Adaptable, Community-Controlled, Language Technologies for Language Maintenance

Lori Levin
Language Technologies Institute
Carnegie Mellon University
lsl@cs.cmu.edu

Abstract

Endangered languages may require more flexible language technologies than stable ones because they may not be standardized and they may be in a cycle of losing, replacing, and borrowing vocabulary and grammar. This paper argues that the coverage and content of language technologies should be in the hands of the speech community, and that it needs to be adaptable and learn from users. This calls for new approaches, possibly based on active learning to allow the language technologies to be as flexible and changeable as languages generally are. The paper also addresses ways in which the development of a machine translation system can be initiated when resources are scarce, including the experience of the AVENUE project with Mapudungun (Chile) and Iñupiaq (Alaska).

1 Introduction

All language professionals have been made aware of the plight of minor and endangered languages. In response, many language technologists have proposed methods for developing systems for languages that lack corpora and other resources. At the same time, speakers of endangered languages have become more aware of the potential of language technologies, bringing us to a point where we may ask ourselves how we can form partnerships that really help with language revitalization and design projects that are more than just exercises in research.

The AVENUE and LETRAS machine translation projects at Carnegie Mellon University[1] have had joint projects with two Native American language communities – the Mapuche in Chile and the Iñupiat in Alaska, speaking Mapudungun and Iñupiaq respectively. We have also had conversations with many others in order to find out what kind of language technologies can be useful.[2] As developers of machine translation, the AVENUE project team would like it to be the case that machine translation would magically revitalize a language by providing access to government, health care, education an the internet, all in the endangered language, thereby eliminating the need to use the surrounding language. However, we know that this position is naive or at least not viable in the near future. It is more likely that the goal should be a stable bilingual situation in which language technologies support the use of the endangered language without totally displacing the surrounding language.

Although promoting conversation with elders is probably the most desirable way to revitalize a language, it is also important for younger speakers to be able to communicate with each other using modern media. Margaret Noori[3] reports that her Ojibwe language students use text messaging, Facebook, Twitter, and adapted versions of video games in Ojibwe. In order to feel comfortable using these tools, non-native speakers need linguistic support. Welsh language revival is farther along in its support of modern media. The language technologies web page at Canolfan Bedwyr[4] lists lo-

[1] NSF grants IIS-0121631 and IIS-0534217.

[2] I would like to thank the following people for sharing their expertise and experience: Eliseo Cañulef, Rosendo Huisca, Edna MacLean, Lawrence Kaplan, Margaret Noori, Delyth Prys, and Per Langaard. I hope that I do not misrepresent their languages or communities. All mistakes are, of course, mine.

[3] http://www.umich.edu/~ojibwe/

[4] http://www.bangor.ac.uk/ar/cb/technolegau_iaith.php.en

calized operating systems and spelling checkers. Particularly important are tools to support texting on mobile phones[5] including access to dictionaries while texting. Speech technology is also important in language maintenance because it can allow speakers to say a word in order to learn how to spell it or spell a word to learn how to say it[6].

There is stable technology for many linguistic support tools such as spelling checkers, grammar checkers, on-line dictionaries, and speech recognition and synthesis. However, it may not be straightforward to adapt these tools to endangered languages. First, the languages may be typologically different from the ones that the technology was developed for. Polysynthetic languages are noticeably underrepresented in the world of language technologies. Second, the languages may not be standardized and there might be variation in everything from pronunciation to grammar. They may have to make up for lost words or make up new words for new things, or they may choose to borrow vocabulary from the surrounding language. They may also, unfortunately, be in the process of losing typologically rare features and gaining features of the surrounding language. Older speakers may have trouble accepting innovations in the language, but in the end, they realize that the language will only survive if it is allowed to change (Littlebear, 1999; Greymorning, 1999).

Setting aside the issue of documenting and preserving older, ”correct” forms of a language, how can we as language technologists support a language that is in the process of rapid change and is being used by speakers who may not be completely fluent? There are many examples work heading in this direction. Three examples from Carnegie Mellon University are summarized here. Schultz and Black (Schultz et al., 2007) describe SPICE, a web based environment for building speech recognition and synthesis. It allows non-experts to enter initial data for training and confirm or disconfirm predicted pronunciations and spellings for additional data. Since the interface is easily usable by people who are not language technologists, the coverage and output of the system can be changed frequently. In statistical machine translation, Rogati (Rogati, 2009) uses active learning to reduce the amount of training data that is needed for domain adaptation by finding data that will have the most impact on performance. Font Llitjos (Font Llitjós, 2007), working with the AVENUE MT system, describes a Translation Correction Tool (TCT) that is operated by an MT user and allows the user to alter the behavior of an MT system. The user corrects erroneous translations and produces correct translations. The transfer rules that produced the erroneous translations are then automatically corrected.

