Linguistic Structure and Bilingual Informants Help Induce Machine Translation of Lesser-Resourced Languages

Christian Monson[1], Ariadna Font Llitjós[2], Vamshi Ambati[1], Lori Levin[1], Alon Lavie[1], Alison Alvarez[1], Roberto Aranovich[3], Jaime Carbonell[1], Robert Frederking[1], Erik Peterson[1], Katharina Probst[4]

[1] Language Technologies Inst.
Carnegie Mellon University
Pittsburgh, PA, USA 15213
cmonson@cs.cmu.edu

[2] Vivísimo, Inc.
1710 Murray Avenue
Pittsburgh, PA 15217
ariadna.font@vivisimo.com

[3] Linguistics Department
University of Pittsburgh
Pittsburgh, PA, 15260
roa6+@pitt.edu

[4] Accenture Technology Labs
161 North Clark Street
Chicago, IL 6060
katharina.a.probst@accenture.com

Abstract

Producing machine translation (MT) for the many minority languages in the world is a serious challenge. Minority languages typically have few resources for building MT systems. For many minor languages there is little machine readable text, few knowledgeable linguists, and little money available for MT development. For these reasons, our research programs on minority language MT have focused on leveraging to the maximum extent two resources that are available for minority languages: linguistic structure and bilingual informants. All natural languages contain linguistic structure. And although the details of that linguistic structure vary from language to language, language universals such as context-free syntactic structure and the paradigmatic structure of inflectional morphology, allow us to learn the specific details of a minority language. Similarly, most minority languages possess speakers who are bilingual with the major language of the area. This paper discusses our efforts to utilize linguistic structure and the translation information that bilingual informants can provide in three sub-areas of our rapid development MT program: morphology induction, syntactic transfer rule learning, and refinement of imperfect learned rules.

Introduction

Speakers of minority languages could benefit from fluent machine translation (MT) between their native tongue and the dominant language of their region. But scarcity in capital and know-how has largely restricted machine translation to the dominant languages of first world nations. To lower the barriers surrounding MT system creation, we must reduce the time and resources needed to develop MT for new language pairs. This paper discusses our application of two underutilized resource reduction techniques in three sub-areas of rapid MT system development. By, first, incorporating linguistic structure into MT knowledge induction algorithms, and, second, strategically employing minimally trained bilingual informants during system creation we: 1. learn morphological paradigms without supervision, 2. induce syntactic transfer rules, and, 3. automatically correct transfer rules in an interactive fashion.

Minor Languages

The AVENUE project has developed prototype machine translation systems for several minor languages from the Americas. We have worked most extensively with Mapudungun, an indigenous language spoken by more than 900,000 people in central Chile and adjacent Argentina. Our project has produced a prototype rule-based Mapudungun-Spanish MT system (Font Llitjós, Levin, and Aranovich, 2005). In addition to our work on Mapudungun, We have built a prototype Quechua to Spanish MT system. Quechua is spoken by several million people in and around Peru and Bolivia. Currently we are working with the Alaska Native Language Center and the Inupiat community to build an MT system for Inupiaq, the most northern indigenous language of Alaska. And we are collaborating with the Universidade de São Paulo to develop MT systems for indigenous languages of Brazil.

Leveraging Linguistic Structure and Bilingual Informants to Learn Machine Translation Systems

Our approach to machine translation seeks to leverage the structure of natural language to automatically induce MT systems; at times with deliberate input from bilingual informants. Our morphology learning system exploits the inherent organizational structure of natural language morphology: the paradigm (Stump, 2001). Similarly, we leverage syntactic structure to automatically learn transfer rules for MT systems. Additionally, input from bilingual informants provides translations of key syntactic structures to seed rule creation. And informants' corrections of translation mistakes facilitate automatic rule refinement.

		Locative	Aspect	Habitual	Possible	Reportative	Polarity and Mood	Tense	Object Agreement	Subject Agreement and Mood
Stem	…	-pa-	-tu-	-ke-	-pe-	-(ü)rke-	-la-	-a-	-fi-	-(ü)n
										-li
		-pu-	-ka-				-ki-	-fu-		-chi
				-Ø-	-Ø-	-Ø-	-nu-	-afu-	-Ø-	-yu
		-Ø-	-Ø-							-liu
							-Ø-	-Ø-		…

Table 1: A portion of the verbal morphology of Mapudungun. Each column headed by one or more morphosyntactic feature categories is a paradigm. Each paradigm consists of at least two cells, the boxes beneath the feature heading. Each cell marks a verb for a specific value of the feature category heading that paradigm. This figure is adapted from Smeets (1989) with personal experience.

