense competition, Quechua in its various forms remains a vital language in Peru and elsewhere. To see statistics on Cuzco Quechua given by the Ethnologue, see Appendix 1.

Problem

Societal problem

Peruvian Spanish-speakers think of Quechua as an inferior, “Indian” language, with no prestige, importance or interest whatsoever. Many Quechua speakers also think of Quechua as an inferior language; they often do not want their children to even learn it at school. The main problem is that Quechua is still associated with rural contexts and a lack of future, whereas Spanish is associated with having a future and being able to find a job in a big city.

Nowadays, technology is often associated with prestigious, important matters that are worth the time, money and effort to spend developing such technology. Technology is always associated with modernization.

Developing technology in languages like Quechua are not cost effective, yet from an anthropological, linguistic, and purely humanitarian point of view, it is something that will help preserve the language and hopefully increase the awareness of its speakers towards its importance. The ultimate goal being to help Andean people gain back the prestige and political strength of past centuries, and improve their self-esteem as Quechua speakers, so that they are not ashamed of speaking their language nor are they afraid of passing it onto future Andean generations.

Technology approach and motivation for picking it

The AVENUE project at the Language Technologies Institute, in the School of Computer Science at Carnegie Mellon University is devoted to the rapid development of Machine Translation (MT) systems for languages with scarce-resources in a cost-effective manner.

The AVENUE MT approach is to combine different types of MT in one "omnivorous" system that will eat whatever resources are available. If a parallel corpus is available in electronic form, we can use example based machine translation (EBMT) (Brown, 1997; Brown and Frederking, 1995), or Statistical machine translation (SMT). If native speakers are available with training in computational linguistics, a human-engineered set of rules can be developed. Finally, if neither a corpus nor a human computational linguist is available, AVENUE uses a machine learning technique called Seeded Version Space Learning (Probst, 2005) to learn translation rules from data that is elicited from a native speaker.

The last approach assumes the availability of a small number of bilingual speakers of the two languages, but these need not be linguistic experts. The bilingual speakers create a comparatively small parallel corpus of phrases and sentences (on the order of magnitude of a few thousand sentence pairs) and align the words of the two languages using a specially designed elicitation tool (Probst *et al.* 2001). From this data, the learning module of our system automatically infers hierarchical syntactic transfer rules, which encode how constituent structures in the source language (SL) transfer to the target language (TL). The collection of transfer rules, which constitute the translation grammar, is then used in our run-time system to translate previously unseen SL text into the TL text (Probst *et al.* 2003).

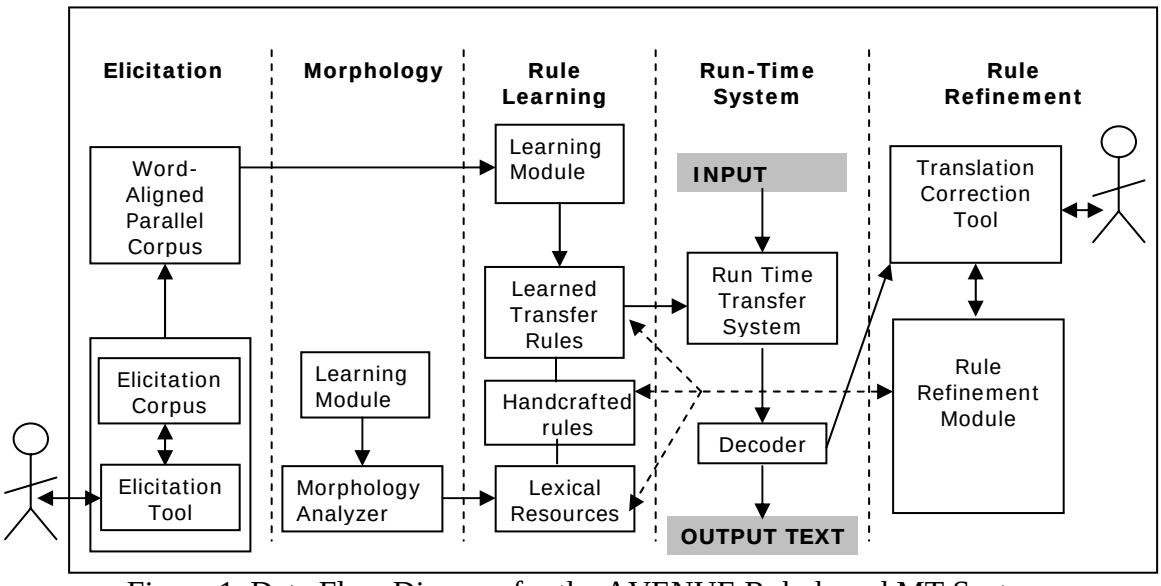

Figure 1. Data Flow Diagram for the AVENUE Rule-based MT System.