2 The AVENUE Iñupiaq and Mapudungun Partnerships

The AVENUE machine translation framework developed at Carnegie Mellon University has been applied to many high resource and low resource languages, including two indigenous Western Hemisphere languages, Mapudungun (Chile) and Iñupiaq (Alaska). The full AVENUE framework includes several steps: (1) elicitation of data from native speakers, (2) automatic learning of transfer rules in a unification based synchronous grammar formalism based on the elicited data, (3)optional hand written transfer rules, (4) decoding, and (5) translation correction (as described above). For high resource languages, other techniques may be used such as statistical word alignment and extraction of syntactic phrases (Lavie, 2008; Hanneman and Lavie, 2009). These steps have not all been implemented for Mapudungun and Iñupiaq, but work is in progress.

2.1 Data Collection

Mapudungun and Iñupiaq are both low-resource languages in the sense that large corpora and dictionaries in electronic form are not available. (Although more resources are becoming available for Mapudungun.) Both languages have descriptive grammars, however, and there are native speakers who are linguists and language experts. Our partners include Edna MacLean and Larry Kaplan for Iñupiaq and Eliseo Cañulef and Rosendo Huisca for Mapudungun. The partner institutions were the Alaska Native Language Center (ANLC) at the University of Alaska at Fairbanks, the Universidad de la Frontera (UFRO) in Temuco, Chile, and the Chilean Ministry of Education.

We have proceeded with data collection in very different ways for Mapudungun and Iñupiaq based on resources that were available. Because we had a reasonable amount of funding for our initial work on Mapudungun, the UFRO team along

[5] http://news.bbc.co.uk/1/hi/wales/4777933.stm
[6] http://www.abair.tcd.ie

with Rodolfo Vega from Carnegie Mellon (CMU) collected and transcribed 170 hours of spoken Mapudungun (Monson et al., 2004). For Iñupiaq we have been pursuing other methods for acquiring data. The AVENUE elicitation corpus (Levin et al., 2006) consists of 3000 simple sentences illustrating grammatical features such as person, number, tense, aspect, animacy, and definiteness, as well as constructions such as relative clauses and questions. Edna MacLean translated the sentences into Iñupiaq and provided interlinear glosses. Some scanned texts were collected from ANLC and were typed by CMU undergraduates[7] resulting in a small corpus of 126K bytes. In addition, Shinjae Yoo at CMU is pursuing OCR with character n-grams for error correction as a method for increasing the size of the corpus.

2.2 Polysynthetic Morphology

Both Mapudungun and Iñupiaq are polysynthetic languages. Mapudungun stems can be simple or compounded. The compounds can involve noun incorporation, although this is becoming more rare, or verb compounding. After a verb stem there can be many closed class morphemes covering things like tense, aspect, agreement, passive and inverse voices, negation, and some adverbial and deictic meanings (Smeets, 1989; Zuñiga, 2000). Iñupiaq verbs also begin with stems followed by a large class of postbases, some of which have meanings related to English modal verbs and derivational morphemes (MacLean, 1993; MacLean, 1995). Inflectional morphemes follow the postbases. Iñupiaq has ergative case marking. Both Mapudungun and Iñupiaq have singular, dual, and plural number. Following are some examples of words in Mapudungun and Iñupiaq[8].

```
Mapudungun:
Treka  -l     -ke   -n.
walk   -CAUS  -HAB  -1.sg.IND
I usually make someone walk.

Mapudungun:
Kintu  -mara  -n.
hunt   -hare  -1sS/IND
I hunted (a/the) hare(s).


Inupiaq:
Imaqpaqaghaluaghniqsuq.

imaq  -qpak -qaq
water -big  -have
-kaluaq       -niq         -suq
-nevertheless -apparently  -past.3.sg

Nevertheless it apparently had
a big body of water.
```

For both languages, building a morphological analyzer was a pre-requisite to doing any other work. The Mapudungun morphological analyzer was built by Carlos Fasola, Roberto Aranovich, and Christian Monson based on data provided by our partners at the Universidad de la Frontera (UFRO). The CMU team sorted the lexical items in the corpus by frequency, and the UFRO team provided morphological segmentation and glossing of the most common words. Because Mapudungun does not have much morpho-phonology at morpheme boundaries, the CMU team built a simple analyzer based on the legal order of morphemes (Aranovich, 2006). It does not take into account co-occurrence restrictions between morphemes and therefore produces spurious analyses of some morpheme sequences, which are then weeded out during machine translation. The Iñupiaq morphological analyzer is being implemented by Aric Bills based on published grammars by Edna MacLean (MacLean, 1993; MacLean, 1995) and is quite different from the Mapudungun system. Iñupiaq has extensive morphophonemic changes at morpheme boundaries. Inspired by Per Langaard's work on a morphological analyzer and spelling checker for Kalaallisut (Greenland), which is related to Iñupiaq, we decided to implement a transducer using the Xerox Finite State tools.

