Web-based translation of Quichua

Martha R. O'Kennon

The 10th Annual Internationalisation and Localisation Conference
13-14 September 2005
University Of Limerick, Ireland

Introduction.

In an age of pandemic language loss, finding mechanisms for rapid and accountable documentation and teaching of less commonly taught languages is more and more crucial. Many people are desperate to learn or contribute to the preservation of languages, and some institutions are attempting to teach these languages, but their materials are often home-grown and hence hard to locate and usually expensive to obtain from outside the institution. Web-based teaching materials are one path to wide, rapid, and cheap dissemination and documentation of these languages, and allow for dynamic feedback flow.

We have in the past decade been developing web-based sentence translators for documenting and teaching various African (Xhosa (2) and Pulaar) and North American (Ojibwe (4)) indigenous languages. In (2) we describe the process of making the translators run on the Web. Lately we have been trying to unify the translators, in the sense that the internal representation of a sentence should be as nearly standard as possible so that as little as possible has to be done from scratch when writing a new translator. This year we have been testing this approach to create translators between Spanish, English, and Quichua.

Quichua (known in Peru and Bolivia as Quechua) is a member (6) of the Amerind family, and of the Andean subfamily of languages. It is the most widely spoken indigenous language in Ecuador. Since it is in constant competition with Spanish as the language of upward mobility, many children are not fluent in this beautiful and expressive language. It is still spoken, however, by about two million people in the Ecuadorian highlands, so there are still many native speakers and teachers. Since Quichua falls into the category of a language (actually a family of dialects) suffering pressure from outside languages but still the mother tongue of many people, we have both time (lacking in the case of many other languages) and incentive to work on publishing a large part of the grammar on the Web.

Like the above African languages, and like many other languages, such as Turkish and Japanese, Quichua's verb structure is highly agglutinative and it shares many slots for verbal extensions with these languages. Like many other languages (including Turkish), Quichua has an elaborate system of cases, actually postpositional affixes. These kinds of

structures are so common that it makes sense to capitalize on the work done for one language when constructing a translator for another. In fact, large sections of the code (especially the code for causative, reciprocal, etc. verbal infixes) for the English to Quichua diagram translator module were borrowed from the corresponding module from the English to Xhosa translator.

We will explain how our translators work below, and give examples of some typical sentence patterns and how they are handled by the computer. But let us admit at the beginning that sentence translation is only one aspect of understanding how a language works. Ideally it should be an adjunct to actual sound and voice recordings of native speakers, films, printed stories and other documents, and other cultural information. And let us further admit that all translation, except for the most basic ideas, has inherent shortcomings and will never be complete. As such, our work is to be able to analyse a useful subset of one language and produce as correctly as possible the corresponding sentences in another.

As explained in [5], the translators work in three steps: **parsing** the source language sentence into the source language sentence diagram, **translating** the source language diagram into the target language diagram, and **synthesizing** the target language diagram into the target language sentence. One big advantage of this parse-first philosophy is that we can avoid many of the errors of some of the web-available translators, which translate on the fly (word-by-word), and hence sometimes produce sentences that are at best in a strange word order, and at worst incomprehensible. Ideally the parsing process, from (say) English to English sentence diagram, should be as nearly independent as possible of the target language. In this way, we get the second advantage that the module for parsing may be reused by translators from English to any other language. Similarly the synthesizing process for a particular target language should be capable of being reused in any translator into that language. This means that with a great deal of collaboration it should be possible to create multiple translators for various dialects (like Xhosa and Zulu) and also of not-so-closely-related languages that share underlying structural similarities.

How the English to Quichua translator works.

We start by introducing the internal structure or sentence diagram which we use to represent a sentence. We distinguish various kinds of sentences: declarative s(nounphrase, verbphrase), yes-no questions q(nounphrase,verbphrase), informational questions w(question word, nounphrase, verbphrase), and a few others.

