Building Machine translation systems for indigenous languages

Ariadna Font Llitjós, Lori Levin
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
aria@cs.cmu.edu

Roberto Aranovich
Department of Linguistics
University of Pittsburgh
roa6+@pitt.edu

Key Words: natural language processing, machine translation, Mapuche, Mapudungun, Quechua, indigenous communities

1. Introduction

In this paper we focus on the cooperation between a team of computational linguists and two communities of indigenous language speakers in Latin America, Mapuche in Chile (2002-2005) and Quechua in Peru (2004-2005). In both cases, this cooperation was embraced by AVENUE, a project devoted to fast and affordable development of Machine Translation (MT) systems for resource-poor languages. With respect to machine translation, "resource poor" refers to the lack of a large corpus in electronic form or lack of native speakers trained in computational linguistics. There may be other difficulties as well, such as spelling and orthographical conventions that are not standardized and missing vocabulary items. As part of our collaboration, the members of the communities compiled corpora and other resources such as vocabulary lists in their languages. The AVENUE team provided expertise in Natural Language Processing in order to develop morphological analysis, spelling correction, and ultimately, an MT system.

1.1. AVENUE project

Machine Translation is not available for the majority of the world's languages, the prohibitive factors being the time and expense involved in acquiring corpora in electronic form and training computational linguists. The AVENUE project is focused on reducing the cost of producing MT systems in an effort to make them available for more languages. There are many types of MT systems, each requiring different resources. The AVENUE 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

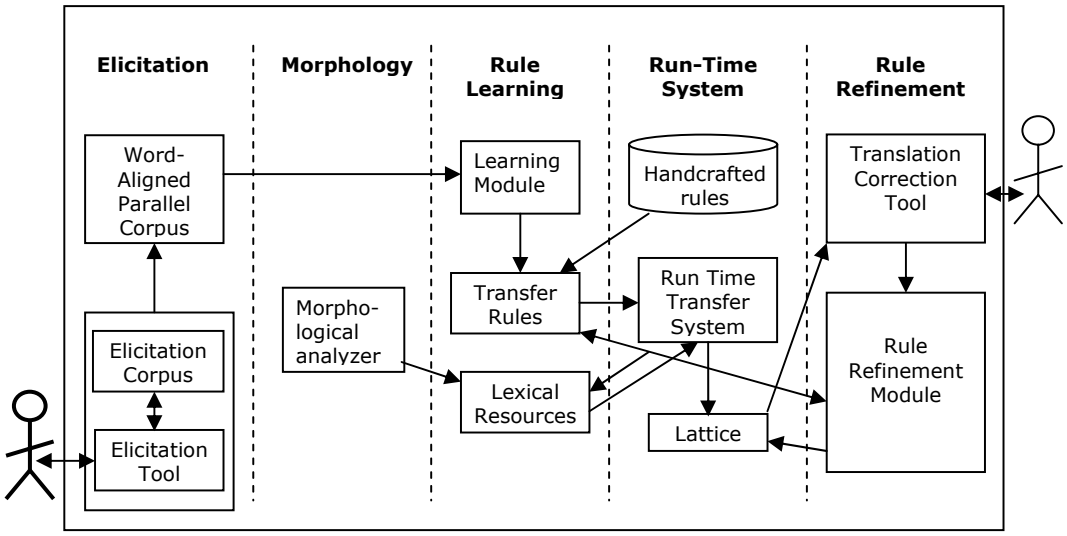
of rules can be developed. Finally, if neither a corpus nor a human computational linguist is available, AVENUE uses a machine learning technique called Seeded Version Space Learning (Probst, 2005) to learn translation rules from data that is elicited from a native speaker.

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).

Since the Mapuche community was able to collect a large parallel corpus of Mapudungun and Spanish, we were able to apply EBMT. Also, since one of the authors of this paper (Aranovich) has knowledge of Mapudungun and computational linguistics, we were able to produce a set of handwritten MT rules. Automatic rule learning has been applied experimentally in Hindi-to-English MT (Lavie *et al.* 2003) and Hebrew-to-English MT.

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 automatically refines the translation rules (Font Lltjós *et al*. 2005).

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



1.2. Collaboration between the CMU team and Indigenous Communities

In a collaboration between the CMU AVENUE team and an indigenous community, each partner brings critical skills. The indigenous community has knowledge of the language and the needs of the community. They must be involved in the design of the machine translation system because they are the speech community who will use it for communication. Even if a government agency is involved (such as the Ministry of Education in Chile), an indigenous community must also be involved. CMU provides expertise on audio recording of data, formatting of data, and, of course, machine translation.

1.3. The CMU team

The members of the AVENUE team at CMU have many sub-specialties in computer science, linguistics, and international development. Jaime Carbonell, the director of the AVENUE project is a computer scientist with expertise in machine learning and many areas of language technologies. Alon Lavie (co-director of AVENUE) is a computer scientist with expertise in parsing algorithms and machine translation. Lori Levin (co-director of AVENUE), a linguist, also has expertise in machine translation, and provides linguistic supervision to the team. Ralf Brown is a computer scientist and leading expert in Example Based MT. Robert Frederking is also a computer scientist with expertise in both rule based MT and Example Based MT. Rodolfo Vega is an expert in international development specializing in the use of technology in education in developing countries; he serves as the liaison between the CMU team and the government agencies and indigenous communities. Ariadna Font Llitjós is a PhD student, working on Mapudungun and Quechua, particularly on interactive and automatic refinement of translation rules. Christian Monson is a PhD student focusing on the implementation of

Mapudungun morphology, coordinating the integration of the components of the Mapudungun MT system, and doing research on the automatic learning of morphemes.
Erik Peterson is the main developer of the interface that is used for eliciting data from informants and also built the transfer engine, which runs the translation rules, and the decoder, which selects the best translation from a large lattice of possible translations.
Kathrin Probst is an alumna of the AVENUE project whose PhD research covered the automatic learning of translation rules. Pascual Masullo is a linguist at the University of Pittsburgh. He contributed knowledge of the linguistic analysis of Mapudungun. Roberto Aranovich, a PhD student at the University of Pittsburgh, implemented the hand written transfer rules for Mapudungun and Spanish and contributed to the development of the morphological analyzer and lexicon. An alumnus of the project, Carlos Fasola, was the main developer of the morphological analyzer for Mapudungun.

