Building NLP Systems for Two Resource-Scarce Indigenous Languages: Mapudungun and Quechua

Christian Monson[1], Ariadna Font Llitjós[1], Roberto Aranovich[2], Lori Levin[1], Ralf Brown[1], Eric Peterson[1], Jaime Carbonell[1], Alon Lavie[1]

1 Language Technologies Institute
School of Computer Science
Carnegie Mellon University

2 Department of Linguistics
University of Pittsburgh

Abstract

By adopting a "first-things-first" approach we overcome a number of challenges inherent in developing NLP Systems for resource-scarce languages. By first gathering the necessary corpora and lexicons we are then enabled to build, for Mapudungun, a spelling-corrector, morphological analyzer, and two Mapudungun-Spanish machine translation systems; and for Quechua, a morphological analyzer as well as a rule-based Quechua-Spanish machine translation system. We have also worked on languages (Hebrew, Hindi, Dutch, and Chinese) for which resources such as lexicons and morphological analyzers were available. The flexibility of the AVENUE project architecture allows us to take a different approach as needed in each case. This paper describes the Mapudungun and Quechua systems.

1. The AVENUE Project

The long-term goal of the AVENUE project at CMU is to facilitate machine translation for a larger percentage of the world's languages by reducing the cost and time of producing MT systems. There are a number of radically different ways to approach MT. Each of these methods of accomplishing machine translation has a different set of strengths and weaknesses and each requires different resources to build. The AVENUE approach combines these different types of MT in one "omnivorous" system that will "eat" whatever resources are available to produce the highest quality MT possible given the resources. If a parallel corpus is available in electronic form, we can use example-based machine translation (EBMT) (Brown et al., 2003; Brown, 2000), 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 newly developed machine learning technique (Probst, 2005) to learn translation rules from data that is elicited from native speakers. As detailed in the remainder of this paper, the particular resources that the AVENUE project produced facilitated developing an EBMT and a human-coded rule-based MT system for Mapudungun, and a hand-built rule-based MT system for Quechua. Automatic rule learning has been applied experimentally for several other language pairs: Hindi-to-English (Lavie et al. 2003) and Hebrew-to-English (Lavie et al. 2004).

The AVENUE project as a whole consists of six main modules (Figure 1), which are used in different combinations for different languages: 1) elicitation of a word aligned parallel corpus (Levin et al. in press); 2) automatic learning of translation rules (Probst, 2005) and morphological rules (Monson et al. 2004); 3) the run time MT system for performing source to target language translation based on transfer rules; 4) the EBMT system (Brown, 1997); 5) a statistical "decoder" for selecting the most likely translation from the available alternatives; and 6) a module that allows a user to interactively correct translations and automatically refines the translation rules (Font Llitjós et al. 2005a).

2. AVENUE and Indigenous Languages of the Western Hemisphere

Over the past six years the AVENUE project at the Language Technologies Institute at Carnegie Mellon University has worked with native informants and the government of Chile to produce a variety of natural language processing (NLP) tools for Mapudungun, an indigenous South American language spoken by less than 1 million people in Chile and Argentina. During the final year and a half of this time, the AVENUE team has also been developing tools for Quechua, spoken by approximately 10 million people in Peru, Bolivia, Ecuador, South of Colombia, and northern Argentina.

Electronic resources for both Quechua and Mapudungun are scarce. At the time the AVENUE team started working on Mapudungun, even simple natural language tools such as morphological analyzers or spelling correctors did not exist. In fact, there were few electronic resources from which such natural language tools might be built. There were no standard Mapudungun text or speech corpora, or lexicons, and no parsed treebanks. The text that does exist is in a variety of competing orthographic formats. More resources exist for Quechua but they are still far from what is needed to build a complete MT system.