Our first example is a basic sentence in which the structure of the English diagram is practically the same as the structure of the Quichua diagram. This does not mean the structure of the sentences is the same, but that they contain semantically the same ingredients.

Example 1. Suppose we are translating from English to Quichua, and input this basic sentence the boys and i gave the girl a cat.

The first stage produces the English sentence diagram

```
s("past",                                                    declarative sentence, past tense
nps([n(det1("the"),[],["boy"],[],"3","plural","p","nom"),    first subject
p(["i"],[],"1","singular","i","nom")],                       second subject
"1","plural",""),                                            aggregate person and number
v([["give"]],"past","ind",[],"active","1","plural",          verb, tense, mood, voice
nps([n(det1("the"),[],["girl"],[],"3","singular","p","dat")],"3","singular","p"),
                                                             indirect object
nps([n(det1("a"),[],["cat"],[],"3","singular","pet","acc")],"3","singular","pet"),
                                                             direct object
nps([],"","","",""),                                         agent of passive sentence
adve(places([]),places([]),places([]),places([]),places([]),times([]),manners([]),using([]),
howmuch([]),with1(epsilon),reasons([]),becauses([]),advs1([])),
                                                             various kinds of adverbs
"t","pos"))                                                  transitive, intransitive, etc.
```

This parsing process consults the lexicon to find that we have an ordinary declarative sentence with a compound subject nps(nounphrase, nounphrase) (nps is our symbol for "nounphrases"): ("the boys and i") where "boys" is a third person singular noun, and "i" is the first person, singular pronoun that occurs in a subject context (or "nominative" case). A calculation is made to resolve the aggregate subject person and number to first plural. This same calculation can be varied for other languages. If we were translating from English to Pulaar, Ojibwe, or Peruvian Quechua, where two forms of "we" ("inclusive" and "exclusive") exist, this would resolve to "first plural exclusive", since we have a combination of first person and third person, and does not involve the person spoken to.

Embedded in the lexicon are simple semantic clues as to the use of the various nouns. In this example both the "boy" and the "girl" are marked as "p" for "persons", while the "cat" is marked as a "pet". We will see later where these semantic markers are necessary for disambiguating parts of sentences.

The lexicon also shows that "gave" matches the template of a verb ("give") in the first person, plural number, simple past tense, indicative mood, and active voice. Now come two noun phrases not connected by "and". We now realize that we have an indirect object ("the girl") and a direct object ("a cat"). These are placed in the verb frame, which contains three fields, the first for the indirect object, the second for the direct, and the third for an agent in a passive construction. This latter is missing in the present example. (A large frame for adverbial structures is also empty, but we shall soon see how it is used.)

In the second phase, this is translated into the Quichua sentence diagram

s("past",
nps([
n(epsilon,[],["huanpra"],[],"3","plural","","nom"),
p(["n~uca"],[],"1","singular","n~uca","nom")],"1","plural",""),
v([["cu"]],"past","ind",[],"active","1","plural",
nps([n(det1("ca"),[],["cuitsa"],[],"3","singular","","dat")],"3","singular",""),
nps([n(det1("shuc"),[],["misi"],[],"3","singular","pet","acc")],"3","singular","pet"),
nps([],"","",""),
adve(places([]),places([]),places([]),places([]),places([]),times([]),manners([]),using([]),howmuch([]),with1(epsilon),reasons([]),becauses([]),advs1([])),"t","pos"))

This involves a second visit to the lexicon, in which we match the items. We find "huanpra" for "boy", "misi" for "cat", and "cuitsa" for "girl" in the list of nouns. We also find that the corresponding first person singular nominative pronoun is "ñuca" (in the older Castellano-based spelling – it would be ñuka in the more modern spelling being adopted in Ecuador). We look up "love" in the verb section of the lexicon and find "cuya" ("kuya") as its equivalent. No big changes in tense, mood, or case are called for, so in this example the Quichua sentence diagram is almost identical to the English one.

In the third step, this Quichua sentence diagram is synthesized into the Quichua sentence

huanpracunapash ñucapash cuitsaman misita curcanchicmi.