Morphology

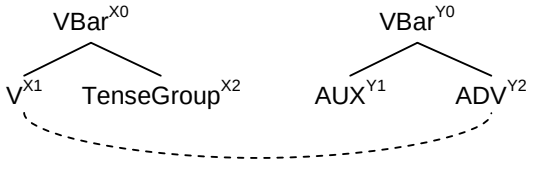
The syntactic-transfer methodology which forms the core of our MT system requires that words first be analyzed into constituent morphemes. Just as machine translation systems are not available for most minority languages, morphological analysis systems have not been developed for these languages either. For our Mapudungun and Quechua MT systems we hand built morphological analyzers. We are currently developing a language independent morphological analysis system that can learn to segment the word forms of a new language by examining a moderate sized monolingual text corpus of that language.

Consider the morphological structure of Mapudungun. Mapudungun morphology is usually described as a slot system with as many as 35 slots (e.g. Smeets, 1989). Each slot is a paradigm, and either the presence or absence of a morpheme in any given slot fills a cell of the paradigm of that slot. Table 1 organizes the trailing paradigms of Mapudungun in slot order, giving the suffixes that can fill the cells of each paradigm.

Our unsupervised morphology induction system follows the lead of other morphology induction systems (Goldsmith, 2001; Snover, 2002) in leveraging the paradigm structure of natural language to learn the morphology of specific languages. Because our unsupervised morphology induction system relies on the paradigm structure of morphology, we christened our system ParaMor. ParaMor discovers the paradigm system of a new language by comparing surface word forms found in a corpus. ParaMor is a two stage algorithm. In the first stage, ParaMor creates paradigm models, while in the second stage ParaMor segments words into morphemes by matching words against the paradigm models. The first stage, paradigm creation, is further broken down into three steps. First, ParaMor greedily and aggressively searches for sets of contrastive word-final strings. Many of the initially selected sets of strings do not represent true paradigms. Of those that do represent paradigms, most capture only a portion of a complete paradigm. Second, ParaMor merges candidate paradigm pieces into larger groups covering more of the affixes in a paradigm. And the third step of paradigm model creation filters out the poorer candidates.

ParaMor placed strongly in Morpho Challenge 2007 (Kurimo, Creutz, and Varjokallio, 2007), a competition pitting unsupervised morphology induction algorithms head to head. Systems participating in Morpho Challenge 2007 were evaluated in two ways: first, in a linguistically motivated assessment of morpheme identification; second, in a task-based evaluation that augmented an information retrieval system with morphological segmentations. Systems competing in Morpho Challenge 2007 segmented wordforms from up to four languages: English, German, Turkish, and Finnish. ParaMor officially competed in the English and German language tracks. In the linguistic assessment of the English track, ParaMor identified morphemes with more accuracy than a state-of-the-art unsupervised morphology induction algorithm which served as a baseline, Morfessor (Creutz, 2006). ParaMor placed fourth among all submitted algorithms in the English linguistic evaluation. In the German linguistic assessment, a system combining the output of ParaMor with output from Morfessor tied for first place. For additional details on ParaMor's algorithms and further analysis of performance in Morpho Challenge 2007 please see Monson et al. (2007a; 2007b).

Following the Morpho Challenge deadline, we adapted ParaMor's algorithms to recognize agglutinative sequences of suffixes and submitted morphological analyses of all four language tracks to the Morpho Challenge committee for evaluation. ParaMor's performance on morpheme identification in Finnish is statistically indistinguishable from the best placing system in Morpho Challenge 2007; in Turkish, ParaMor achieves much higher morpheme recall than any officially competing system, significantly outperforming the highest placing system at F_1 over morpheme identification, 52.0% vs. 24.7%. The adapted version of ParaMor also performs well in the information retrieval (IR) evaluation. An IR system augmented with morphological analyses from the adapted version of ParaMor consistently improves the average precision of retrieved newswire documents over an unaugmented IR system. In the English and German tracks, ParaMor's IR performance is as good as the performance of the top systems that were officially submitted to the Morpho Challenge. Particular results covering ParaMor adapted for agglutination will appear in Monson (2008, In Press).




```
(X1 lexicalaspect)   =c  (X2 myVerbLexAsp)
(X2 voice)           =   *UNDEFINED*
(X1 morph)           =c  deadv

(X0 morph)           =   (X1 morph)
(X0 person)          =   (X2 person)
(X0 number)          =   (X2 number)
(X0 mood)            =   (X2 mood)
…

(Y0 person)          =   (X0 person)
(Y0 number)          =   (X0 number)
…

(Y0 person)          =   (Y1 person)
(Y0 number)          =   (Y1 number)
…

(Y1 form)            =c  (Y2 AUX form)
```

Figure 1: This syntactic transfer rule translates Mapudungun deadjectival verbs into Spanish adverbial constructions, i.e. Mapudungun: kümelen becomes Spanish: estoy bien meaning I'm fine.