In addition to these practical challenges facing construction of natural language systems for Mapudungun and Quechua, there are also theoretical and human factor challenges. Both Mapudungun and Quechua pose unique challenges from a linguistic theory perspective, since they have complex agglutinative morphological structures. In addition Mapudungun is polysynthetic, incorporating objects into the verb of a sentence. Agglutination and polysynthesis are both properties that the majority languages, for which most natural language resources have been built, do not possess. Human factors also pose a particular challenge for these two languages. Namely, there is a scarcity of people trained in computational linguistics who

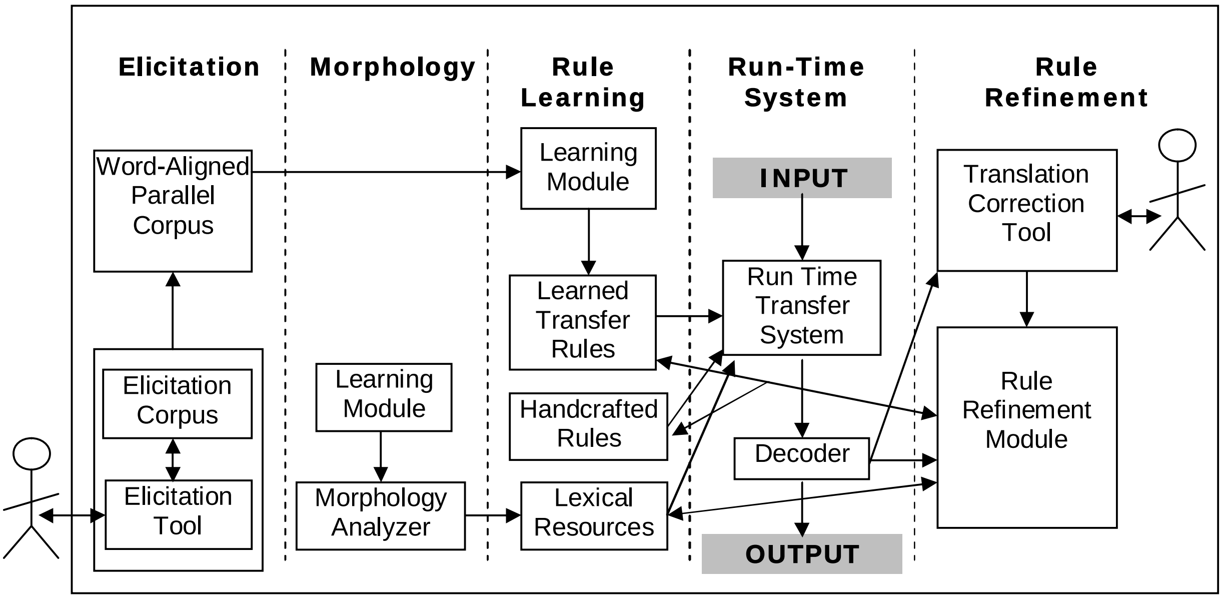

Figure 1: *Data Flow Diagram for the* AVENUE *Rule-Based MT System*

are native speakers or have a good knowledge of these indigenous languages. And finally, an often over looked challenge that confronts development of NLP tools for resource-scarce languages is the divergence of the culture of the native speaker population from the western culture of computational linguistics.

Despite the challenges facing the development of natural language processing systems for Mapudungun and Quechua, the AVENUE project has developed a suite of basic language resources for each of these languages, and has then leveraged these resources into more sophisticated natural language processing tools. The AVENUE project led a collaborative group of Mapudungun speakers in first collecting and then transcribing and translating into Spanish by far the largest spoken corpus of Mapudungun available. From this corpus we then built a spelling checker and a morphological analyzer for Mapudungun. For Quechua, we have created parallel and aligned data as well as a basic bilingual lexicon with morphological information. And with these resources we have built two prototype machine translation (MT) systems for Mapudungun and one prototype MT system for Quechua. This paper will detail the construction of these resources focusing on overcoming the specific challenges Mapudungun and Quechua each present as resource-scarce languages.

3. Mapudungun

Since May of 2000, in an effort to ultimately produce a machine translation system for Mapudungun and Spanish, computational linguists at CMU's Language Technologies Institute have collaborated with Mapudungun language experts at the Instituto de Estudios Indigenas (IEI - Institute for Indigenous Studies) at the Universidad de la Frontera (UFRO) in Chile and with the Bilingual and Multicultural Education Program of the Ministry of Education (Mineduc) in Chile (Levin et al., 2000).

From the very outset of our collaboration we battled the scarcity of electronic resources for Mapudungun. The first phase of the AVENUE collaboration was to collect and produce parallel Mapudungun-Spanish language data from which higher-level language processing tools and systems could be built.