The AVENUE project as a whole consists of six main modules, which are used in different combinations for different languages: elicitation of a word aligned parallel corpus (Levin *et al*. in press); automatic learning of translation rules (Probst, 2005) and morphological rules (Monson *et al*. 2004); the run time MT system for application of SL-to-TL transfer rules; the EBMT system (Brown, 1997); a statistical decoder for selecting the most likely translation from the available alternatives; and a module that allows a user to interactively correct translations and that uses their feedback to automatically refine incorrect translation rules and lexical entries (Font Llitjós *et al*. 2005). Figure 1 shows an Overview of the Avenue Architecture.

Accomplishments

1. Quechua to Spanish MT system prototype (25 grammar rules; 683 lexical entries (40 manually and 643 semi-automatically created))
2. Test sets:
    a. from Elicitation Corpora (Functional Corpus: 1743 sentences; Structural Corpus: 198 sentences)
    b. 7 simple dialogs
3. User study for correcting Quechua-Spanish translations using the Translation Correction Tool (part of the Rule Refinement module).

Quechua to Spanish MT system

The first step in order to develop a Quechua to Spanish MT system prototype involved creating initial resources and infrastructure. The following tasks describe the different resources and steps that were taken to develop such a system.

Tasks 1: Obtaining parallel corpora

Elicitation Corpus

The AVENUE team worked together with two bilingual speakers to translate and align the AVENUE Elicitation Corpus. The purpose of the elicitation system is to collect a parallel corpus whose content is controlled in order to ensure that it illustrates the basics of the language being elicited. The elicitation system (Probst et al., 2001; Probst and Levin, 2002) can be used by an informant who is bilingual in the language of elicitation and the language being elicited. The informants are required only to translate Spanish sentences into Quechua and to align Spanish words to Quechua words as well as they can. Because a human linguist may not be available to supervise the elicitation, a user interface is available for presenting sentences to an informant and allowing the informant to translate and align sentences. Some potential pitfalls of automated elicitation are described in Probst and Levin, 2002.

The elicitation corpus follows two organizational principles. The first is compositionality. Small phrases are elicited first, and are then combined into larger phrases. For example,

simple noun phrases are elicited first followed by noun phrases containing possessors, simple sentences, and multi-clausal sentences. Compositionality in the corpus facilitates the learning of compositional transfer rules.

The second organizational principle of the elicitation corpus is creation of minimal pairs of sentences. Minimal pairs of sentences differ in only one feature such as tense, number of the subject, gender of the possessor, etc. A process of feature detection compares the members of the minimal pairs in order to make a first guess at what grammatical features (verb agreement with subjects and objects, number, tense, etc.) are marked in the language being elicited.

The Functional Elicitation corpus has now 1,700 sentences. The current coverage includes basic transitive and intransitive sentences, animate and inanimate subjects and objects, definite and indefinite subjects and objects, present/ongoing and past/completed events, singular, plural, and dual nouns, simple noun phrases with determiners and adjectives, and possessive noun phrases. Following guides for field workers such as Comrie-smith and BouquiauxThomas we expect the elicitation corpus to grow to several thousand sentences.

There has been a recent addition to the Elicitation Corpus, what we call the Structural Elicitation Corpus, a smaller corpus designed to cover the major structures present in the Penn Treebank (Mitchell *et al.*, 1992). Out of 122,176 sentences from the Brown section of the Penn Treebank, 222 different basic structures and substructures were extracted. Namely, 25 AdvPs, 47 AdjPs, 64 NPs, 13 PPs, 23 SBARs, and 50 Ss. For more information about how this corpus was created and what its properties are, see Probst and Lavie (2004). The final Structural Elicitation Corpus which was translated into Quechua had 146 Spanish sentences.

The elicitation corpus can be used as development and test sets to write translation rules and create a lexicon to start developing an MT system prototype.
It can also be used for training automatic acquisition of MT transfer rules and an Example-Based MT system.

A native Quechua speaker, Irene Gómez, and a linguist with good knowledge of Quechua, Marilyn Feke, translated both the Functional Elicitation Corpus and the Structural Elicitation Corpus. In addition to that, a non-native speaker of Quechua (with knowledge of it as a second language), Yenny Ccolque, work with focus groups, mainly from the *Casa del Cargador* in Cusco, in order to translate several of the sentences in the Elicitation Corpora.