So that the reader can compare the two sentences, we have color-coded the various items: the subject in red, plural in pink, indirect object in purple with the dative postfix "man" in magenta, direct object in blue with the accusative postfix "ta" in sky blue. The verb is marked in teal, with the past tense in aqua, and the first person plural by brown. A focus particle "mi", marked in bright green, is added to the end of the verb. The focus particle tells us that the verb is stressed in this sentence. In the Quichua sentence this focus particle can be added in a number of different places, as we will see.

Note that most of the diagram is occupied by slots or place holders for other syntactic/semantic fields, many of them adverbial in nature. We will later discuss how these adverbial structures can be used in producing the correct Quichua answer to an informational question (as opposed to a yes-no question). Also note that the order of the words in this sentence is identical to the order in a Japanese or Turkish sentence, with the verb final as is usual. One further note: we have not stressed the article "a" ("shuc", which also can be rendered as "one" or "another").

Since there are two competing systems of orthography, the output on the website is printed in both. So the output is first given as “ñucaca canta cuyanimi”, and then converted by replacing “c” by “k” (unless it occurs in the combination “ch”), “hua” by “wa”, “ui” by “uy”. So the target sentence is postprocessed to give

wanprakunapash ñukapash cuytsaman misita kurkanchikmi.

A complete list of spelling changes is given in http://mokennon2.albion.edu/qhelp3.htm. Note that in the newer orthography, the name of the language would be “Kichwa”, but in this paper we retain the spelling “Quichua” as being more familiar for most people.

Quichua also has an elaborate system of postpositional affixes (“case endings”), which we model as part of a large frame for holding various types of adverbial constructions, mostly prepositional/postpositional phrases. In the previous example, the frame was empty. But in the following examples, we will see how various kinds of adverbs are sorted into the frame.

We will also see in the next examples how verbal extensions are used to code for a number of semantic functions.

Example 2.

i am going with my mother to town in a car

s("prog",
nps([p(["i"],[],"1","singular","i","nom")],"1","singular","i"),
v([["go"]],"prog","ind",[],"active","1","singular",
nps([],"","","",""),nps([],"","","",""),nps([],"","","",""),
adve(places([pp(["in"],nps([n(det1("a"),[],["car"],[],"3","singular","","in")],"3","singular",""))]),places([pp(["to"],nps([n(epsilon,[],["town"],[],"3","singular","","man")],"3","singular",""))]),places([]),places([]),places([]),times([]),manners([]),using([]),howmuch([]),
with1(pp(["with"],nps([n(poss([p(["my"],[],"1","singular","i","gen")]),[],["mother"],[],"3","singular","pf","com")],"3","singular","pf"))),reasons([]),becauses([]),advs1([])),"i","pos"))

When translating this sentence diagram to the Quichua diagram, note that the progressive “tense” in the English diagram changes to present tense in the Quichua diagram. The reason for this is that the progressive is rendered in Quichua by the addition of a verbal extension “cu” to the verb. So in addition to changing the tense, we add this extension to the verb in the new diagram.

s("pres",nps([p(["n~uca"],[],"1","singular","n~uca","nom")],"1","singular","n~uca"),v([["ri"],["cu"]],"pres","ind",[],"active","1","singular",nps([],"","","",""),nps([],"","","",""),nps([],"","","",""),adve(places([pp(["pi"],nps([n(det1("shuc"),[],["carro"],[],"3","singular","","in")],"3","singular",""))]),places([pp(["man"],nps([n(epsilon,[],["llacta"],[],"3","singular","","man")],"3","singular",""))]),places([]),places([]),places([]),times([]),manners([]),using([]),howmuch([]),with1(pp(["huan"],nps([n(poss([p(["n~uca"],[],"1","singular","n~uca","gen")]),[],["mama"],[],"3","singular","pf","com")],"3","singular","pf"))),reasons([]),becauses([]),advs1([])),"i","pos"))

ñukaka llactaman carropi ñuka mamawan rikunimi

Here the framework distinguishes the three prepositional phrases by semantics. (The five slots labeled "place" correspond to locative, motion to, motion from, motion along, and motion up to.)