One barrier we faced in the collection of Mapudungun language data is that there are currently several competing orthographic conventions for written Mapudungun. Early in the AVENUE Mapudungun collaboration the IEI-UFRO team established a set of orthographic conventions. All the data collected under the AVENUE project conforms to this set of orthographic conventions. If the Mapudungun data was originally in some other orthographic convention then we manually converted it into our own orthography. Recently, however, a different orthography, Azümchefi, has been chosen by the Chilean government for official documents. Portions of our data have been automatically converted into Azümchefi using automatic substitution rules.

3.1. Corpora and Lexicons

Recognizing the scarcity of Mapudungun language data, the AVENUE team began collecting Mapudungun data soon after our collaboration began in 2000. Directed from CMU and conducted by native speakers of Mapudungun at the Universidad de la Frontera in Temuco, Chile, our data collection efforts ultimately resulted in three separate corpora: 1) a small parallel Mapudungun-Spanish corpus of historical texts and newspaper text, 2) 1700 sentences in Spanish that were manually translated and aligned into Mapudungun (Elicitation Corpus) and 3) a relatively large parallel corpus consisting of 170 hours of transcribed and translated Mapudungun speech. These corpora are described by Monson et al. (2004).

3.1.1. Frequency Based Lexicons

As a first step toward higher level NLP resources for Mapudungun the AVENUE team converted the transcribed spoken corpus text into a lexicon for Mapudungun. All the unique words in the spoken corpus were extracted and then ordered by frequency. The first 117,003 most frequent of these fully-inflected word forms were hand-checked for spelling according to the adopted orthographic conventions for 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)

Figure 2: *Examples of Mapudungun verbal morphology taken from the AVENUE corpus of spoken Mapudungun*

Because Mapudungun has a rich morphology, many NLP applications could benefit from knowing not just fully inflected word forms but also knowing lexical stems. To this end 15,120 of the most frequent fully inflected word forms were hand segmented into two parts: The first part consisting of the stem of the word and the second part consisting of one or more suffixes. This produced 5,234 stems and 1,303 suffix groups. These 5,234 stems were then translated into Spanish.

3.2. Basic Language Tools

With the basic Mapudungun corpora and lexicons in hand, the AVENUE team has developed two basic language tools for Mapudungun: a spelling checker (Monson et al., 2004) for use in a word processing application and a morphological analyzer that produces a syntactic description of individual Mapudungun words.

3.2.1. Mapudungun morphological analyzer

In contrast to the stand-alone spelling checker, the AVENUE team has also developed a morphological analyzer for Mapudungun designed to be integrated into machine translation systems. Mapudungun is an agglutinative and polysynthetic language. A typical complex verb form occurring in our corpus of spoken Mapudungun consists of five or six morphemes. Since each morpheme may alone contain several morpho-syntactic features, it is difficult for an MT system to translate Mapudungun words directly into another language. By identifying each morpheme in each Mapudungun word and assigning a meaning to each identified morpheme, a machine translation system can translate individually each piece of meaning. Figure 2 contains glosses of a few morphologically complex Mapudungun verbs that occur in the spoken corpus.

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

- A single stem in that word
- Each suffix in that word
- A syntactic analysis for the stem and each identified suffix.

To identify the morphemes (stem and suffixes) in a Mapudungun word, a lexicon of stems works together with a fairly complete lexicon of Mapudungun suffixes. The first version of the stem lexicon contains the 1,670 cleanest stems, and their Spanish translations, that were segmented and translated during the lexicon production for Mapudungun. Each entry in this lexicon lists the part of speech of the stem as well as other features associated with the stem such as lexical aspect in the case of verb stems. The suffix lexicon, built by hand by computational linguists on the AVENUE team, is fairly complete. Unlike the suffix groups used in the spelling checker, each suffix entry in the suffix lexicon for the morphological analyzer is an individual suffix. 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 morphological analyzer performs 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. Because the morphological analyzer also takes into account constraints on the allowable ordering of Mapudungun suffixes, most Mapudungun words for which the stem is in the stem lexicon receive a single analysis. A few truly ambiguous suffix combinations may cause a Mapudungun word to receive perhaps as many as five distinct analyses.

Once the morphological analyzer has found all possible and correct segmentations of a word, it combines the feature information from the stem and the suffixes encountered in the analyzed word to create a syntactic analysis that is returned. For an example, see Figure 3.