2. Mapudungun Cooperation

Mapudungun is spoken by over 900,000 people (Mapuche) in Chile and Argentina. The Chilean Ministry of Education has created programs to ensure that each child's cultural and linguistic needs are met in school. In collaboration with the AVENUE team at LTI, the Ministry of Education has provided support for the collection of data and other tasks related to the building of an MT system for Mapudungun. The data collection was carried out by native speakers of Mapudungun at the Universidad de la Frontera in Temuco, Chile. The products of the collaboration have been a small Mapudungun-Spanish parallel corpus of historical texts and newspaper text and a large parallel corpus consisting of 150 hours of transcribed speech in Mapudungun, which has been translated into Spanish. In addition, frequency-ordered word lists have been created from the corpus. A spelling checker was developed based on the stems and suffix groups in the word list (Monson et al. 2004). The spelling checker uses one boundary between a stem and a list of suffixes. A more sophisticated morphological analyzer was also developed, which identifies all of the suffixes attached to a stem. Experimental MT systems (EBMT and handwritten rules) are currently being tested.

The collaboration between UFRO and CMU consisted of planning meetings and training sessions in Temuco, extended visits to Temuco by Spanish speaking members of the AVENUE CMU team, and heavy use of email and telephone.

2.1. Chilean team

In a preliminary meeting in May, 2000, representatives of CMU's Language Technologies Institute met with Mapudungun language experts at the Instituto de Estudios Indigenas (IEI - Institute for Indigenous Studies) at the Universidad de la Frontera (UFRO). We agreed to collaborate in building language technologies to respond the demands of intercultural bilingual education programs for the Mapuche. Soon afterward, the Bilingual and Multicultural Education Program of Ministry of Education (Mineduc) agreed to participate in the project, and to fund most of the research that was planned to take place in Chile.

The Chilean AVENUE team includes members of the Ministry of Education, UFRO, and the Mapuche community. Carolina Huenchullan Arrúe is the National

Coordinator of the Bilingual Multicultural Education Program in the Ministry of Education in Chile. Also in the Ministry of Education, Claudio Millacura Salas is Pedagogical Coordinator (encargado pedagógico). At IEI-UFRO, the team coordinator is Eliseo Cañulef, a specialist in intercultural bilingual education. Rosendo Huisca is an expert in the Mapudungun language and a long-time proponent of its use. Hugo Carrasco, a Linguist, is UFRO's Dean of the Humanities and Education Faculty. Hector Painequeo, also a linguist is a professor at UFRO. Flor Caniupil is the senior member of the transcription and translation team. Luis Caniupil Huaiquiñir, the data collection specialist, conducted most of the interviews in the spoken language corpus. Marcela Collio Calfunao and Cristian Carrillan Anton are members of the transcription and translation team. Salvador Cañulef is a computer and software support specialist. Except for Dr. Carrasco, All members of the IEI-UFRO team are of Mapuche descent. Several are native speakers.

2.2. Mapudungun Database

The AVENUE-Mapudungun team (consisting of the US and Chilean participants) collected, transcribed, and translated a Spanish-Mapudungun parallel corpus that could be used for corpus-based language technologies (language technologies that do not involve human rule engineering) and could also be used for corpus linguistics or corpus-based computer-assisted language learning. The corpus has two main parts: written texts and transcribed speech. Both parts of the corpus (written and spoken) were collected the IEI-UFRO team.

2.2.1. Written Corpus

The written Mapudungun corpus consists of historical documents and current newspaper articles. The two historical texts are *Memorias de Pascual Coña*, the life story of a Mapuche leader written by Ernesto Wilhelm de Moessbach; and *Las Últimas Familias* by Tomás Guevara. The two historical texts were first typed into electronic form as exact copies of the originals and then were transliterated into the orthographical conventions chosen by AVENUE-Maupdungun. The modern newspaper, *Nuestros Pueblos* is published by the Corporación Nacional de Desarrollo Indígena (CONADI). The length of the text corpus is about 200,000 words.

2.2.2. Speech corpus

The spoken Mapudungun corpus consists of 170 hours of Mapudungun speech. The corpus consists of interviews, most of which were conducted by Luis Caniupil Huaiquiñir, a native speaker of Mapudungun. The recordings were transcribed and translated into Spanish at the IEI, UFRO. They cover three dialects, 120 hours of Nguluche, 30 hours of Lafkenche and 20 hours of Pewenche. The Williche variant was not collected because it presents some morpho-syntactic differences, specifically in the pronouns and verb conjugations.

The subject matter of the spoken corpus is primary and preventive health, both Western and Mapuche traditional medicine. The informants are asked to tell their experiences on illnesses and remedies that they or their relatives have experienced. They

are asked to provide a complete account of symptoms, diagnostics, treatments, and results. For an excerpt from the spoken corpus, see Figure 2.

The ages of informants are between 21 to 75 years old, most of them between 45

Figure 2: *Excerpt from the corpus of spoken Mapudungun*

nmlch-nmjm1_x_0405_nmjm_00:
M:<SPA>no pütokovilu kay ko
C: no, si me lo tomaba con agua

M:chumgechi pütokoki femuechi pütokon pu <Noise>
C: como se debe tomar, me lo tomé pués

nmlch-nmjm1_x_0406_nmlch_00:
M:Chengewerkelafuymiürke
C: Ya no estabas como gente entonces!

and 60 years old. All informants are fully native speakers. Most informants work as auxiliary nurses in rural areas of the Chilean Public Health System, or are knowledgeable in traditional Mapuche medicine. They did not reveal any culturally sensitive information about Mapuche medicine.