Extracting written parallel text: OCR correction

Besides the three versions of the Elicitation Corpora, the Avenue team did not have access to any other Quechua text on electronic format, and so the following three books containing Spanish and Quechua parallel text were scanned:

Cuento Cusqueños
Cuentos de Urubamba
Gregorio Condori Mamani

Because optical character recognition (OCR) for non-major languages is not a solved problem, the scanned Quechua text contained several OCR errors per page, and thus two people with good knowledge of Quechua went over the Quechua text (360 pages total) and correct the OCR errors, given an image of the original text.

A third of the manual correction was done by Salomé Gutierrez (from University of Pittsburgh) and the remaining two thirds were completed by Yenny Ccolque (from Cuzco). Neither of them is a native speaker of Quechua, however, both have good knowledge of Quechua and were given the images of the original Quechua text to compare them with the scanned text.

Task 2: Word Segmentation and Translation

In order to build a translation and morphology lexicon, as many examples as possible of segmented Quechua words translated into Spanish are needed.

For this purpose, all the types of words from the three Quechua books scanned were automatically extracted and ordered by frequency. The total number of types were 31,986 (*Cuento Cusqueños* 9,988; *Cuentos de Urubamba* 12,223; *Gregorio Condori Mamani* 12,979), with less than 10% overlap between books, only 3,002 word types were in more than one book.[1] Since 16,722 word types were only seen once in the books, we decided to segment and translate only the first 10,000 most frequent words in the list, hoping to reduce the amount OCR errors and misspellings.

Additionally, all the different types of words from the Elicitation Corpora translated by Irene Gómez were also extracted (1,666 word types) to make sure our lexicons covered everything in our Elicitation Corpora.

I worked very close with Irene Gómez to segment and translate the word types extracted from the Elicitation Corpora as well as the first 3,000 most frequent word types from the Quechua books. This was done having the list of words in Excel files with the following fields:

1. Word
2. Segmentation
3. Root translation
4. Root POS
5. Word Translation
6. Word POS
7. Translation of the final root if there has been a POS change.

[1] This was done before the OCR correction was completed and thus this list contained OCR errors.

The reason for the last field (Translation of the final root if there has been a POS change) is that if the POS fields for the root and the word differ, the translation of the final root might have changed and thus the translation in the lexical entry actually needs to be different from the translation of the root specified in the 3rd field.

In Quechua, this is important for words such as "machuyani" (I age/get older), where the root "machu" is an adjective meaning "old" and the word is a verb, whose root really means "to get old" ("machuyay")[2], and instead of having a lexical entry like V-machuy-viejo (old), we are interested in having a lexical entry V-machu(ya)y-envejecer (to get old).

During this close cooperation, some issues with the Spanish version of the Elicitation Corpora became apparent, as well as how to improve the Elicitation Tool so that context cannot be so easily bypassed by users.

Task 3: Quechua Morphology Analyzer

While Spanish and English are analytic languages, Quechua is an agglutinative and polysynthetic language, and so many of the concepts that are expressed as independent words in Spanish, are realized as dependent suffixes in Quechua, and in some cases it even incorporates lexical nucleus to verb stems.

For example, many of the concepts that are realized as adverbials in Spanish and English are suffixes that can attach to a verb or a noun. Furthermore, negation, aspect, reflexivity, passive voice and many other grammatical categories are marked in Quechua by means of verbal suffixes and not stand-alone words, as in Spanish and English.

From this, it follows that an MT system should not try to translate Quechua words directly into Spanish words. There is the need to identify each element with meaning in a Quechua sentence, so that the system can then properly translate it into the corresponding Spanish word or phrase. For this, a morphological analyzer is needed.

In order to test the Quechua morphology analyzer, a small Suffix Lexicon and Stem Lexicon were manually created. The Quechua morphology analyzer is now integrated and running with the Transfer Engine. At a later stage, the code of the morphology analyzer was modified to work optimally for Quechua. Initially, however, no morphology analyzer was working for Quechua and thus initial development and test sets were segmented manually, in order to start testing all the other components of the MT prototype.

The next step will be to test the MT system with unsegmented input and the Quechua morphology analyzer. In order to take advantage of the full power of the morphological analyzer, though, the morphology lexicon needs to be augmented to match the translation

[2] -ya- is a verbalizer in Quechua.

lexicons. This can be done writing a script to format the existing lexical entries into the format expected by the morphology analyzer.

Task 4: Stem Lexicon

Form the list of segmented and translated words result of Task 2, two lexicons containing mostly stems from the 100 most frequent words and from the two different types of the Elicitation Corpora (643 lexical entries) were automatically generated and manually corrected.

This was done with several Perl scripts to automatically parse the Excel files containing the segmentation and translation information and to generate flat lexical entries for each different POS and alternative translation specified for each word type.

