Data Collection and Language Technologies for Mapudungun

Lori Levin*, Rodolfo Vega*, Jaime Carbonell*, Ralf Brown*, Alon Lavie*
Eliseo Cañulef†, Carolina Huenchullan**

*Language Technologies Institute
Carnegie Mellon University
{lsl, rmvega, alavie, jgc}@cs.cmu.edu

†Instituto de Estudias Indígenas
Universidad de la Frontera
** Ministerio de Educación, Chile

Abstract

Mapudungun is spoken by over 900,000 people (Mapuche) in Chile and Argentina. Thanks to an active bilingual and multicultural education program, Mapuche children are now being taught to be literate in both Mapudungun and Spanish. The Chilean Ministry of Education has teamed up with the Language Technologies Institute's AVENUE project to collect data and produce language technologies that support bilingual education. The main resource that has come out of the Mineduc-LTI partnership is Mapudungun-Spanish parallel corpus consisting of approximately 200,000 words of text and 120 hours of transcribed speech. Plans are being made for machine translation and computer-assisted instruction.

1. Introduction

This paper describes a partnership between the Ministry of Education in Chile and Carnegie Mellon University's Language Technologies Institute in the United States. The goal of the joint project is to produce corpora and language technologies that support bilingual education in Spanish and Mapudungun, the language of over 900,000 Mapuche people in Chile and Argentina. The joint project takes place in the context of two ongoing programs: the Bilingual and Multicultural Education Program in Chile (Section 2.), and the AVENUE project (Section 3.) at Carnegie Mellon University. Section 4. of this paper describes the Mineduc-LTI partnership and Section 5. describes the main product of the Mineduc-LTI partnership — a parallel corpus of Spanish and Mapudungun including historical texts (around 200,000 words) and transcripts of spoken dialogue (120 hours of speech). Section 6. describes plans for the partnership, including eliciting sentences for machine learning of translation rules.

2. Bilingual-Multicultural Education in Chile

Bilingual and multicultural education is an educational practice committed to relevance to the students, contextualization of educational content, and the centrality of the child in the pedagogical practice with the participation of the family and community. Bilingual multicultural education allows indigenous people to exercise their right to learn their own language based in their own culture, and at the same time, learn the national language and culture.

The Chilean government (Law 19.253, 1993) recognizes eight ethnic groups: the Mapuche, Aymara, Rapa Nui or Pascuense, Likay Antai, Quechua, Colla, Kawashkar or Alacalufe, and Yamana or Yagan. In this context, the Ministry of Education has proposed to improve the quality of learning for the members of the recognized ethic groups by favoring practices that strengthen customs and world views of the recognized ethnic groups. The challenge taken on by the Ministry of Education is to assure that each child's cultural and linguistic needs are met in school. The goals of the Bilingual-Multicultural Education Program are as follows:

- Strengthen the identity and autonomy of indigenous children through the incorporation of educational material relevant to their culture and language.
- Improve learning for children from culturally and linguistically diverse groups by teaching the native language and a second language (the indigenous language and/or Spanish).
- Improve pedagogical practices of teachers in areas with indigenous populations.
- Incorporate methods of teaching and learning developed by the families and indigenous communities.
- Encourage the participation of indigenous communities in the construction of the curriculum in order to incorporate their knowledge, and world view into the educational process.

The strategy of the Bilingual and Multicultural Education Program is to contextualize and adapt the curriculum to the learning needs of indigenous children and to expand the curriculum in participation with indigenous families and communities. In this way the indigenous communities, rather than the teachers, have the primary responsibility for transmitting their own knowledge and language.

3. The AVENUE Project

The AVENUE Project at Carnegie Mellon University focuses on affordable machine translation for languages with scarce resources. With respect to machine translation, "scarce resources" refers to 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.

AVENUE uses a multi-engine approach to machine translation (Frederking and Nirenburg, 1994) in order to make the best use of whatever resources are available:

1. 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) (Tillmann et al., 2000).

2. If native speakers are available with training in computational linguistics, a human-engineered set of rules can be developed.

3. Finally, if neither a corpus nor a human computational linguist is available, AVENUE uses a machine learning technique called Seeded Version Space Learning (Probst, 2002) to learn translation rules from data that is elicited from a native speaker.

Because AVENUE-Mapudungun is a test case for AVENUE, we are experimenting with all three approaches, although normally we would not expect to be able to pursue all three approaches for a language with scarce resources.

