Data Collection and Analysis of Mapudungun Morphology for Spelling Correction

Christian Monson[1], Lori Levin[1], Rodolfo Vega[1], Ralf Brown[1], Ariadna Font Llitjos[1], Alon Lavie[1], Jaime Carbonell[1], Eliseo Cañulef[2], Rosendo Huisca[2]

[1]Language Technologies Institute, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
[2]Instituto de Estudios Indígenas, Universidad de La Frontera
Montevideo 0870, Temuco, Chile
cmonson@cs.cmu.edu, lsl@cs.cmu.edu, rmvega@cs.cmu.edu, ralf@cs.cmu.edu, aria+@cs.cmu.edu, alavie+@cs.cmu.edu, jgc+@cs.cmu.edu, ecanulef@ufro.cl, rhuisca@entelchile.net

Abstract

This paper describes part of a three year collaboration between Carnegie Mellon University's Language Technologies Institute, the Programa de Educación Intercultural Bilingüe of the Chilean Ministry of Education, and Universidad de La Frontera (Temuco, Chile). We are currently constructing a spelling checker for Mapudungun, a polysynthetic language spoken by the Mapuche people in Chile and Argentina. The spelling checker will be built in MySpell, the spell checking system used by the open source office suite OpenOffice. This paper also describes the spoken language corpus that is used as a source of data for developing the spelling checker.

Introduction

This paper describes part of a three year collaboration between Carnegie Mellon University's Language Technologies Institute, the Programa de Educación Intercultural Bilingüe of the Chilean Ministry of Education, and Universidad de La Frontera (Temuco, Chile). In a previous paper (Levin et al., 2002) we provided an overview of the project. In this paper, we will focus on the preparation of corpora and lexica that will support an on-line lexicon and a spelling corrector for Mapudungun, an indigenous language of Chile.

Our project has scientific and social significance. The scientific novelty of the project is in the application of computational tools (such as morphological analysis, Example-Based Machine Translation, and Transfer Based MT) to a polysynthetic language. We are also working on new techniques for automatically learning transfer rules from word-aligned bilingual data (Carbonell et al., 2002; Probst et al., 2001, 2002a, 2002b, 2003; Lavie et al., in press).

The social significance of the project stems from the Chilean Ministry of Education's commitment to bilingual education in Spanish and Mapudungun for Mapuche children, where computer-based tools are a welcome part of the bilingual education program. Chile's electronic education network project, ENLACES, for example, provides computers and networking to all Chilean schools, including those in rural areas.

Mapudungun

Mapudungun, a polysynthetic language with noun and verb incorporation, is the language of over 900,000 Mapuche people in Chile and Argentina. While 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 1 contains glosses of a few morphologically complex Mapudungun verbs taken from our bilingual lexicon.

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 1: Examples of Mapudungun verbal morphology taken from our 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!

Figure 2: Excerpt from the corpus of spoken Mapudungun

Corpora and Lexica

The CMU-Chile project, Avenue-Mapudungun, is planning two tools for the near future: an on-line bilingual lexicon with examples of usage from a corpus of spoken Mapudungun, and a spelling checker for Mapudungun built on MySpell, the spell checking system used by the open source office suite OpenOffice. In support of these tools we are developing a number of corpora and lexica.

The Corpus of Spoken Mapudungun

In the last three years, the Chilean Ministry of Education and CMU's Avenue project have supported the collection of 170 hours of spoken Mapudungun. The recordings (all on the topic of health care) have been transcribed and translated into Spanish at the Instituto de Estudios Indígenas at Universidad de La Frontera. The corpus covers three dialects of Mapudungun: 120 hours of Nguluche, 30 hours of Lafkenche and 20 hours of Pewenche. A small excerpt from this spoken Mapudungun corpus can be found in Figure 2. The corpus is described in more detail in Levin et al. 2002.