The dialogues were recorded using a Sony DAT recorder (48kHz) and Sony digital stereo microphone. The tapes are downloaded using CoolEdit 2000 v.1.1 (http://www.syntrillium.com/cooledit). For transcription, we use the TransEdit transcription tool v.1.1 beta 10, developed by Susanne Burger and Uwe Meier[1]. The software synchronizes the transcribed text and the wave file. It also shows the actual wave, making it easy to identify each speaker turn as well as simultaneous speakers.

The transcribers use the LTI's transcription conventions for noises and disfluencies including aborted words, mispronunciations, poor intelligibility, repeated and corrected words, false starts, hesitations, undefined sound or pronunciations, non-verbal articulations, and pauses.

Foreign words, in this case Spanish words, are also labeled. The entire corpus is transcribed using orthographic conventions that were established by the IEI-UFRO team.

However, recently a different orthography, azümchefi, has been chosen by the government. The corpus has been converted automatically into azümchefi using substitution rules.

[1] For more information about TransEdit, contact sburger@cs.cmu.edu.

2.3. Developing Natural Language Processing Tools

2.3.1. *Bilingual Lexicons*

Bilingual lexicons were constructed from the spoken language corpus. All the unique words were extracted from the spoken corpus, and then they were ordered by frequency. This word frequency list was then used as a guide for translation dictionary development. There were two main different dictionary development efforts. One effort was lead by the Chilean team, to create an online translation dictionary with examples of usage (1,926 entries). See Figure 3 below for all the mandatory fields included in the dictionary. Optional fields included POS, Pronunciation, Explanation (encyclopedic and cultural description; for example, *machi*: specialist in Mapuche medicine and ritualism), Connotation (in case the Spanish translation looses part of the connotations contained in the Mapudungun word) and Synonyms.

Figure 3: *Fields in the Mapudungun-Spanish dictionary elaborated by the Chilean team.*

1. Full form Mapudungun word (in supra-dialectal alphabet)
2. A segmentation of the word into morphemes (root + suffixes)
3. A gloss for each morpheme
4. Translation into Spanish
5. Example of usage:
   - A sentence from the corpus of spoken Mapudungun containing the word form, where it has the translation indicated in 4.
   - A Spanish translation of the sentence, and
   - A reference into the corpus of spoken Mapudungun identifying the specific cited sentence

4 contains sample entries from among the 1,926 in the translation dictionary. The dictionary is in a very general text-only format that can be re-configured for any computer-based lexicon interface. The morphemes were labeled by native speakers who are not linguists. They used glosses that are consistent, but do not follow linguistic terminology. For example, *él(ella).a.ti* means third person singular acting on second person singular. (A more detailed segmentation might be *e-ymu* where the first morpheme indicates that the object, in this case second person, outranks the subject, in this case third person, and the scond morpheme agrees with the higher ranking noun, in this case, second person.) The Chilean team is currently finalizing the last design and implementation details to be able to put translation dictionary online.

The other dictionary development effort was lead by the LTI team, originally derived from the first one, to create a translation lexicon for the MT systems, which included just the translations as well as some additional features necessary for the correct application of the translation rules. This effort is on a larger scale (66,413 Mapudungun fully-inflected word forms, automatically extracted from the spoken corpus), but with only grammatical features such as number and person in each lexical entry.

Figure 4: *Entries from the UFRO Translation Dictionary*

Kümekünueymu: küme-künu-eymu.bien-quedar-él(ella).a.ti .? . / /. te ha dejado muy bien. Ka kümekünueymu tati. (Y te ha dejado muy bien). nmlch-nmpll1_x_0070_nmlch_00. EC/RH03-02-03.

Lichi: .? . / /. leche. Feychi lichi, ¿chem lichingey? (Esta leche ¿qué leche es?) nmlch-nmfhp1_x_0051_nmlch_00. Ec/Rh/Fc. Ec/ Rh02-01-03.

Mongepeürkelayan: monge-pe-ürke-la-y-a-n.sanar-tal.vez-acaso-no-0-futuro-yo .? . / /. no mejoraré tal vez. Feytüfachi operalayaymi, operaeliyu l'ayaymi" pieneu. "Mongepeürkelayan may" pin. Fey l'awen'tueneu, l'awen'tueneu; fey ka tripantun.("Esta vez no te vas a operar, si te opero te vas a morir" me dijo. "No mejoraré tal vez, entonces", dije. Entonces me medicinó, me medicinó; entonces también estuve un año).
nmlch-nmpll1_x_0042_nmpll_00. Ec/Rh/Fc. Ec/ Rh23-12-02.

2.3.2. *Spelling Checker*

The Mapudungun spelling checker is prototype software that detects spelling errors in Mapudungun text within OpenOffice, a freely available graphical text editor (http://www.openoffice.org/). With the Mapudungun spelling checker installed, OpenOffice automatically and interactively underlines misspelled words in red squiggles. Right clicking on a word that has been underlined brings up a menu that lists correctly spelled words that are the closest matches to the misspelled word. If the spelling checker mistakenly underlines a correctly spelled word, the right-click menu also allows adding the word to the dictionary.