Note that the "with" in English semantically can stand for either a comitative or instrumental preposition, not to mention many other shades of meaning ("with a glance", "with a heavy heart", etc.). Since "mother" is marked as "pf", ("person in family"), and hence a person, we know that "with" is comitative, and the prepositional phrase is placed in the slot "with1" (accompaniment). If the sentence had been "i ate my food with a spoon", this "with" would have been marked as "instrumental", and the prepositional phrase would have been placed in the appropriate slot. In fact, in Quichua, the same suffix ("huan" or "wan") will be added to the noun. However in other languages, the difference between the comitative and the instrumental is registered by such other mechanisms as verbal extensions (Pulaar), different prepositions (Xhosa – "na" vs "nga"), or different case endings (Hungarian). Here is a pair of Pulaar sentences to show this difference (5, p. 6).

Example 3.

i ate with a fork
mi ñaamrii salndu

i ate with John
mi ñaamdii e John

This pair of examples shows that it is important to get as much semantic information as possible into the sentence diagram structure if we wish to generalize to many different languages.

Focus in Quichua.

Note that in the absence of an informational question, we do not know in what kind of dialogue this sentence "I went with my mother to town in a car" might have appeared. It could have been the answer to "where did you go with your mother in a car?", "in what did you go to town with your mother?", or "with whom did you go to town in a car?". In English we do not have a way in writing to distinguish this focus. But in Quichua it must be displayed. The focus could have been in any position, affixed to "llactaman", "carropi", "mamawan", or the verb "rikuni". We have made the choice in the circumstance where the sentence appears in isolation to affix the "mi" to the verb unless it is the verb "kana" ("to be"), where the "mi" is never added.

Before we discuss how the program identifies the focus in a question-answer pair, let us first say a bit about the informational question-answer pair. Most Quichua "question words" are made up of at three parts: the "basic question type", such as "where", "what", "who"; a "case" part, such as "to", "from", "with"; and a question marker ("tak" for present or past, "shi" for future or conditional). In example 4a, "maymantak" is made up of "may" ("where"), "man" (motion to or dative), and "tak" (question-former). When the answer is processed, the case ending is added to the answer noun, and the final position is occupied by the focus particle "mi". Note that the answer to the question "maymantak" is "llactamanmi", where the "may" is replaced by the answer "llacta" ("town"), the "man" ("towards") occupies the middle, and the "tak" is replaced by the "mi" (3, p. 6-7).

Example 4a.

? where are you going with your mother in a car ?
i am going with my mother to town in a car

¿maymantak carropi kanpa mamawan rikunki ? ?
ñukaka llactamanmi carropi ñuka mamawan rikuni

As in the other examples, the English sentence pair is first parsed into a sentence diagram which expresses the syntax and to some extent the semantics of the sentence. The only

complicated part of this process is correctly interpreting “where” as “to where”. (This was a problem not encountered in the Spanish parser, since the question word would have been “a dónde?” or “adónde?”) To disambiguate this use of “where”, we have to examine the verb “go”. Since “go” is semantically associated with motion towards something, we supply the motion preposition “to”.

This diagram is translated into the corresponding Quichua sentence diagram, with “to where” replaced by “maymantak”. In order to translate the Quichua sentence diagram into Quichua, since the sentence is in reality a sentence pair a special rule is activated. Upon seeing that the first of the sentences is an informational question sentence, after translating the question itself, the question word is put into temporary memory as a focus reminder to guide in translating the adverb structure of the answer. Finally, when the adverb structure is being rendered into Quichua, only the section relevant to the question focus has the “mi” attached to it.

Example 4b. with whom are you going to town in a car ? i am going with my mother to town in a car

¿ piwantak llaktaman carropi rikunki ? ?