4. Collaboration of LTI and Mineduc

In order to produce language technologies for bilingual education, we need people with several kinds of expertise, including computational linguists (who don't need to know the language we are processing), bilingual education experts, and native speakers with conscious and implicit linguistic knowledge about their language. The first partnership we established was between the LTI and Instituto de Estudios Indigenas (IEI — Institute for Indigenous Studies) at the Universidad de la Frontera. In a preliminary meeting in May, 2000 we agreed to collaborate in building language technologies to respond the demands of intercultural bilingual education programs for the Mapuche. After LTI and IEI agreed on the AVENUE-Mapudungun vision and goals and established a plan of action for the year 2001, the Intercultural Bilingual Education Program of Mineduc agreed to participate in the project, and to fund 90% of the Avenue/Mapudungun expenses for the year 2001. This support has been extended for the year 2002. Mineduc also provides a policy framework that allows the AVENUE-Mapudungun project to be in tune with the national plans to improve the quality and equity of the Chilean education system with respect to ethnic communities.

5. The Mapudungun Database

The first plan of action of AVENUE-Mapudungun was to make a parallel corpus of Spanish and Mapudungun 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 by a team that was assembled at the IEI. The IEI team consists of near-native speakers of at least one major dialect of Mapudungun. All of the team members but one are of Mapuche descent. They are also bilingual in Spanish, and accustomed to writing in Mapudungun. The team also includes Mapuches with training in linguistics and involvement in bilingual education.

5.1. The Corpus of Written Mapudungun

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.

5.2. The Corpus of Spoken Mapudungun

The corpus of spoken Mapudungun consists of 120 dialogues, each of which is one hour long. The content and recording methods for the spoken corpus are based on several decisions made by LTI and IEI: restricting the corpus to a limited semantic domain, inclusion of the major dialects of Mapudungun, recording quality that is suitable for speech recognition, and design of orthographical conventions to be used by AVENUE-Mapudungun.

The subject matter of the speech corpus: Since machine translation systems for restricted domains can usually achieve higher quality than general purpose machine translation, we chose to record a corpus in a limited domain, specifically primary and preventive health, both Western and Mapuche traditional medicine. A Mapudungun native speaker from the IEI team conducts conversations with informants based on a guide composed by the IEI team to grasp keywords and narrative styles used in the target domain. 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. Figure 1 contains an excerpt from the 70 questions that were used to prompt the discussion. In accordance with Mapuche culture, the interviews were scheduled ahead of time and took place in the informant's house, or in rare cases, in the informant's place of work.

The informants for the speech corpus: The age of informants are between 21 to 75 years old, most of them between 45 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. Among the informants are some machi, the Mapuches' specialized medicine wisewomen, who are asked to answer the interviewer's questions without providing specialized knowledge that is only known by and transmitted to initiated people.

I. Mantención de la salud y enfermedades
1. Chumkeymi tami külfünküleal. (Cómo hace para mantenerse as de bien.)
2. Rüfkünungey am tami amulngen kiñe machimew.
(Es verdad que el médico lo mandó donde una machi.)
. . .
II. Embarazo - Niepeklen
1.Tunten püñeñ dew nieymi. (Cuántos hijos ha tenido.)
2.Tunten mongeley. (Cuántos estn vivos.)
3.Chumngekefui tami niepüñekülen, kutrankawkefuimi kam femkelafuimi.
(Cómo eran sus embarazos. Tuvo algún problema.)
...
III. Las enfermedades - Puke kutran
1. Chumngey tami kutran. (En qué consiste su enermedad.)
2. Chem. Üy niey tami kutran ? (Cómo se llama su enfermad?)
3. Chem. Dewmangekey pelontual chem. Kutran niel?
(Qué tipo de exámenes se necesitan para efectuar el diagnostico?)
...

Figure 1: **Examples of conversation topics in the Spanish-Mapudungun parallel corpus**

The dialects included in the speech corpus: There are four major Mapudungun variants: Lafkenche, Nguluche, Pewenche and Williche. For the oral corpus the IEI team choose to work with three dialects (Lafkenche, Nguluche, Pewenche) that are quite similar with some minor semantic and phonic differences. The Williche variant presents some morpho-syntactic differences, specifically in the pronouns and verb conjugations. The IEI team will return to Williche at a later stage in the project.

Recording and transcription methods: 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 labelled.

The orthography chosen for the speech corpus: Language technologies for languages with scarce resources often suffer from the lack of a standardized character set and spelling conventions. Because of the availability of experts on the IEI team, AVENUE-Mapudungun decided to create an orthographically uniform corpus. However, because there are competing orthographies for Mapudungun, we agreed to develop orthographical conventions that would be for use only by AVENUE-Mapudungun. At a latter time, we will evaluate the social and cultural acceptability of the AVENUE-Mapudungun orthography.

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

He has sold both of his cars.
El ha vendido sus dos automóviles
fey welui ñi epu awtu

He can move both of his thumbs.
El puede mover sus dos pulgares
fey pepi newüleli ñi epu fütrarumechangll

He loves both of his sisters.
El ama a sus dos hermanas
fey poyey ñi epu deya

He loves both of his brothers.
El ama a sus dos hermanos
fey poyey ñi epu peñi

Figure 2: **Example from the Elicitation Corpus**

The IEI developed a supra-dialectal alphabet that comprises 28 letters that cover 32 phones used in the three Mapudungun variants. The main criterion for choosing alphabetic characters is to use the current Spanish keyboard that we find in all computers in Chilean offices and schools. The alphabet uses the same letters used in Spanish for those phonemes that sound like Spanish phonemes. Diacritics such as apostrophes are used for sounds that are not found in Spanish.