The spelling checker is written for MySpell, the spelling checker file format that OpenOffice uses. Two files comprise the MySpell Mapudungun spelling checker. The first file contains two lists: a list of Mapudungun stems, and a list of Mapudungun words. The second file is a list of Mapudungun suffixes. While Mapudungun words frequently contain more than one suffix, MySpell is limited to accepting only a single suffix string per word. For this reason each entry in the suffix list may actually consist of several suffixes. To spell check a Mapudungun text, the spelling checker compares each word in the text to the list of Mapudungun words. If an exact match is found then the word is correctly spelled. If no exact match is found then the spelling checker tries to match the word using any stem in the stem list and any suffix in the suffix list. If no match can be found then the spelling checker believes the word is incorrectly spelled.

The IEI-UFRO team manually checked the spelling of 117,003 full form words that were extracted from corpus. They segmented 15,120 of these. Based on this segmentation, the Mapudungun Spelling Checker contains a list of 5,234 stems which can each combine with 1,303 suffix groups. Additionally, there are 53,094 unsegmented full form words. The single most helpful way to improve the spelling checker would be to increase the number of segmented words used to generate the stem and suffix group lists. Increasing the number of unsegmented words would also help. Additionally, the spelling checker could be extended to understand suffix sequences, since Mapudungun

words frequently contain more than one suffix. Another enhancement would be to inform the spelling checker of the part of speech of the stems, i.e. which stems are nouns, which are verbs, etc. For more details, see Monson et al. 2004.

2.3.3. Machine Translation Systems

2.3.3.1. *Example-Based Machine Translation system*

Example-Based Machine Translation (EBMT) relies on previous translations performed by humans to create new translations without the need for human translators. The previous translations are called the training corpus. For the best translation quality, the training corpus should be as large as possible, and as similar to the text to be translated as possible. When the exact sentence to be translated occurs in the training material, the translation quality is human-level, because the previous translation is re-used. As the sentence to be translated differs more and more from the training material, quality decreases because smaller and smaller fragments must be combined to produce the translation, increasing the chances of an incorrect translation. As the amount of training material decreases, so does the translation quality; in this case, there are fewer long matches between the training texts and the input to be translated. Conversely, more training data can be added at any time, improving the system's performance by allowing more and longer matches.

EBMT usually finds only partial matches, which generate lower-quality translations. When only part of a sentence can be matched against the training corpus, the unmatched words are translated one by one using the most probable target language word from the training corpus. Because EBMT uses probabilities of matches, it can usually find some candidates for translation that are somewhat probable. Thus EBMT is a high coverage approach; most of the text will be translated.

EBMT is not, however, always a high quality approach. While the translation quality can be human-level, any mistakes in the human translations used for training ~ spelling errors, omissions, mistranslations ~ will become visible in the EBMT system's output. Thus it is important that the training data be as accurate as possible. The training corpus we are currently using for EBMT is the spoken language corpus described earlier. This corpus still contains some errors and awkward translations.

Where there are legitimate variants of spelling or word choice in the source language, all of them can be added to increase translation coverage. However, among variant choices in the target language, a single standard translation should be chosen whenever possible to avoid producing conflicting translation candidates among which the EBMT system must choose (possibly incorrectly).

Highly agglutinative languages post a challenge for Example Based MT. Because there are so many inflected versions of each stem, most inflected words are rare. If the rare words do not occur in the corpus at all, they will not be translatable by EBMT. If they occur only a few times, it will also be hard for EBMT to have accurate statistics about how they are used. We are currently working to address this issue by splitting Mapudungun words into stems and affixes. Each individual stem and suffix is not as rare as the combinations of stems and suffixes. For this segmentation, we are currently using

the lists of words segmented into stems and suffix groupings that are used for the spelling checker.

We currently have an EBMT prototype which needs improvement. The improvements will come from the use of morphological analysis, the inclusion of common phrases in the corpus, and fixing translation errors and awkward translations in the corpus.

2.3.3.2. *Rule-Based MT system*

Simultaneously to the development of EBMT, we are working on a prototype rule-based machine translation system for Mapudungun. Rule-based machine translation, which requires a detailed comparative analysis of the grammar of source and target languages, can produce high quality translation but takes a longer amount of time in order to be implemented. It also has lower coverage than EBMT because there is no probabilistic mechanism for filling in the parts of sentences that are not covered by rules. Up to now, the rule system that has been developed for Mapudungun covers the basic grammatical constructions (simple sentences with intransitive and transitive verbs, nominal phrases with determiners and modifiers, verbal phrases with different temporal and aspectual values, passive voice, inverse marking etc.).

The rule-based machine translation system is composed of a series of programs and databases. The input to the system is a Mapudungun sentence, phrase or word, which is processed in different stages until turned into a Spanish output. The MT system consists of three programs: the Mapudungun morphological analyzer, the transfer system, and the Spanish morphological analyzer. Each of these programs makes use of different data bases (lexicons or grammars). The Mapudungun morphological analyzer makes use of two separate Mapudungun lexicons, one containing a list of stems specified for part of speech, and a second one containing a list of suffixes, each one specified for grammatical features. The input to the morphological analyzer is a Mapudungun expression and its output is a morphologically segmented expression plus a specification of the grammatical features of each morpheme, which constitutes the input for the transfer system. The transfer system makes use of a transfer grammar and a transfer lexicon, which contain syntactic and lexical rules in order to map Mapudungun expressions into Spanish expressions. The output of the transfer system is a Spanish expression composed of uninflected words plus grammatical features, which constitutes the input for the Spanish morphological generator. The morphological generator makes use of a Spanish lexicon of inflected words (developed by the Universitat Politècnica de Catalunya). Each of these programs and databases, as well as its interactions, will be described in more detail in the following sections of this paper.

2.3.3.2.1. *Mapudungun morphological analyzer*

While Spanish is an analytic language, Mapudungun is an agglutinative and polysynthetic language with noun and verb incorporation. Even though the morphology of other parts of speech is relatively simple, Mapudungun has a complex agglutinative suffixal verb morphology—some analyses provide as many as 36 verb suffix slots

(Smeets, 1989). A typical complex verb form occurring in our corpus of spoken Mapudungun consists of five or six morphemes.