Frequency Rank	Transcribed Word Form	Spelling Corrected Word Form
1	ta	ta
2	ka	ka
3	fey	fey
4	tati	tati
5	l'awen'	l'awen'
101	Ngünechen	Ngünechen
102	ko	ko
103	fel**i**	fel**ey**
104	pichikeche	pichikeche
105	kümey	kümey
10,001	chumk**u**nual	chumk**ü**nua**e**l
10,002	puedelafuy	puedelafuy
10,003	tulayi**n**	tulayi**ñ**
10,004	kimngepelay	kimngepelay
10,005	en<*SPA>la	**<*SPA>**en <*SPA>la

Table 1: Entries from the Spelling Corrected Full Form Word List

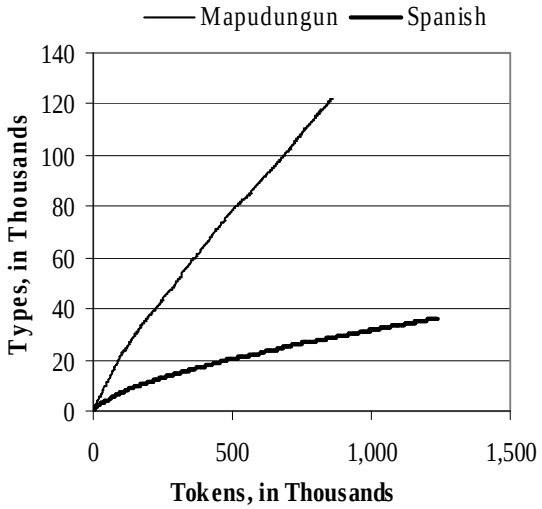

Figure 3: Type-Token Curve over the Corpus of Spoken Mapudungun and its Spanish translation

To better understand the nature of this spoken language corpus it is interesting to compare the plots, shown in Figure 3, of vocabulary size (types) vs. corpus size (tokens) over the transcribed Mapudungun and its Spanish translation. There are two reasons for the much steeper slope of the Mapudungun curve compared to that for Spanish. First, as a polysynthetic language, the absolute number of potential types is larger for Mapudungun. And second, because the orthography of Mapudungun is not standardized, alternate spellings occur in the corpus. We estimate that as many as half of the types in the transcribed corpus are misspellings or alternate spellings of words.

A set of spelling conventions has been devised by Mapuche linguists at Universidad de La Frontera for use in this project. These spelling conventions will be applied to future versions of the transcribed corpus and will be the basis of the spelling corrector discussed below.

Full Form Word List

To support a spelling checker for Mapudungun, the 70,000 most frequent full form words (stem plus inflections) were extracted from the corpus of spoken Mapudungun. These 70,000 most frequent full form words cover 57% of the word forms or types in the corpus but 94% of the tokens.

The word forms were hand checked for spelling using the spelling conventions agreed upon for this project. Selected entries from the full form word list appear in Table 1. The first column gives the frequency rank of the word form; the second column lists the word form as it appears in the transcribed spoken Mapudungun corpus; and the third column gives the spelling that follows the project's spelling conventions. (Spelling changes appear in bold in the table.) There will eventually be a version of the corpus in which all words conform to the project's spelling conventions.

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.

Figure 4: Entries from the Bilingual Lexicon

Bilingual Lexicon

Using the Corpus of spoken Mapudungun the Instituto de Estudios Indígenas at Universidad de La Frontera has begun to build a bilingual Mapudungun-Spanish lexicon. Each entry in the bilingual lexicon consists of:

- A full form Mapudungun word
- A segmentation of the word into morphemes
- A gloss for each morpheme
- A Spanish translation of the word
- A sentence from the corpus of spoken Mapudungun containing the word form
- A Spanish translation of the sentence, and
- A reference into the corpus of spoken Mapudungun identifying the specific cited sentence

Figure 4 contains sample entries from among the 1,600 currently in the lexicon. The lexicon is in a very general text only format that can be re-configured for any computer-based lexicon interface.

The morphemes were labeled by project members who have experience with the spoken language corpus, but are not linguists. For this reason, the glosses of the morphemes 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 second morpheme agrees with the higher ranking noun, in this case, second person.)

Spelling Checker

Building on the Full Form Word List, and the morphological segmentations in the Bilingual Lexicon we are currently developing a Mapudungun spelling checker to be used inside a word processor. In general, a good spelling checker will reject typos and misspelled words while accepting well formed words, even if they are morphologically complex.

One approach to building a spelling checker is to simply collect a large list of full form words (stems with affixes). While our project has built a full form word list of about 70,000 frequent word forms from the health care corpus, this is not large enough to cover the productive word formation processes of Mapudungun in domain independent text. Hence, to produce a reasonable spelling checker, we need to robustly model morphology.