For example, from the word type “chayqa” and the specifications given for all the other fields as shown below, six different lexical entries were automatically created, one for each POS and each alternative translation (Pron-ese, Pron-esa, Pron-eso, Adj-ese, Adj-esa, Adj-eso):

Word	Segmentation	Root translation	Root POS	Word Translation	Word POS
chayqa	*chay+qa*	*ese* | *esa* | *eso*	*Pron* | *Adj*	*ese* | *es ese*	*Pron* | *Adj*

In some cases, when the word has a different POS, it actually is translated differently in Spanish. For these cases, the native speaker was asked to use || instead of |, and the post-processing scripts were designed to check for the consistency of || in both the translation and the POS fields. When the script encounters ||, it assigns the first translation to the lexical entry with the first POS, and the second translation with the seconds POS of speech, for example.

The scripts allow for fast post-processing of thousands of words, however manual checking is still required to make sure there are no spurious lexical entries.

These are some examples of automatically generated lexical entries:

V |: [ni] -> [decir]
((X1::Y1))

N |: [nina] -> [fuego]
((X1::Y1))

N |: [pacha] -> [tiempo]
((X1::Y1))

N |: [pacha] -> [tierra]
((X1::Y1))

Adj |: [hatun] -> [grande]
((X1::Y1))

Adj |: [hatun] -> [alto]
((X1::Y1))

Adv |: [kunan] -> [ahora]
((X1::Y1))

Adv |: [allin] -> [bien]
((X1::Y1))

```
Pron |: [noqa] -> [yo]
((X1::Y1))

Interj |: [alli] -> ["a pesar"]
((X1::Y1))

Adv |: [ama] -> [no]
((X1::Y1))
```

In addition to the automatically generated lexicons, for the current prototype, a small Stem Lexicon with has 19 stems was also created and tested.

Task 5: Suffix Lexicon

On the other hand, most of the suffix lexical entries were hand-crafted, since they are more complex there are only about 150, as listed in Cusihuaman's grammar (2001).

For the current MT prototype, the Suffix Lexicon has 21 entries. And the necessary features for 36 other suffixes have already been listed and they just need to be converted into the lexicon format.

```
; "dicen que" on the Spanish side
Suff::Suff |: [s] -> [""]
((X1::Y1)
((x0 type) = reportative))

; when following a consonant
Suff::Suff |: [si] -> [""]
((X1::Y1)
((x0 type) = reportative))

Suff::Suff |: [qa] -> [""]
((X1::Y1)
 ((x0 type) = emph))


Suff::Suff |: [chu] -> [""]
((X1::Y1)
((x0 type) = interr))

VSuff::VSuff |: [nki] -> [""]
((X1::Y1)
((x0 person) = 2)
((x0 number) = sg)
((x0 mood) = ind)
((x0 tense) = pres)
((x0 inflected) = +))

NSuff::NSuff |: [kuna] -> [""]
((X1::Y1)
((x0 number) = pl))

NSuff::Prep |: [manta] -> [de]
((X1::Y1)
((x0 form) = manta))
```

In the future, we plan to keep expanding the suffix lexicon to include all 150 suffixes listed in the Cusihuaman grammar.

Task 5: Development and Test Sets

The sentences from both types of Elicitation Corpora were extracted and were used as development and test sets to further expand the coverage of the MT system.

In order to actually run this sets through the MT system, the translation lexicon needs to contain all the words that appear in them, namely all the stems and all the suffixes.

Since vocabulary from the EC has already been translated and segmented during Task 2, Lexical entries for all of the stems that appear in the development and test sets were automatically created (described in Task 4). However, many of the suffixes that appear in these sets are still not in the Suffix Lexicon, especially all verb conjugations, and thus the system currently only produces good translations for the first 50 sentences of the Development set.

Task 6: Grammar Rules

In the AVENUE system, translation rules have 6 components: a) the type information, which in most cases corresponds to a syntactic constituent type; b) part-of speech/constituent sequence for both the source language (SL), in this case Quechua, and the target language (TL), in this case Spanish; c) alignments between the SL constituents and the TL constituents; d) x-side constraints, which provide information about features and their values in the SL sentence; e) y-side constraints, which provide information about features and their values in the TL sentence, and f) xy-constraints, which provide information about which feature values transfer from the source into the target language.

The NP rule below illustrates an example of a Quechua to Spanish translation rule noun-phrases containing a noun and an adjective in inserting a determiner in Spanish.

```
        ; Example: hatun wasi → la casa grande
a)      {NP,4}
b)      NP::NP : [Adj NBar] → [Det NBar Adj]
        ; x0  y0     x1    x2          y1    y2     y3
c)      ( (X1::Y3)  (X2::Y2)
d)      ((x0 number) = (x2 number))
d)      ((x0 person) = 3)

f)      ((y2 number) = (x2 number))

e)      ((y0 gender) = (y2 gender))
e)      ((y1 gender) = (y2 gender)) ; det-n agreement in Spanish
e)      ((y1 number) = (y2 number)) ; det-n agreement in Spanish

e)      ((y3 gender) = (y2 gender)) ; n-adj agreement in Spanish
e)      ((y3 number) = (y2 number)) ); n-adj agreement in Spanish
```

This translation rule swaps the original Quechua word order, from adjective-noun to noun-adjective, inserts a determiner on the Spanish side, enforces their agreement in Spanish and ensures that the noun is the head of the NP, passing its features up to the mother node ((x0 number) = (x2 number)).