A verb begins with a stem and ends with an obligatory morpheme-sequence marking, in the case of finite clauses, the person and number of the subject together with the mood of the verb or, in the case of non-finite clauses, adverbialization or nominalization. A number of morphemes may occur between the verb stem and the verb-final morpheme cluster, including aspect, tense, applicative, voice, directional, and object agreement markers. If incorporation occurs, the incorporated noun or verb is placed immediately following the verb stem. The relative order of the verbal morphemes is usually fixed, and there are only a few simple morphophonemic changes at morpheme boundaries. Figure 5 contains glosses of a few morphologically complex Mapudungun verbs taken from our bilingual lexicon.

From this, it follows that an MT system cannot translate Mapudungun words directly into Spanish words. There is the need, therefore, to identify each morpheme with meaning in a Mapudungun sentence, so that the system can then properly translate it into the corresponding Spanish word or phrase. As for EBMT, a morphological analyzer is needed, but in this case the analyzer is more sophisticated because it needs to provide syntactic and semantic features for each morpheme.

Figure 5: *Examples of Mapudungun verbal morphology taken from the AVENUE corpus of spoken Mapudungun*

Amu	-ke	-yngün
go	-habitual	-3plIndic
They (usually) go		
ngütrümtu	-a	-lu
call	-fut	-adverb
While calling (tomorrow), ...		
nentu	-ñma	-nge -ymi
extract	-mal	-pass -2sgIndic
you were extracted (on me)		
ngütramka	-me	-a -fi -ñ
tell	-loc	-fut -3obj -1sgIndic
I will tell her (away)		

The morphological analyzer takes a Mapudungun word as an input and as output it produces all possible segmentations of the word. Each segmentation identifies:

a. a single stem in that word
b. each suffix in that word
c. a semantic analysis for the stem and each identified suffix.

A lexicon of stems works together with a fairly complete lexicon of Mapudungun suffixes. The first version of the stem lexicon contains 1,670 Mapudungun stems. Each entry in this lexicon lists the part of speech of the stem. The suffix lexicon is fairly complete. There are 105 Mapudungun suffixes in the suffix lexicon. Each suffix lists the part of speech that the suffix attaches to: verb, noun, adjective, etc. Each suffix also lists the linguistic features, such as person, number, or mood, that it marks. The software's algorithm does a recursive and exhaustive search on all possible segmentations of a given Mapudungun word. The software starts from the beginning of the word and identifies each stem that is an initial string in that word. Next, the candidate stem from the word is removed. The software then examines the remaining string looking for a valid combination of suffixes that could complete the word. The software iteratively and exhaustively searches for sequences of suffixes that complete the word. For example, after it identifies a first suffix that matches the beginning of the string after the stem, the software resumes the search for the second suffix, and so on, until it exhausts all possibilities. The morphological analyzer also takes into account the allowable ordering of Mapudungun suffixes.

Once the analyzer has found all possible and correct segmentations of a word, it creates a semantic analysis of the complex of suffixes encountered in the analyzed word. For an example, see Figure 6.

Figure 6. *Example showing the output of the morphological analyzer for Mapudungun.*

pekelan	pe-ke-la-n	lexeme = pe (see) Sujeto Persona = 1 Sujeto Número = singular Modo = indicativo Negación = + Aspecto = habitual

2.3.3.2.2. *Run-time Transfer System*

At run time, the translation module translates a source language sentence into a target language sentence. The output of the run-time system is a lattice of translation alternatives. The alternatives arise from syntactic ambiguity, lexical ambiguity, multiple synonymous choices for lexical items in the dictionary, and multiple competing hypotheses from the transfer rules (see next section).

The run-time translation system incorporates the three main processes involved in transfer-based MT: parsing of the source language input, transfer of the parsed constituents of the source language to their corresponding structured constituents on the target language side, and generation of the target language output. All three of these processes are performed based on the transfer grammar – the comprehensive set of transfer rules that are loaded into the run-time system. In the first stage, parsing is performed based solely on the SL side, also called x-side, of the transfer rules. The

implemented parsing algorithm is for the most part a standard bottom-up Chart Parser, such as described in Allen (1995). A chart is populated with all constituent structures that were created in the course of parsing the SL input with the source-side portion of the transfer grammar. Transfer and generation are performed in an integrated second stage. A dual TL chart is constructed by applying transfer and generation operations on each and every constituent entry in the SL parse chart. The transfer rules associated with each entry in the SL chart are used in order to determine the corresponding constituent structure on the TL side. At the word level, lexical transfer rules are accessed in order to seed the individual lexical choices for the TL word-level entries in the TL chart. Finally, the set of generated TL output strings that corresponds to the collection of all TL chart entries is collected into a TL lattice, which is then passed on for decoding (choosing the correct path through the lattice of translation possibilities.) A more detailed description of the runtime transfer-based translation sub-system can be found in Peterson (2002).

2.3.3.2.3. *Transfer Rules*

The function of the transfer rules is to decompose the grammatical information contained in a Mapudungun expression into a set of grammatical properties, such as number, person, tense, subject, object, lexical meaning, etc. Then, the rule builds an equivalent Spanish expression, copying, modifying, or rearranging grammatical values according to the requirements of Spanish grammar and lexicon.

In the AVENUE system, translation rules have six components[2]: a. rule identifier, which consists of a constituent type (Sentence, Nominal Phrase, Verbal Phrase, etc.) and a number; b. constituent structure for both the source language (SL), in this case Mapudungun, 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. transfer equations, which provide information about which feature values transfer from the source into the target language.