3.3. Machine Translation Systems

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

		Lexeme = pe (see)
		subject person = 1
pekelan	pe-ke-la-n	**subject number = singular**
		mode = indicative
		negation = +
		aspect = habitual

Figure 3. *Example showing the output of the morphological analyzer for Mapudungun.*

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. Further, EBMT searches for phrases of two or more words, and thus there can be portions of the input which do not produce phrasal translations. For unmatched portions of the input, EBMT falls back on a probabilistic lexicon trained from the corpus to produce word-for-word translations. This fall-back approach provides some translation (even though of lower quality) for any word form encountered in the training data. While the translation quality of an EBMT system can be human-level when long phrases or entire sentence are matched, 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 use for EBMT is the spoken language corpus described earlier. As discussed in section 1.1.1, the corpus of spoken Mapudungun contains some errors and awkward translations.

Highly agglutinative languages pose 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. Additionally, word-level alignment between the two languages, which is used to determine the appropriate translation of a partial match, becomes more difficult when individual words in one language correspond to many words in the other language. We address both issues by using the morphological analyzer for Mapudungun to split words into stems and suffixes. Each individual stem and suffix is more common than the original combination of stem and suffixes, and the individual parts are more likely to map to single words in Spanish.

We currently have a prototype EBMT system trained on approximately 204,000 sentence pairs from the corpus of spoken Mapudungun, containing a total of 1.25 million Mapudungun tokens after splitting the original words into stems and one or more suffixes. This separation increased BLEU scores (Papineni et al., 2001) on a held out portion of the speech corpus by 5.48%, from 0.1530 to 0.1614. We expect further increases from improved use of morphological analysis, the inclusion of common phrases in the corpus, and fixing translation errors and awkward translations in the corpus.

3.3.2. Rule-Based MT system

Simultaneous to the development of the example based machine translation system for Mapudungun we have been working on a prototype rule-based MT system. 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 to implement. Hand-built rule-based MT 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.

The rule-based machine translation system is composed of a series of components and databases. The input to the system is a Mapudungun sentence, phrase or word, which is processed in different stages until a Spanish string is output. The MT system consists of three main components: the Mapudungun morphological analyzer discussed in section 3.2.1, the transfer system, and the Spanish morphological analyzer. Each of these programs makes use of different data bases (lexicons or grammars). 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 their interactions, will be described in more detail in the following sections of this paper.

3.3.2.1. Run-time Transfer System

At run time, the transfer 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 subsystem can be found in Peterson (2002).

{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)**

Figure 4. Plural noun marked by particle pu. Example: pu ruka::casas ('houses')

3.3.2.2. 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, each particular 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[1]: a. rule identifier, which consists of a constituent type (Sentence, Nominal Phrase, Verbal Phrase, etc.) and a unique ID 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 4) 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 be transfered 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 would 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 5.

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).

3.3.2.3. 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 6.

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: 3^{rd}. person object; -ñ: 1st. 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.

3.3.2.4. 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 7 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 8 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 brevity). Analogous rules deal with the other temporal specifications.

3.3.2.5. Typological divergence

As an agglutinative language, Mapudungun has many grammatical constructions that are expressed by morphological, rather than syntactic, means. For instance, passive

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

```
{VSuffG,1}                            {VSuffG,2}
VSuffG::VSuffG : [VSuff] -> [""]      VSuffG::VSuffG : [VSuffG VSuff] -> [""]
((X0 = X1))                           ((X0 = X1)
                                      (X0 = X2))
```

Figure 6. *Verbal Suffix Group Rules.*

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 7. *Tense in Mapudungun.*

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

Figure 8. *Past tense rule (transfer of features omitted)*

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.

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 9 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 brevity).

3.3.2.6. 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 generate Spanish morphology, 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.

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

```
{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)
  …)
```

Figure 9. *Passive voice rule (transfer of features omitted).*

Generator, once the translation has been completed, and have it generate the appropriate surface, inflected forms.