The feature unification equations used in the rules follow a typical unification grammar formalism. For more details about the rule formalism, see Probst *et al*. 2003.

The current translation grammar for Quechua-Spanish contains 25 rules and it covers subject-verb agreement, agreement within the NP (Det-N and N-Adj), intransitive VPs, copula verbs, verbal suffixes, nominal suffixes and enclitics. These are a couple other examples of rules in the translation grammar:

```
{S,2}
S::S : [NP VP] -> [NP VP]
(  (X1::Y1)   (X2::Y2)

  ((x0 type) = (x2 type))

  ((y1 number) = (x1 number))
  ((y1 person) = (x1 person))
  ((y1 case) = nom)

; subj-v agreement
  ((y2 number) = (y1 number))
  ((y2 person) = (y1 person))

; subj-embedded Adj agreement
  ((y2 PredAdj number) = (y1 number))
  ((y2 PredAdj gender) = (y1 gender)))
```

```
{SBar,1}
SBar::SBar : [S] -> ["Dice que" S]
( (X1::Y2)
   ((x1 type) =c reportative) )

{VBar,4}
VBar::VBar : [V VSuff VSuff] -> [V]
( (X1::Y1)
    ((x0 person) = (x3 person))
    ((x0 number) = (x3 number))
    ((x2 mood) = (*NOT* ger))
    ((x3 inflected) =c +)
    ((x0 inflected) = +)
    ((x0 tense) = (x2 tense))
    ((y1 tense) = (x2 tense))
    ((y1 person) = (x3 person))
    ((y1 number) = (x3 number))
    ((y1 mood) = (x3 mood)))
```

The next version of the grammar will also cover transitive VPs (NSuff "-ta") and a few more complicated structures that appear in the development set.

Task 7: Spanish Morphology Generator

Even though Spanish is not as highly inflected as Quechua, there is still a lot to be gained from just listing the stems in the translation lexicon, and having a Spanish morphology generator take care of inflecting all the words according to the relevant features.

In order to do this, we obtained a morphologically inflected dictionary from the Universitat Politècnica de Catalunya (UPC) in Barcelona under a research license. Each citation form (infinitive for verbs and masculine, singular for nouns, adjectives, determiners, etc.) has all the inflected words listed with a PAROLE tag (http://www.lsi.upc.es/~nlp/freeling/parole-es.html) that contains the values for the relevant feature attributes. For example, here are some of the entries listed for the stem citation form “cantar”:

```
cantar#NCMP000 cantares
```

cantar#NCMS000 cantar
cantar#VMG0000 cantando
cantar#VMIC1P0 cantaríamos
cantar#VMIC1S0 cantaría
cantar#VMIC2P0 cantaríais
…
cantar#VMIF1P0 cantaremos
cantar#VMIF1S0 cantaré
…

Where the first slot corresponds to the part-of-speech (POS) and the rest of the slots are dependent on the POS. For example, the second slot to the fourth entry, the second slot represents type (main), the third mood (indicative), the fourth tense (conditional), the fifth person (first), the sixth number and the last slot gender.
In order to be able to use these Spanish dictionary, we mapped the PAROLE tags for each POS into feature attribute and value pairs in the format that our MT system is expecting. This way, the AVENUE transfer engine can easily pass all the citation forms to the Spanish Morphology Generator, once the translation has been completed, and have it generate the appropriate surface, inflected forms.

For examples of MT output which has been generated by the system with the morphology generator, see Appendix 2.

Task 8: Translation Correction Tool preliminary user studies

A preliminary user study of the correction of Quechua to Spanish translations was also conducted. For this user study, three Quechua speakers with good knowledge of Spanish evaluated and corrected nine machine translations, when necessary, through a user-friendly interface called Translation Correction Tool (TCTool) that was designed as part of Avenue's Rule Refinement module (Font Llitjós & Carbonell, 2004).

It was crucial for the development of the Rule Refinement moduleh to see how Quechua speakers used the TCTool and whether they had any problems with the TCTool interface. The user study showed that the representation of stem and suffixes as separate words in Quechua doesn't seem to pose a problem and that it was relatively easy to use for non-technical users.