ñukaka llaktaman carropi ñuka mamawanmi rikuni

Here the question was “with whom” and so only the part of the sentence which answers this question (the “with my mother” or “ñuka mamawan”) is marked with the “mi”.

Example 4c.

We would like to give special mention of one of the adverbial slots, the one which answers the question “for what purpose?”, and is of the general form “in order” followed by an infinitive phrase.

? why are you coming to Albion ? i am coming to Albion in order to eat with you in your house

Note that in this example, there are two sets of adverbs, one occurring inside the other. The outside adverbial frame contains the slot for **reasons**, this being actually an infinitive

phrase “in order to eat…”. But this infinitive phrase itself contains not only a verb, but also the set of adverbs belonging to this verb.

¿ imapaktak Albionman shamukunki ?
ñukaka kanpa wasipi kanwan mikunkapakmi Albionman shamukuni

(Note that in this example “your” = “kan + pa” (“you” + “r”), whereas the “pa” is not needed in the previous example for the first person possessive.)

Verbal infixes.

As mentioned before, Quichua is an agglutinative language, like Turkish, Japanese, and many African languages. It possesses at least 10 distinct morphemes which may be infixed between the verb and its tense endings to refine the meaning of the verb. Among these, the most ubiquitous are the causative, the reflexive, the reciprocal, the progressive (already seen in example 2), and the inchoative (“to be going to”).

Many English words with causative effect are translated into Quichua words which contain the causative infix. For instance we have many pairs like “eat” – “feed” (“miku” – “mikuchi”), “see” – “show” (“riku” – “rikuchi”), “die” – “kill” (“wañu” – “wañuchi”). Unlike in English, this infix is completely productive in Quichua and the above languages.

Example 5. i made the children eat

s("past",nps([p(["i"],[],"1","singular","i","nom")],"1","singular","i"),c(["make"],"past","ind",[],"active","1","singular",
nps([n(det1("the"),[],["child"],[],"3","plural","","acc")],"3","plural",""),v([["eat"]],"past","subj",[],"active","3","plural",nps([],"","","",""),nps([],"","","",""),nps([],"","","",""),adve(places([]),places([]),places([]),places([]),places([]),times([]),manners([]),using([]),howmuch([]),with1(epsilon),reasons([]),becauses([]),advs1([])),"i","pos"),"t","pos"))

s("past",nps([p(["n~uca"],[],"1","singular","n~uca","nom")],"1","singular","n~uca"),c(["make"],"past","ind",[],"active","1","singular",
nps([n(epsilon,[],["huahua"],[],"3","plural","","acc")],"3","plural",""),v([["micu"],["chi"]],"past","subj",[],"active","1","singular",nps([],"","","",""),nps([],"","","",""),nps([],"","","",""),adve(places([]),places([]),places([]),places([]),places([]),times([]),manners([]),using([]),howmuch([]),with1(epsilon),reasons([]),becauses([]),advs1([])),"i","pos"),"t","pos"))

ñukaka wawakunata mikuchirkanimi

In this example, we chose to use “make eat” instead of “feed” (“mikuchi”) to show how the English sentence is transferred into a special causative frame, which contains the “children” as the subject of the verb phrase “eat”. This sentence diagram is then translated into the Quichua diagram in which the “chi” has been added to the list of extensions.

Example 6. This example uses three verbal infixes.