When the Spanish morphological generation is integrated with the run-time transfer system the final rule-based Quechua-Spanish MT system produces output such as the following:

```
sl: kümelen (I'm fine)
tl: ESTOY BIEN
   tree: <((S,5 (VPBAR,2 (VP,1 (VBAR,9 (V,10
          'ESTOY') (V,15 'BIEN') ) ) ) ) )>

sl: ayudangelay (he/she were not helped)
tl: NO FUE AYUDADA
   tree: <((S,5 (VPBAR,2 (VP,2 (NEGP,1 (LITERAL
          'NO') (VBAR,3 (V,11 'FUE')
          (V,8 'AYUDADA') ) ) ) ) ) )>
tl: NO FUE AYUDADO
   tree: <((S,5 (VPBAR,2 (VP,2 (NEGP,1 (LITERAL
          'NO') (VBAR,3 (V,11 'FUE')
          (V,8 'AYUDADO') ) ) ) ) ) )>

sl: Kuan ñi ruka (John's house)
tl: LA CASA DE JUAN
   tree: <((S,12 (NP,7 (DET,10 'LA') (NBAR,2 (N,8
          'CASA') ) (LITERAL 'DE')
          (NP,4 (NBAR,2 (N,1 'JUAN') ) ) ) ) )>
```

4. Quechua

Data collection for Quechua started in 2004, when the AVENUE team established a collaboration with bilingual speakers in Cusco (Peru). In 2005, one of the authors (Ariadna Font Llitjós) spent the summer in Cusco to set up basic infrastructure and to develop a first Quechua-Spanish MT prototype system, with the main goal to have an initial system for testing the Translation Correction Tool (Font Llitjós & Carbonell, 2004) and the Rule Refinement module (Font Llitjós et al., 2005a). Translation and morphology lexicons were automatically created from data annotated by a native speaker using Perl scripts. A small translation grammar was written. Additionally, a preliminary user study of the correction of Quechua to Spanish translations was also conducted using the Translation Correction Tool (TCTool), an online user-friendly interface.

4.1. Text Corpora

As part of the data collected for Quechua, the AVENUE Elicitation Corpora (EC) were translated and manually aligned by a both a native Quechua speaker and a linguist with good knowledge of Quechua. The EC is used when there is no natural corpus large enough to use for development of MT. The EC resembles a fieldwork questionnaire containing simple sentences that elicit specific meanings and structures. It has two parts. The first part, the Functional Elicitation Corpus, contains sentences designed to elicit 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. 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.

Besides the Elicitation Corpora, there was no other Quechua text readily available on electronic format, and thus three books which had parallel text in Spanish and Quechua were scanned: Cuento Cusqueños, Cuentos de Urubamba, and Gregorio Condori Mamani. Quechua speakers examined the scanned Quechua text (360 pages), and corrected the optical character recognition (OCR) errors, with the original image of the text as a reference.

4.2. A Rule-Based MT Prototype

Similar to the Mapudungun-Spanish system, the Quechua-Spanish system also contains 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 (Section 3.3.2.6).

4.2.1. Morphology and Translation Lexicons

In order to build a translation and morphology lexicon, the word types from the three Quechua books were extracted and ordered by frequency. The total number of types was 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.[2] Since 16,722 word types were only seen once in the books (singletons), 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 unique word types from one of the versions of the Elicitation Corpora were also extracted (1,666 types) to ensure basic coverage.

10,000 words were segmented and translated by a native Quechua speaker. The (Excel) file used for this task contained 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 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. 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")[3]. 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).

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

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

```
{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)))
```

Figure 13. *Manually written grammar rules for Quechua-Spanish translation..*

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

Figure 11. *Automatically generated lexical entries from segmented and translated word*

```
; "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))
```

Figure 12. *Manually written suffix lexical entries.*

From the list of segmented and translated words, a stem lexicon was automatically generated and manually corrected. For example, from the word type “chayqa” and the specifications given for all the other fields as shown in Figure 10, 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). 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. The scripts allow for fast post-processing of thousands of words, however manual checking is still required to make sure that no spurious lexical entries have been created. Some examples of automatically generated lexical entries are presented in Figure 11. Suffix lexical entries, however, were hand-crafted, see Figure 12. For the current working MT prototype the Suffix Lexicon has 36 entries. Cusihuaman’s grammar (2001) lists a total of 150 suffixes.

4.2.2. Translation Rules

The translation grammar, written with comprehensive rules following the same formalism described in subsection 3.3.2.2 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 13 shows a couple of examples of rules in the translation grammar.