The log files, resulting from user corrections still need to be fully analyzed to see what sorts of errors they corrected and how they corrected them.

Personal Statement (anecdotal)

Being able to spend almost three months in Cuzco over summer was definitely a worth while experience, not only from a work and research perspective, but also from a personal perspective.

Besides achieving the technical goals I had set for myself for the V-Unit project, I was able to work with native Quechua speakers in the data collection and got to know potential users of our Quechua-Spanish MT system well.

I believe that this collaboration provided for an enriching experience not just for me, but also for them. It definitely helped me grow a better vision of what technology is needed and what are the real technical needs of the Quechua-speaking community. A speech-to-speech Quechua-Spanish MT system in the medical domain, especially to assist Quechua monolingual women with their delivery, is a real improvement for the Quechua society in big cities like Cuzco. Currently, there are often no bilingual speakers even in several of the major hospitals who can assist with the communication between Quechua monolingual patients and Spanish monolingual doctors and nurses. This takes an often high tall on women's health and even lives.

Moreover, Quechua speakers I interacted with have shown a real interest for language and communication technology, especially when they found out it affects their language. It was easy to explain to them and have them immediately agree that it would be a great thing to have the local daily newspaper (El Sol) translated into Quechua, for example.

Crucially, this project has added a new dimension to my thesis research. I can now test my Rule Refinement approach on more than just English to Spanish. Since my thesis goal is to implement methods that are as language independent as possible, having a significantly different language pair, especially one that contains a real resource-poor language like Quechua, is essential to prove my thesis approach's generality.

During most of the time I was in Cuzco (June 27–August 20), I took an intensive Quechua course at the *Centro Bartolomé de las Casas* (CBC), especially designed for international students to dive into the Quechua language and culture. The CBC is a well established and well known institution that has been preserving and promoting the Quechua language and culture for several years now. In addition to the Andean School, it also has a Publishing House and what is called Casa Campesina (Rural House), which is a forum for Quechua-speaking farmers and rural people, and it provides them with legal advice and with a way to sell their products at a fair price.

The teacher in my class, Gina Maldonado, who has been teaching there since the beginning of the foundation of the Andean School, was the most pedagogical and fun language teacher that I ever had. The classes were excellent and always tried to bring in as many elements of the Quechua culture as possible. Gina proved to be a great informant for my project work, since she had good knowledge of the linguistic phenomena that occur in Quechua.

In particular, the Quechua classes proved to be an excellent resource when creating the suffix lexicon and in order to supervise and contribute to all the other tasks described above, since the teacher was the perfect informant to complement the information that can be found in the grammars.

During these three months, I also managed to find time to visit some of the most culturaly important places in the Peruvian Andes as well as to participate in a Conference on Bilingual and Multicultural Education. On July 11, I was invited to give a talk at the *2nd Congreso de Educación Bilingue Intercultural* in Velille – Chumbivilcas (7 hours by car from Cuzco), which I entitled ***Educación en lengua autóctona: el caso del catalán (Education in the autoctonous language: the case of Catalan)***. See Appendix 3 for the full paper in Spanish.

All in all, the V-unit has given me the opportunity to see the human face of my research and to better understand it as well as to discover a language and a people with enormous cultural heritage. For this, I am very grateful to Bernardine Dias and Manuela Veloso.

Conclusion

Machine Translation is a very difficult task and it is not one that can be satisfactorily solved for any given language pair in just a few months, or even a year. A prototype system like the one described here is by no means a final solution, but it is a first step towards the right direction. The V-Unit was the perfect chance for me to create the necessary infrastructure and electronic resources to start developing a Quechua to Spanish MT system.

There is now a Quechua-Spanish MT system prototype which can be expanded to start bridging the technology gap and thus empower the Quechua-speaking indigenous communities, through making their language and culture accessible to the rest of the world, and at the same time enabling them to assimilate crucial information only available in other languages.

More specifically, having an Quechua-Spanish MT system will allow them to assimilate government texts (and one day hopefully also speech) in their native language and thus have better access to fundamental issues that affect their lives, such as health care, plagues, land laws, and much more.

On the other hand, having a Quechua-Spanish MT prototype will allow me to run user studies and test the Rule Refinement module to automatically correct errors identified by non-expert Quechua and Spanish bilingual users.

One of the most vivid memories of my stay in Cuzco were the internet cafes, from where I often worked, filled with teenagers writing to each other messages of love… in Spanish.

The good news were young people were already massively using technology to communicate among them. However, Spanish seemed to be the only vehicular language, even though several of those teenagers were actually also Quechua speakers.