2.3 Machine Translation

We have not yet implemented automatic rule learning for Mapudungun or Iñupiaq. However, Roberto Aranovich (Aranovich, 2006; Font Llitjòs et al., 2005) has produced a hand-written transfer grammar for Mapudungun-to-Spanish MT. The system is currently small and is awaiting further development. It was tested on simple but unseen sentences from a textbook with about 65% accuracy after unkown vocabulary items were added. The main issue that was encountered was translation of negation, tense, and aspect morphemes. The AVENUE grammar formalism is synchronous, assuming corresponding source and target language rules applying in step with each other. Spaces were inserted between Mapudungun morphemes before translation so that each morpheme

[7] We would like to thank Ida Mayer, J. Eliot DeGolia, and Sai Venkateswaran for this work.

[8] The digraph *gh* is used in place of a dotted *g* in Iñupaiq.

would appear as a separate word, but we could not write transfer rules for every possible combination of Mapudungun morphemes and every possible corresponding sequence of Spanish words. Furthermore, in order to determine the tense of a Spanish sentence, it is sometimes necessary to look at multiple, non-adjacent morphemes in Mapudungun. The problem was solved using feature structures and unification, which are a part of the AVENUE transfer rule formalism. The Mapudungun morphemes were parsed and their features were stored in a feature structure until there was sufficient information to generate corresponding inflections, auxiliary verbs, negation, and adverbs in Spanish. In effect, transfer using synchronous grammars was not found to be useful for languages that are as different as Spanish and Mapudungun, but unification was found to be helpful.

3 Concluding Remarks

So far, this paper has recommended that language technologies for endangered languages be adaptable and community controlled in order to match the dynamic nature of language change and revitalization. Two additional issues related to endangered languages which are evident in our experience with Mapudungun and Iñupiaq are lack of electronic resources and typological divergences from major languages. It was suggested by Per Langaard that these problems could be solved by translating between related endangered languages. For example, Kalaallisut and Iñupaiq are related but are not equal in resources. Kalaallisut has newspapers, literature, and a textbooks for a full school curriculum. MT from Kalaallisut to Iñupiaq would probably produce more authentic and native sounding output than translation from English to Iñupiaq and could produce much needed literature and educational materials in Iñupiaq. Many other language families could also benefit from pooling resources in this way.

References

Aranovich, Roberto. 2006. *Handling of Translation Divergences in the Mapudungun/Spanish AVENUE Transfer Grammar and Lexicon*. Comprehensive Exam Paper, University of Pittsburgh.

Font Llitjòs, A., R. Aranovich, and L. Levin. 2005. Building machine translation systems for indigenous languages of latin america. In *Second Conference on the Indigenous Languages of Latin America (CILLA II)*.

Font Llitjós, Ariadna. 2007. *Automatic Improvement of Machine Translation Systems*. Ph.D. Thesis, Carnegie-Mellon University, School of Computer Science.

Greymorning, Stephen. 1999. Running the gauntlet of an indigenous language program. In *Revitalizing Indigenous Languages*, pages 6–16. Internet publication.

Hanneman, G. and A. Lavie. 2009. Decoding with syntactic and non-syntactic phrases in a syntax-based machine translation system. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*.

Lavie, Alon. 2008. Stat-xfer: A general search-based syntax-driven framework for machine translation. In Gelbuch, editor, *Computational Linguistics and Intelligent Text Processing*, pages 362–375. Springer, LNCS 4919.

Levin, L., J. Good, A. Alvarez, and R. Frederking. 2006. Parallel reverse treebanks for the discovery of morpho-syntactic markings. In *Proceedings of Treebanks and Linguistic Theories*.

Littlebear, Richard. 1999. Some rare and radical ideas for keeping indigenous languages alive. In Reyner, J., G. Cantoni, R.N. St. Clair, and E. Parsons Yazzie, editors, *Revitalizing Indigenous Languages*, pages 1–5. Internet publication.

MacLean, Edna. 1993. *North Slope Inupiaq Grammar: First Year (Revised)*. Alaska Native Language Center.

MacLean, Edna. 1995. *North Slope Inupiaq Grammar: Second Year (Revised)*. Alaska Native Language Center.

Monson, C., L. Levin, R. Vega, R. Brown, A. Font Llitjos, A. Lavie, C. Carbonell, E. Canulef, and R. Huisca. 2004. Data collection and analysis of mapudungun morphology for spelling correction. In *LREC*.

Rogati, Monica. 2009. *Domain Adaptation of Translation Models for Multilingual Applications*. Ph.D. Thesis (in progress), Carnegie-Mellon University, School of Computer Science.

Schultz, T., A. Black, S. Badaskar, M. Hornyak, and J. Kominek. 2007. Spice: Web-based tools for rapid language adaptation in speech processing systems. In *Interspeech*.

Smeets, I. 1989. *A Mapuche Grammar*. Ph.D. Thesis, University of Leiden.

Zuñiga, F. 2000. *Mapudungun*. Muenchen: Lincom Europa.