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 discussed in section 3.3.2.6:

sl: taki sha ra ni (I was singing)

tl: ESTUVE CANTANDO
tree: <((S,1 (VP,0 (VBAR,5 (V,0:0 "ESTUVE") (V,2:1 "CANTANDO")))))>

sl: taki ra n si (it is said that s/he sang)
tl: DICE QUE CANTÓ
tree: <((SBAR,1 (LITERAL "DICE QUE") (S,1 (VP,0 (VBAR,1 (VBAR,4 (V,2:1 "CANTÓ")))))))>

sl: noqa qa barcelona manta ka ni (I am from Barcelona)
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")))))))>

4.3. Preliminary User Studies

A preliminary user study of the correction of Quechua to Spanish translations was conducted where three Quechua speakers with good knowledge of Spanish evaluated and corrected nine machine translations, when necessary, through a user-friendly interface called the Translation Correction Tool (TCTool). This small user study allowed us to see how Quechua speakers used the TCTool and whether they had any problems with the interface. It 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.

5. Conclusions and Future Work

The "first-things-first" approach the AVENUE project has taken to building NLP systems for scarce-resource languages has proven effective. By first focusing effort on producing basic NLP resources, the resultant resources are of sufficient quality to be put to any number of uses: from building all manner of NLP tools to potentially aiding linguists in better understanding an indigenous culture and language. For both Mapudungun and Quechua, separate work on morphology analysis and on a transfer grammar modularized the problem in a way that allowed rapid development. Besides a spelling-checker for Mapudungun, the AVENUE team has developed computational lexicons, morphology analyzers and one or more Machine Translation systems for Mapudungun and Quechua into Spanish.

The AVENUE team has recently put many of the resources we have developed for Mapudungun online at http://www.lenguasamerindias.org/mapudungun. The AVENUE interactive website, which is still in an experimental phase, contains a basic Mapudungun-Spanish lexicon, the Mapudungun morphological analyzer, and the example-based MT system from Mapudungun to Spanish.

The AVENUE team continues to develop the resources for both Mapudungun and Quechua. We are actively working on improving the Mapudungun-Spanish rule-based MT system by both increasing the size of the lexicon as well as improving the rules themselves. The Mapudungun-Spanish example-based MT system can be improved by cleaning and increasing the size of the training text. 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.

6. Acknowledgements

Many people beyond the authors of this paper contributed to the work described. The Chilean AVENUE team includes Carolina Huenchullan Arrúe, Claudio Millacura Salas, Eliseo Cañulef, Rosendo Huisca, Hugo Carrasco, Hector Painequeo, Flor Caniupil, Luis Caniupil Huaiquiñir, Marcela Collio Calfunao, Cristian Carrillan Anton, and Salvador Cañulef. Rodolfo Vega was the liaison between the CMU team and the government agency and indigenous community in Chile. People that participated in the development of Quechua data are Marilyn Feke, Salomé Gutierrez, Irene Gómez and Yenny Ccolque.

This research was funded in part by the DARPA TIDES program and by NSF grant number IIS-0121-631.

7. References

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

Brown, Ralf D., Rebecca Hutchinson, Paul N.Bennett, Jaime G. Carbonell, and Peter Jansen. (2003). "Reducing Boundary Friction Using Translation-Fragment Overlap", in Proceedings of the Ninth Machine Translation Summit, New Orleans, USA. pp. 24-31.

Brown, Ralf D. (2000). "Example-Based Machine Translation at Carnegie Mellon University". In The ELRA Newsletter, European Language Resources Association, vol 5:1, January-March 2000.

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

Font Llitjós, Ariadna, Jaime Carbonell, Alon Lavie. (2005a). 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, Roberto Aranovich, and Lori Levin (2005b). Building Machine translation systems for indigenous languages. Second Conference on the Indigenous Languages of Latin America (CILLA II). Texas, USA.

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

Lavie, A., S. Wintner, Y. Eytani, E. Peterson and K. Probst. (2004) "Rapid Prototyping of a Transfer-based Hebrew-to-English Machine Translation System". In Proceedings of the 10th International Conference on

Theoretical and Methodological Issues in Machine Translation (TMI-2004), Baltimore, MD, October 2004. Pages 1-10.

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

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

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

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022).

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