We are not, however, currently building a comprehensive model of Mapudungun morphology for two reasons. First, a simple theoretical model of morphology would be too brittle. For example, while morpheme order in Mapudungun is generally fixed and while morphophonemic changes are few, there are exceptions to both of these rules. And second, the spelling correction system we have chosen has inherent limitations. We wish to create a spelling checker for a major word processor. Unfortunately commercial word processors use proprietary spelling correction systems that we currently do not have access to. Hence, we have opted to build a spelling corrector for OpenOffice, an open source graphical word processor. The spelling correction system within OpenOffice, MySpell, is limited to appending a single affix (or affix group) to a stem.

For these reasons, as a first pass at the spelling checker we will use a simple system of two lists, a list of stems, and a list of suffix groups. We will, then, allow any stem to combine with any suffix group. For example, in the last lexical entry in Figure 4, the stem is *monge* and the suffix group is *peürkelayan* .

Taking such a naïve approach to Mapudungun morphology would not be a good idea for a system designed to generate word forms. We, however, are designing a spelling recognizer and assume users will not intentionally attach verb suffixes to nouns.

In order to empirically compile the list of stems and the list of suffix groups we are following an iterative process of semiautomatic segmentation of full form words. The previously built Mapudungun-Spanish Bilingual Lexicon contains complete morphological segmentations for each of its entries. Using an uncomplicated algorithm for matching sequences of Mapudungun suffixes, these complete morphological segmentations were reduced to initial lists of stems and suffix groups.

Using these initial lists, the most frequent 1,000 word forms in the Corpus of Spoken Mapudungun were automatically segmented into stems and suffix groups. Mapuche linguists then verified and corrected the automatic segmentations. We then updated the initial lists of stems and suffix groups with the hand corrected segmentations and automatically segmented the next several thousand most frequent word forms. This second group of automatically segmented word forms is currently being corrected by native speakers. The future plan is to

iterate this process until all 70,000 most frequent word forms are correctly segmented.

Acknowledgements

This research was funded in part by NSF grant number IIS-0121-631. We would also like to thank the Chilean Ministry of Education funding the team at the Instituto de Estudios Indígenas, especially Carolina Huenchullán, the National Coordinator of the Chilean Ministry of Education's Programa de Educación Intercultural Bilingüe for her continuing support, and the team in Temuco—Flor Caniupil, Cristián Carrillán, Luis Canuipil, and Marcella Collío for their hard work in collecting, transcribing and translating the data. And Pascual Masullo for his expert linguistic advice.

References

Carbonell, J., K. Probst, E. Peterson, C. Monson, A. Lavie, R. Brown, and L. Levin. (2002). Automatic Rule Learning for Resource-Limited MT. In *Proceedings of AMTA 2002*. (Copyright Springer Verlag)

Lavie, A., S. Vogel, L. Levin, E. Peterson, K. Probst, A. Font Llitjos, R. Reynolds, J. Carbonell, and R. Cohen. (in press). Experiments with a Hindi-to-English Transfer-based MT System under a Miserly Data Scenario. *ACM Transactions on Asian Language Information Processing (TALIP)*.

Levin, L., A. Lavie, R. Vega, J. Carbonell, R. Brown, E. Canulef, and C. Huenchullan. (2002). Data Collection and Language Technologies for Mapudungun. In *Proceedings of the International Workshop on Resources and Tools in Field Linguistics*, LREC.

Probst, K., R. Brown, J. Carbonell, A. Lavie, L. Levin, and E. Peterson. (2001). Design and Implementation of Controlled Elicitation for Machine Translation of Low-density Languages. In *Proceedings of the MT 2010 Workshop at MT Summit*.

Probst, K. (2002). Semi-Automatic Learning of Transfer Rules for Machine Translation of Low-Density Languages. In *Proceedings of the ESSLLI 2002 Student Session.*

Probst, K., and L. Levin. (2002). Challenges in Automated Elicitation of a Controlled Bilingual Corpus. In *Proceedings of TMI.*

Probst, K., L. Levin, E. Peterson, A. Lavie, and J. Carbonell. (2003). MT for Resource-Poor Languages Using Elicitation-Based Learning of Syntactic Transfer Rules. To appear in: *Machine Translation, Special Issue on Embedded MT*.

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