In Mapudungun, plurality in nouns is marked, in some cases, by the pronominal particle *pu*. The NBar rule below (Figure 7) illustrates a simple example of a Mapudungun to Spanish transfer rule for plural Mapudungun nouns (following traditional use, in this Transfer Grammar, NBar is the constituent that dominates the noun and its modifiers, but not its determiners).

According to this rule, the Mapudungun sequence PART N will turn into a noun in Spanish. That is why there is only one alignment. The x-side constraint is checked in order to ensure the application of the rule in the right context. In this case, the constraint is that the particle should be specified for (number = pl); if the noun is preceded by any other particle, the rule will not apply. The number feature is passed up from the particle to the Mapudungun NBar, then transferred to the Spanish NBar and passed down to the Spanish noun. The gender feature, present only in Spanish, is passed up from the Spanish noun to the Spanish NBar. This process is represented graphically by the tree structure showed in Figure 8.

---

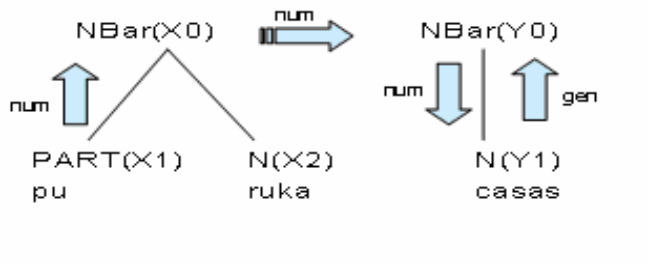
[2] This is a simplified description, for a full description see Peterson (2002) and Probst et al. (2003).

Figure 7. *Plural noun marked by particle pu. Example: pu ruka::casas ('houses')*

{NBar,1}	**(identifier)**
Nbar::Nbar: [PART N] -> [N]	**(x-side/y-side constituent structures)**
((X2::Y1)	**(alignment)**
((X1 number) =c pl)	**(x-side constraint)**
((X0 number) = (X1 number))	**(passing feature up)**
((Y0 number) = (X0 number))	**(transfer equation)**
((Y1 number) = (Y0 number))	**(passing feature down)**
((Y0 gender) = (Y1 gender)))	**(passing feature up)**

Some of the problems that the Transfer Grammar has to solve, among others, are the agglutination of Mapudungun suffixes, that have been previously segmented by the morphological analyzer; the fact that tense is mostly unmarked in Mapudungun, but has to be specified in Spanish; and the existence of a series of grammatical structures that have a morphological nature in Mapudungun (by means of inflection or derivation) and a syntactic nature in Spanish (by means of auxiliaries or other free morphemes).

Figure 8. *Rule for plural NP's with particle pu.*



2.3.3.2.3.1. *Suffix Agglutination*

The transfer grammar manages suffix agglutination by constructing constituents called Verbal Suffix Groups (VSuffG). These rules can operate recursively. The first VSuffG rule turns a Verbal Suffix (VSuff) into a VSuffG, copying the set of features of the suffix into the new constituent. Notice that at this level there are no transfer of features to the target language and no alignment. See Figure 9.

The second VSuffG rule combines a VSuffG with another VSuff, passing up the feature structure of both suffixes to the parent node. For instance, in a word like *pe-fi-ñ* (*pe-*: to see; *-fi:* 3rd. person object; *-ñ*: 1rst. person singular, indicative mood; 'I saw he/she/them/it'), the rule {VSuffG,1} is applied to *-fi*, and the rule {VSuffG,2} is applied

to the sequence *-fi-ñ*. The result is a Verb Suffix Group that has all the grammatical features of its components. This process could continue recursively if there are more suffixes to add.

Figure 9. Verbal Suffix Group Rules.

```
{VSuffG,1}
VSuffG::VSuffG : [VSuff] -> [""]
((X0 = X1))

{VSuffG,2}
VSuffG::VSuffG : [VSuffG VSuff] -> [""]
((X0 = X1)
(X0 = X2))
```

2.3.3.2.3.2. *Tense*

Tense in Mapudungun is mostly morphologically unmarked. The temporal interpretation of a verb is determined compositionally by the lexical meaning of the verb (the relevant feature is if the verb is stative or not) and the grammatical features of the suffix complex. Figure 10 lists the basic rules for tense in Mapudungun.

Since tense should be determined taking into account information from both the verb and the VSuffG, it is managed by the rules that combine these constituents (called VBar rules in this grammar). For instance, Figure 11 displays a simplified version of the rule that assigns the past tense feature when necessary (transfer of features from Mapudungun to Spanish are not represented in the rule for space reasons).

Figure 10. *Tense in Mapudungun.*

Lexical/grammatical features	Temporal interpretation
a. Unmarked tense + unmarked lexical aspect + unmarked grammatical aspect	past (kellu-n::ayudé::(I)helped)
b. Unmarked tense + stative lexical aspect	present (niye-n::poseo::(I)own)
c. Unmarked tense + unmarked lexical aspect + habitual grammatical aspect	present (kellu-ke-n::ayudo::(I)help)
d. Marked tense (for instance, future)	future (pe-a-n::veré::(I)will see)

Figure 11. *Past tense rule (transfer of features omitted)*

```
{VBar,1}
VBar::VBar :  [V VSuffG] -> [V]
((X1::Y1)                                   (alignment)
((X2 tense) = *UNDEFINED*)                  (x-side constraint on morphological
((X1 lexicalaspect) =                       tense)
*UNDEFINED*)                                (x-side constraint on verb's aspectual
((X2 aspect) = (*NOT* habitual))            class)
((X0 tense) = past) …)                      (x-side constraint on grammatical
                                            aspect)
```

(tense feature assignment)

Analogous rules deal with the other temporal specifications.

2.3.3.2.3.3. *Typological divergence*

As an agglutinative language, Mapudungun has many grammatical constructions that are expressed by morphological, rather than syntactic, means. For instance, passive voice in Mapudungun is marked by the suffix *-nge*. On the other hand, passive voice in Spanish, as well as in English, requires an auxiliary verb, which carries tense and agreement features, and a passive participle.