The ultimate goal of my research is to have a fully fledged Quechua-Spanish MT system which has high translation quality so that the gap between monolingual Quechua Hospital patients and monolingual doctors, say, can be effectively bridged. However, for such a system to be reliable and robust enough to be used in critical situations, a lot more work remains to be done.

Whereas my research remains just experimental and in the current state of things, I believe all it would take to start a real change in Andean society and future generations of Quechua speakers is for a company like Microsoft to offer a Quechua version of its tools, so that monolingual and native Quechua speakers could access and use technology in their own language.[3] This way, perhaps a third of the teenagers I saw impatiently and loudly gathering around the computers at the internet cafes, would feel more inclined to use their mother tongue to communicate with other teenagers.

Changes like this, which do not require the development of new technology but rather require making the existing technology accessible to resource-poor languages, would have a powerful effect in how the Quechua language is perceived by the young generation of Andean people. I believe that being able to use technology in their own language would increase the chances for Quechua to gain back its linguistic and cultural prestige.

References

Brown, Ralf D. (1997). *Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation*. Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-97).

Brown, Ralf and Robert Frederking. (1995). *Applying Statistical English Language Modeling to Symbolic Machine Translation*. Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95), pp. 221-239.

Centro Bartolomé de las Casas: http://www.cbc.org.pe/
Av. Tulluymayo 465, Tee: (0051) 84-233472

Cusihuaman, Antonio. (2001). *Gramatica Quechua. Cuzco Callao*. 2a edición. Centro Bartolomé de las Casas.

Font Llitjós, Ariadna; Carbonell, Jaime and Lavie Alon. (2005). *A Framework for*

[3] This would hopefully be done mainly for humanitarian and linguistic purposes, and not so much to make a profit out of it

Interactive and Automatic Refinement of Transfer-based Machine Translation. European Association of Machine Translation (EAMT) 10th Annual Conference. Budapest, Hungary.

Font Llitjós, Ariadna and Carbonell, Jaime. (2004). *The Translation Correction Tool: English-Spanish user studies*. International Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal.

Monson, Christian ; Levin, Lori; Vega, Rodolfo; Brown, Ralf; Font Llitjós, Ariadna; Lavie, Alon; Carbonell, Jaime; Cañulef, Eliseo and Huesca, Rosendo. (2004). *Data Collection and Analysis of Mapudungun Morphology for Spelling Correction*. International Conference on Language Resources and Evaluation (LREC).

Levin, Lori; Alison Alvarez, Jeff Good and Robert Frederking. (In Press). *Automatic Learning of Grammatical Encoding*. To appear in Jane Grimshaw, Joan Maling, Chris Manning, Joan Simpson and Annie Zaenen (eds) Architectures, Rules and Preferences: A Festschrift for Joan Bresnan , CSLI Publications.

Probst, Katharina. (2005). *Automatically Induced Syntactic Transfer Rules for Machine Translation under a Very Limited Data Scenario*. PhD Thesis. Carnegie Mellon.

Probst, Katharina and Lavie, Alon. (2004). *A structurally diverse minimal corpus for eliciting structural mappings between languages*. Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-04).

Probst, Katharina; Levin, Lori; Peterson, Erik; Lavie, Alon and Carbonell, Jaime. (2003). *MT for Resource-Poor Languages Using Elicitation-Based Learning of Syntactic Transfer Rules*. Machine Translation, Special Issue on Embedded MT.

Probst, Katharina; Brown, Ralf, Carbonell, Jaime; Lavie, Alon; Levin, Lori and Peterson, Erik. (2001). *Design and Implementation of Controlled Elicitation for Machine Translation of Low-density Languages*. Proceedings of the MT2010 workshop at MT Summit

The World Bank. *Indigenous People Leadership Capacity Building Program for Andean Countries*. Newsletter no. 1 (July 2003)

Appendix 1: Quechua statistics from the Ethnologue

http://www.ethnologue.com/show_language.asp?code=quz

Population	1,500,000 (1989 UBS). 300,000 to 500,000 monolinguals. Total Quechua speakers in Peru 3,500,000 to 4,400,000 including Quechua I 750,000, Quechua II 2,675,000 (2000 Adelaar). Ethnic population: 1,500,000.
Region	Departments of Cusco, half of Puno, and northeast Arequipa.
Alternate names	Cuzco Quechua, Quechua Qosqo-Qollaw, Runasimi Qusqu Qullaw, Quechua de Cusco-Collao, Qheswa, Quechua Cusco, Quechua de Cuzco
Dialects	Caylloma Quechua, Eastern Apurímac Quechua, Puno Quechua. Some dialect differences, but not as distinct as elsewhere. Substantial phonological and morphological differences with Ayacucho Quechua.
Classification	Quechuan, Quechua II, C
Language use	Official language. All ages. People in towns and cities generally want their children to primarily speak Spanish. In rural areas 65% may be bilingual, and in urban areas it might be 90% to 95%.
Language development	Literacy rate in first language: 1% to 5%. Literacy rate in second language: 62%. Taught in primary schools. Poetry. Radio programs. Dictionary. Grammar. Bible: 1988.