6. Plans for Machine Translation

As mentioned above, Mapudungun is a test case for AVENUE in which we are experimenting with three approaches to machine translation. In this section of the paper we will focus on the most experimental of our machine translation methods — automatic learning of transfer rules from carefully elicited sentences. There are four main components of the AVENUE rule-learning system: the elicitation system, morphology learning, Seeded Version Space Learning of transfer rules, and the run-time transfer rule system.

The purpose of the elicitation system is to collect a parallel corpus whose content is controlled in order to ensure

that it illustrates the basics of the language being elicited. The elicitation system (Probst et al., 2001; Probst and Levin, 2002) can be used by an informant who is bilingual in the language of elicitation and the language being elicited. In the case of AVENUE-Mapudungun, the language of elicitation is Spanish and the language being elicited is Mapudungun. The informants are required only to translate Spanish sentences into Mapudungun and to align Spanish words to Mapudungun words as well as they can. Because a human linguist may not be available to supervise the elicitation, a user interface is available for presenting sentences to an informant and allowing the informant to translate and align sentences. Some potential pitfalls of automated elicitation are described in (Probst and Levin, 2002).

A fragment of the elicitation corpus is shown in Figure 2. In each example, the elicitation sentence is shown in English and Spanish. In actual use, however, Mapudungun informants would see only the Spanish elicitation sentence. The third line of each example shows the Mapudungun translation provided by the Mapudungun informant.

The elicitation corpus follows two organizational principles. The first is compositionality. Small phrases are elicited first, and are then combined into larger phrases. For example, simple noun phrases are elicited first followed by noun phrases containing possessors, simple sentences, and multi-clausal sentences. Compositionality in the corpus facilitates the learning of compositional transfer rules.

The second organizational principle of the elicitation corpus is creation of minimal pairs of sentences. Minimal pairs of sentences differ in only one feature such as tense, number of the subject, gender of the possessor, etc. A process of feature detection compares the members of the minimal pairs in order to make a first guess at what grammatical features (verb agreement with subjects and objects, number, tense, etc.) are marked in the language being elicited. Figure 2 shows a fragment from the elicitation corpus illustrating the notion of inclusion (*both*) and alienable and inalienable (kinship and body parts) possession.

Because AVENUE's automated rule learning is still in the research stages, we are working with an elicitation corpus of around 850 sentences. The current coverage includes basic transitive and intransitive sentences, animate and inanimate subjects and objects, definite and indefinite subjects and objects, present/ongoing and past/completed events, singular, plural, and dual nouns, simple noun phrases with determiners and adjectives, and possessive noun phrases. Following guides for field workers such as (Comrie and Smith, 1977; Bouquiaux and Thomas, 1992) we expect the elicitation corpus to grow to several thousand sentences.

The elicitation corpus will be used for training automatic aquisition of MT transfer rules. We do not expect the coverage of this system to be very broad within the next year. In the mean time, we will implement EBMT based on the parallel corpus and prepare to add transfer rules into the multi-engine architecture. Immediate plans for AVENUE-Mapudungun also include making a bilingual word list based on the corpus that can be used for MT as well as bilingual education.

Acknowledgements

The AVENUE project is supported by NSF grant IIS-0121631. Preliminary funding for work on Mapudungun was also provided by DARPA.

7. References

Luc Bouquiaux and Jacqueline M.C. Thomas. 1992. *Studying and Describing Unwritten Languages*. Summer Institute of Linguistics.

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

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

Bernard Comrie and Norval Smith. 1977. Lingua Descriptive Series: Questionnaire. *Lingua*, 42:1–72.

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

Erik Peterson. Forthcoming technical report. Adapting a Transfer Engine for Rapid Machine Translation Development. Technical report, Language Technologies Institute, Carnegie Mellon University, Carnegie Mellon University.

Katharina Probst and Lori Levin. 2002. Challenges in automated elicitation of a controlled bilingual corpus. In *TMI 2002*.

Katharina Probst, 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. In *Workshop MT2010, Machine Translation Summit 2001*.

Katharina Probst. 2002. Semi-automatic learning of transfer rules for machine translation of low-density languages. To appear in Proceedings of the Student Session at ESSLLI 2002.

Christoph Tillmann, Stephan Vogel, Hermann Ney, and Hassan Sawaf. 2000. Statistical Translation of Text and Speech: First Results with the RWTH System. *Machine Translation*, 15:1–2:43–74.