Figure 12. Passive voice rule (transfer of features omitted).

```
{VBar,6}
VBar::VBar :  [V VSuffG] -> [V V]        (insertion of aux in Spanish side)
((X1::Y2)                                (Mapudungun verb aligned to Spanish verb)
((X2 voice) =c passive)                  (x-side voice constraint )
((Y1 person) = (Y0 person))              (passing person features to aux)
((Y1 number) = (Y0 number))              (passing number features to aux)
((Y1 mood) = (Y0 mood))                  (passing mood features to aux)
((Y2 number) =c (Y1 number))             (y-side agreement constraint)
((Y1 tense) = past)                      (assigning tense feature to aux)
((Y1 form) =c ser)                       (auxiliary selection)
((Y2 mood) = part)                       (y-side verb form constraint)
  …)
```

For instance, *pe-nge-n* (*pe-*: to see; *-nge:* passive voice; *-n*: 1rst. person singular, indicative mood; 'I was seen') has to be translated as *fui visto* o *fue vista.* The rule for passive (a VBar level rule in this grammar) has to insert the auxiliary, assign it the right grammatical features and inflect the verb as a passive participle. Figure 12 shows a simplified version of the rule that produces this result (transfer of features from Mapudungun to Spanish are not represented in the rule for space reasons).

2.3.3.2.4. *Spanish Morphology generation*

Even though Spanish is not as highly inflected as Mapudungun or Quechua, there is still a great deal to be gained from listing just 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
…

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 for the fourth entry 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 this 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.

3. Quechua Cooperation

In the case of Quechua, there are two projects that allowed the cooperation between a team of computational linguists and some members of the Quechua community: AVENUE and TechBridgeWorld. TechBridgeWorld is a fairly new initiative started at Carnegie Mellon University and it embraces several programs. The one of interest here is called the V-Unit (for Vision Unit), which allows graduate students at Carnegie Mellon University to self-define and implement a project related to non-traditional uses of technology during a Semester as a regular course.

We have been coordinating the Quechua data collection with some partners in Cusco (Peru) for over a year, with the ultimate goal of building a Quechua-Spanish MT system. One of the authors (Ariadna Font Llitjós) spent last summer in Cusco (from the beginning of June until the end of August 2005) to set up the infrastructure required to develop all the necessary NLP tools and databases as well as to implement a first prototype for the Quechua-Spanish MT system.

The main purpose of the trip was getting the basic resources (such as a lexicon and morphology) together with members of the Quechua community, as well as developing a test suite to serve as training and test set data for MT system development. Translation and morphology lexicons were automatically created from the data annotated by a native speaker using several scripts. Grammar writing also started during that period.

A preliminary user study of the correction of Quechua to Spanish translations was conducted towards the end of the trip. For this user study, three Quechua speakers with

good knowledge of Spanish evaluated and corrected machine translations, when necessary, through a user-friendly interface called Translation Correction Tool, designed by one of the authors (Font Llitjós & Carbonell, 2004).

3.1. Obtaining Parallel Written Corpus

3.1.1. *Elicitation Corpus*

Part of the data collected in Cusco was a translation of the AVENUE Elicitation Corpus (EC). The EC is used when there is no natural corpus large enough to use for development of MT. The EC is like a fieldwork questionnaire containing simple sentences that elicit specific meanings and structures. The EC has two parts. The first part, the Functional Elicitation Corpus, runs through functional/communicative features such as number, person, tense, and gender. The version that was used in Peru had 1,700 sentences. The second part, the Structural Elicitation Corpus, is a smaller corpus designed to cover the major structures present in the Penn Treebank (Marcus *et al*., 1992). Out of 122,176 sentences from the Brown Corpus 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. Some examples of elicitation sentences and phrases can be seen in Figure 13. For more information about how this corpus was created and what its properties are, see Probst and Lavie (2004).

Figure 13: *Some elicitation sentences from the structural corpus*

```
SL: to the election
C-Structure:(<PP> (PREP to-1) (<NP> (DET the-2) (N election-3)))
CompSeq: PP-> PREP NP

SL: the chair in the corner
C-Structure:(<NP> (DET the-1) (N chair-2) (<PP> (PREP in-3)
  (<NP> (DET the-4) (N corner-5))))
CompSeq: NP-> DET N PP

SL: attorneys for the mayor
C-Structure:(<NP> (N attorneys-1) (<PP> (PREP for-2) (<NP>
  (DET the-3) (N mayor-4))))
CompSeq: NP-> N PP

SL: I can not run
C-Structure:(<S> (<NP> (PRO I-1)) (<AUX> (AUX can-2)) (<NEG>
  (ADV not-3)) (<VP> (V run-4)))
CompSeq: S-> NP AUX NEG VP
```

We had a native Quechua speaker (Irene Gómez) and a linguist with good knowledge of Quechua (Marilyn Feke) translate both the Functional Elicitation Corpus and the Structural Elicitation Corpus. We also had non-native speaker of Quechua (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. The final Structural Elicitation Corpus which was translated into Quechua had 146 Spanish sentences.

3.1.2. *Scanned Text*

Besides the Elicitation Corpora, we did not have access to any other Quechua text on electronic format, so we looked for written text and we found three books which had parallel text in Spanish and Quechua: *Cuento Cusqueños*, *Cuentos de Urubamba*, *Gregorio Condori Mamani*. We scanned these books and had Quechua speakers (both in Pittsburgh and in Cusco) go over the Quechua text (360 pages total), so as to correct the optical character recognition (OCR) errors. 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 Cusco). Neither of them are native speakers 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.