Appendix 2: MT output

Below are a few correct translations as output by the Quechua-Spanish MT system. For these, the input of the system was already segmented (and so they weren't run by the Quechua Morphology Analyzer), and the MT output is the result of inflecting the Spanish citation forms using the Morphological Generator:

1:
sl: taki ni
tl: CANTO
tree: <((S,1 (VP,0 (VBAR,2 (V,2:1 "CANTO")))))>

2:
sl: taki sha ni
tl: ESTOY CANTANDO
tree: <((S,1 (VP,0 (VBAR,3 (V,0:0 "ESTOY") (V,2:1 "CANTANDO")))))>

3:
sl: taki ra ni

tl: CANTÉ
tree: <((S,1 (VP,0 (VBAR,4 (V,2:1 "CANTÉ")))))>

4:
sl: taki sqa ni
tl: CANTABA
tree: <((S,1 (VP,0 (VBAR,4 (V,2:1 "CANTABA")))))>

5:
sl: taki sha ra ni
tl: ESTUVE CANTANDO
tree: <((S,1 (VP,0 (VBAR,5 (V,0:0 "ESTUVE") (V,2:1 "CANTANDO")))))>

6:
sl: taki ni taq
tl: Y CANTO
tree: <((SBAR,2 (LITERAL "Y") (S,1 (VP,0 (VBAR,1 (VBAR,2 (V,2:1 "CANTO"))))))))>

7:
sl: taki ra n si
tl: DICE QUE CANTÓ
tree: <((SBAR,1 (LITERAL "DICE QUE") (S,1 (VP,0 (VBAR,1 (VBAR,4 (V,2:1 "CANTÓ"))))))))>
…

10:
sl: taki nki taq
tl: Y CANTAS
tree: <((SBAR,2 (LITERAL "Y") (S,1 (VP,0 (VBAR,1 (VBAR,2 (V,2:1 "CANTAS"))))))))>
tl: Y CANTARÁS
tree: <((SBAR,2 (LITERAL "Y") (S,1 (VP,0 (VBAR,1 (VBAR,2 (V,2:1 "CANTARÁS"))))))))>

11:
sl: taki ra nki taq
tl: Y CANTASTE
tree: <((SBAR,2 (LITERAL "Y") (S,1 (VP,0 (VBAR,1 (VBAR,4 (V,2:1 "CANTASTE"))))))))>

12:
sl: taki ra nki chu
tl: CANTASTE ?
tree: <((SBAR,0 (S,1 (VP,0 (VBAR,1 (VBAR,4 (V,2:1 "CANTASTE"))))) (LITERAL "?")))>

…

15:
sl: noqa taki sha ni
tl: YO ESTOY CANTANDO
tree: <((S,2 (NP,1 (PRONBAR,1 (PRON,0:1 "YO"))) (VP,0 (VBAR,3 (V,0:0 "ESTOY") (V,2:2 "CANTANDO")))))>

16:
sl: qan taki ra nki taq
tl: Y TU CANTASTE
tree: <((SBAR,2 (LITERAL "Y") (S,2 (NP,1 (PRONBAR,1 (PRON,1:1 "TU"))) (VP,0 (VBAR,1 (VBAR,4 (V,2:2 "CANTASTE")))))))>
…

20:
sl: wasi
tl: CASA
tree: <((NP,2 (NBAR,1 (N,3:1 "CASA"))))>
tl: LA CASA
tree: <((NP,3 (DET,0:0 "LA") (NBAR,1 (N,3:1 "CASA"))))>

21:
sl: hatun wasi
tl: LA CASA GRANDE
tree: <((NP,4 (DET,0:0 "LA") (NBAR,1 (N,3:2 "CASA")) (ADJ,1:1 "GRANDE")))>
…

24:
sl: noqa qa barcelona manta ka ni
tl: YO SOY DE BARCELONA
tree: <((S,2 (NP,6 (NP,1 (PRONBAR,1 (PRON,0:1 "YO")))) (VP,3 (VBAR,2 (V,3:5 "SOY")) (NP,5 (NSUFF,1:4 "DE") (NP,2 (NBAR,1 (N,2:3 "BARCELONA")))))))>
tl: YO SOY DE LA BARCELONA
…

Appendix 3: ***Educación en lengua autóctona: el caso del catalán***.

Please find attached a Spanish version of the paper I presented in the 2nd Congreso de Educación Bilingue Intercultural in Velille – Chumbivilcas (Cusco).