i am making the children help each other

s("prog",nps([p(["i"],[],"1","singular","i","nom")],"1","singular","i"),
c(["make"],"prog","ind",[],"active","1","singular",nps([n(det1("the"),[],["child"],[],"
3","plural","","acc")],"3","plural",""),v([["help"],["recip"]],"prog","subj",[],"activ
e","3","plural",nps([],"","","",""),nps([],"","","",""),nps([],"","","",""),adve(places([]),places([]),pl
aces([]),places([]),places([]),times([]),manners([]),using([]),howmuch([]),with1(epsilon),r
easons([]),becauses([]),advs1([])),"t","pos"),"t","pos"))
s("past",nps([p(["n~uca"],[],"1","singular","n~uca","nom")],"1","singular","n~uca"),c([
"make"],"past","ind",[],"active","1","singular",
nps([n(epsilon,[],["huahua"],[],"3","plural","","acc")],"3","plural",""),v([["yanapa
"],["nacu"],["chi"],[“cu”],"past","subj",[],"active","1","singular",nps([],"","","",""),nps([]
,"","","",""),nps([],"","","",""),adve(places([]),places([]),places([]),places([]),places([]),times([]
),manners([]),using([]),howmuch([]),with1(epsilon),reasons([]),becauses([]),advs1([])),"t"
,"pos"),"t","pos"))

ñukaka wawakunata yanapanakuchikunimi

The program is capable of several levels of infixes. In actual Quichua, three or four is usually the upper limit. In this example, we have the progressive, causative, and reciprocal used together. Note that the English translation of the infixes runs from right to left.

In translating present or past progressive verbs, the English sentence diagram shows tense “prog” or “pastprog”, but when we form the Quichua sentence diagram the “prog” turns into “pres” (and “pastprog” turns into “past”) but the infix “cu” is added to the list of extensions.

There is overlap, though not perfect, in the sets of verbal infixes used in Quichua and other languages we have studied. In languages such as Xhosa and Turkish, the passive is realized as a verbal infix, but in Quichua it is realized as in English, that is, as a periphrasis of the perfect participle plus the verb to be. In Quichua, on the other hand, the progressive is realized as a verbal infix, which is not the case in the other languages. However, we were still able to copy a substantial amount of code from our Xhosa translator with only tiny changes, economizing greatly the effort to produce the English to Quichua sentence diagram translators.

Conclusion and a look to the future.

According to (1) and others, a finite number of "parameters" suffice to classify languages, not by their family trees, but by their structural characteristics (head-directionality, polysynthesis, etc.). As we conclude in (5, p. 8), we want to continue investigating these structural categories to start making a library of modules that can be "mixed and matched" to produce new programs for languages which are possibly not related, but share basic characteristics. This method should make it possible for us not only to easily produce multiple translators for languages that are "dialects" of each other but also to clone code for some parts of languages that are not related. In the face of great urgency we would like to "identify collaborators for a long-term project, especially people who can teach others how to make translators to document and teach their own languages".

Acknowledgement.

In the spring of 2005, I went to Quito in search of a teacher of Quichua with whom to work. At the Universidad San Francisco de Quito, I was lucky to find José Maldonado Córdova, a gifted teacher of Imbabura Quichua. His passion for his language, good humor, and willingness to work with me through my many many questions has been the major contributing factor to the present state of this project.

Biography:

Martha O'Kennon is Professor Emerita of Mathematics and Computer Science from Albion College in Michigan. She has developed and taught, in addition to more standard classes, language-related courses such as "Computer Understanding of Human Language", in which one project has students write small translators from English to a language of their own interest; and "Survey of Human Languages", an introduction to comparative language for undergraduates. In the past decade, all her research has been dedicated to documenting and teaching less commonly taught languages through web-based translation programs.

References:

1. M. Baker, *The Atoms of Language*, Basic Books, New York, 2001.

2. D. Barber and M. O'Kennon, "Xhosa on the Web", Journal for Computing in Small Colleges, Fall 1997.

3. J. Maldonado, *Kichwata yachakunkapak*, notes; to appear in book form, December, 2005.

4. M. O'Kennon, "Anishinaabemowin on the Web" (abstract), Michigan Academy Conference, March 5, 2004

5. ._, "Documenting and Teaching Languages through Web-based translation", Fourth Conference on Preserving African Languages, University of Maryland - Eastern Shore, Salisbury, MD, Nov. 4-7, 2004.

6. M. Ruhlen, *A Guide to the World's Languages, Vol. 1: Classification*, Stanford University Press, 1991.

7. http://mokennon2.albion.edu/language.htm has links to the translators that are in more advanced stages.