3.2. *Segmentation and Translation of Word Types*

In order to build a translation and morphology lexicon, we need to have as many examples as possible of segmented words translated into Spanish. When counting words, we distinguish between types and tokens. The number of types does not count repetitions of words. The number of tokens counts each instance of each word.

For this project, we extracted all the types of words from the three Quechua books, and ordered them by frequency. The total number of types are 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.[3] Since 16,722 word types were only seen once in the books, we decided to segment and translate only the 10,000 most frequent words in the list, hoping to reduce the number of 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.

During this summer, Ariadna Font Llitjós and Irene Gómez segmented and translated 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: Word Segmentation, Root translation, Root POS, Word Translation, Word POS and Translation of the final root if there has been a POS change.

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”

---

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

("machuyay")[4]. 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)

3.3. *A Rule-Based MT Prototype*

Similarly to the Mapudungun-Spanish system, the Quechua-Spanish system also has a Quechua morphological analyzer which pre-processes the input sentences to split words into roots and suffixes. The lexicon and the rules are applied by the transfer engine, and finally, the Spanish morphology generation module is called to inflect the corresponding Spanish stems with the relevant features.

3.3.1. *Stem and Suffix Lexicons*

Form the list of segmented and translated words, we automatically generated and manually corrected two lexicons containing mostly stems from the 100 most frequent words and from the two different types of the Elicitation Corpora. For example, from the word type "chayqa" and the specifications given for all the other fields as shown in Figure 14, 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).

Figure 14. *Example of segmented and translated word type.*

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 aren't any spurious lexical entries.

Some examples of automatically generated lexical entries see Figure 15.

Figure 15. *Automatically generated lexical entries from segmented and translated word list*

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

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

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

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

Adv |: [kunan] -> [ahora]

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

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

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

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

((X1::Y1))

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

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

Most of the suffix lexical entries, however, are hand-crafted, since they are only about 150, as listed in Cusihuaman's grammar (2001). See Figure 16.

For the current working MT prototype, the Suffix Lexicon has 36 entries.

Figure 16. *Manually written suffix lexical entries.*

```
; "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))
```

3.3.2. *Translation Rules*

The translation grammar, written with comprehensive rules following the same formalism described in subsection 2.3.3.2.3 above, currently 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. Figure 17 shows a couple of examples of rules in the translation grammar.

Figure 17. *Manually written grammar rules for Quechua-Spanish translation..*

```
{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)))
```

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:

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

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

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

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

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

sl: taki ni taq
tl: Y CANTO

tree: <((SBAR,2 (LITERAL "Y") (S,1 (VP,0 (VBAR,1 (VBAR,2 (V,2:1 "CANTO")))))))>

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")))))))>

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

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")))))))>

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

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")))))))>

We are also planning to expand the translation grammar and lexicon to be able to cover simple dialogs.

3.4. *User Studies*

A preliminary user study of the correction of Quechua to Spanish translations was conducted towards the end of the trip. 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), developed by Ariadna Font Llitjós as part of her Ph.D. research (Font Llitjós & Carbonell, 2004).

It was very important for our research to see how Quechua speakers used the TCTool and whether they had any problems with the interface. The user study already showed that the Quechua representation of stem and suffixes as separate words does not seem to pose a problem and that it was relatively easy to use for non-technical users.

However, we still need to analyze the log files from the user study in detail to see what sorts of errors they corrected and how they corrected them.

4. Conclusions and Future Work

The cooperation with Mapudungun and Quechua speakers has been fruitful. The AVENUE partners in Chile have just released their Mapudungun-Spanish dictionary

online (http://www.estudiosindigenas.cl/), and the AVENUE team in Pittsburgh is currently working on putting the different MT systems for Mapudungun-Spanish online as well. To see the AVENUE MT website, which is still in an experimental phase, go to http://www.lenguasamerindias.org/.

For the official release of the AVENUE MT website, the EBMT team has worked on cleaning the data to improve alignment accuracy. (One problem for the initial system was posed by untranslated sentences in the speech corpus.) We are also working on adding our morphological analyzer to the MT web site.

For the next version of the MT website, we plan to plug in the Translation Correction Tool to allow bilingual users interested in translating sentences to give us feedback about the correctness of the automatic translation produced by our systems in a simple and user-friendly way.

5. Bibliography

Allen, James. (1995). *Natural Language Understanding*. Second Edition ed. Benjamin Cummings.

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.

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

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

Frederking, Robert and Nirenburg, Sergei. (1994). Three Heads are Better than One. Proceedings of the fourth Conference on Applied Natural Language Processing (ANLP-94), pp. 95-100, Stuttgart, Germany.

Mitchell, Marcus, Taylor A., MacIntyre, R., Bies, A., Cooper, C., Ferguson, M., Littmann, A. (1992). The Penn Treebank Project. http://www.cis.upenn.edu/ treebank/home.html.

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).

Lavie, Alon and Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna

Font Llitjós, Rachel Reynolds, Jaime Carbonell, and Richard Cohen. (2003). Experiments with a Hindi-to-English Transfer-based MT System under a Miserly Data Scenario". ACM Transactions on Asian Language Information Processing (TALIP), 2(2).

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.

Levin, Lori; Vega, Rodolfo; Carbonell, Jaime; Brown, Ralf; Lavie, Alon; Cañulef, Eliseo and Huenchullan, Carolina. (2000). Data Collection and Language Technologies for Mapudungun. International Conference on Language Resources and Evaluation (LREC).

Peterson, Erik. (2002). Adapting a transfer engine for rapid machine translation *development.* M.S. thesis, Georgetown University.

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; 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

Smeets, I. (1989). A Mapuche Grammar. Ph.D. Dissertation. University of Leiden.

6. Contact Information

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave. NSH 4611
Pittsburgh PA, 15213
USA
(http://www.cs.cmu.edu/~aria/)