Syntax

For minor languages there is often little if any machine readable text data from which to induce an MT system. This text shortage necessitates bilingual speakers creating new parallel data. We maximize the usefulness of the relatively few translation examples that a bilingual informant can produce, by adopting a syntactic transfer formalism. And we have specially designed a corpus to contain sentences that are targeted to elicit common syntactic structures (Alvarez et al., 2006; Probst et al., 2001). The targeted elicitation corpus, in combination with the syntactic nature of our MT rules, facilitates automatic induction of translation rules.

An example of a syntactic transfer rule our MT engine can interpret is given in Figure 1. The transfer rule in Figure 1 is part of a hand-written grammar to translate Mapudungun into Spanish. Concepts that Spanish (and English) express with a copula and adverb, Mapudungun expresses with an inflected verb. For example, the Mapudungun verb *kümelen*, from the adjective stem *küme* 'good,' translates to Spanish as *estoy bien* 'I'm fine'. The syntactic transfer rule in Figure 1 describes this translation mismatch between Mapudungun and Spanish. At the top of Figure 1 are two context-free phrase structure trees. The tree on the left breaks Mapudungun deadjectival verbs into a verb stem (V) and the verbal suffixes (TenseGroup); the tree on the right captures the syntax for Spanish copula plus adverb constructions. The dotted line connecting the Mapudungun V to the Spanish ADV is our formalism explicitly aligning the Mapudungun verb to the Spanish adjective with the corresponding meaning. Beneath the phrase structure rules in Figure 1 are equations for manipulating feature structures associated with the phrase structure nodes. The feature equations in Figure 1 unify the agreement features from the Mapudungun verb with the agreement features of the Spanish auxiliary—for in Spanish it is not meaning bearing adverb that inflects but the copula.

By leveraging the syntactic structure that our translation formalism captures, we have developed two separate algorithms that induce syntactic transfer MT systems. We have applied our first induction algorithm Probst (2005) to learn machine translation systems for Hebrew and Hindi and obtained encouraging translation results Lavie et al. (2004; 2003). This paper will focus on the our recently designed second syntactic induction algorithm. Our second induction algorithm is intended to suit a variety of scenarios of Machine Translation including resource-rich to resource-poor language scenarios, with syntax information for only one language, and resource-rich to resource-rich scenarios, with syntax information for both languages. In this paper we concentrate on the approach taken for the resource-rich to resource-poor rule induction.

Rule Induction by Projection

Our rule learning algorithm takes as input a source and target sentence pair along with word level alignment information. We also require a full syntactic parse to be available for the resource-rich side of the parallel sentence pair, which is not difficult to obtain. Given this information we start by traversing the source side syntax tree beginning from the root. At each node of the source tree we calculate the smallest contiguous segment in the target side that is "consistently" aligned with all the words in the yield of this source node. Consistent alignment is the well-formedness constraint which requires all the words in a particular segment of the source side to align with a particular contiguous segment of the target sentence, as decided by the word-level alignment. For example, in Figure 2 the Urdu string 'ek seb khaya' is contiguously aligned to the yield of the English VP 'ate an apple'. If a consistent alignment is found we mark the source node as a decomposition point. We then traverse further down the tree to identify all such points along all the children of every node until we reach the leaves of the tree.

Once the decomposition points and the corresponding target sub-sentential segments have been identified, we extract syntactic rules. For each source node we obtain the 'minimal' tree segment that has only decomposition nodes and leaf notes at its frontier. Such a tree fragment can be flattened out on the source side to form a context free rule. For the target side context free rule we project the source side syntactic categories across the decomposition nodes of the minimal tree. An example tracing the rule extraction process is shown in Figure 2.

The rule learning process outputs three kinds of primary resources. First, we obtain synchronous transfer rules that are responsible for the reordering in the translation process. Note that our induced rules do not currently contain feature unification equations. Second, we obtain syntactic phrasal lexicons or tables. An entry in the phrasal lexicon is the yield of a particular decomposition node in a source tree and its equivalent contiguous trans-

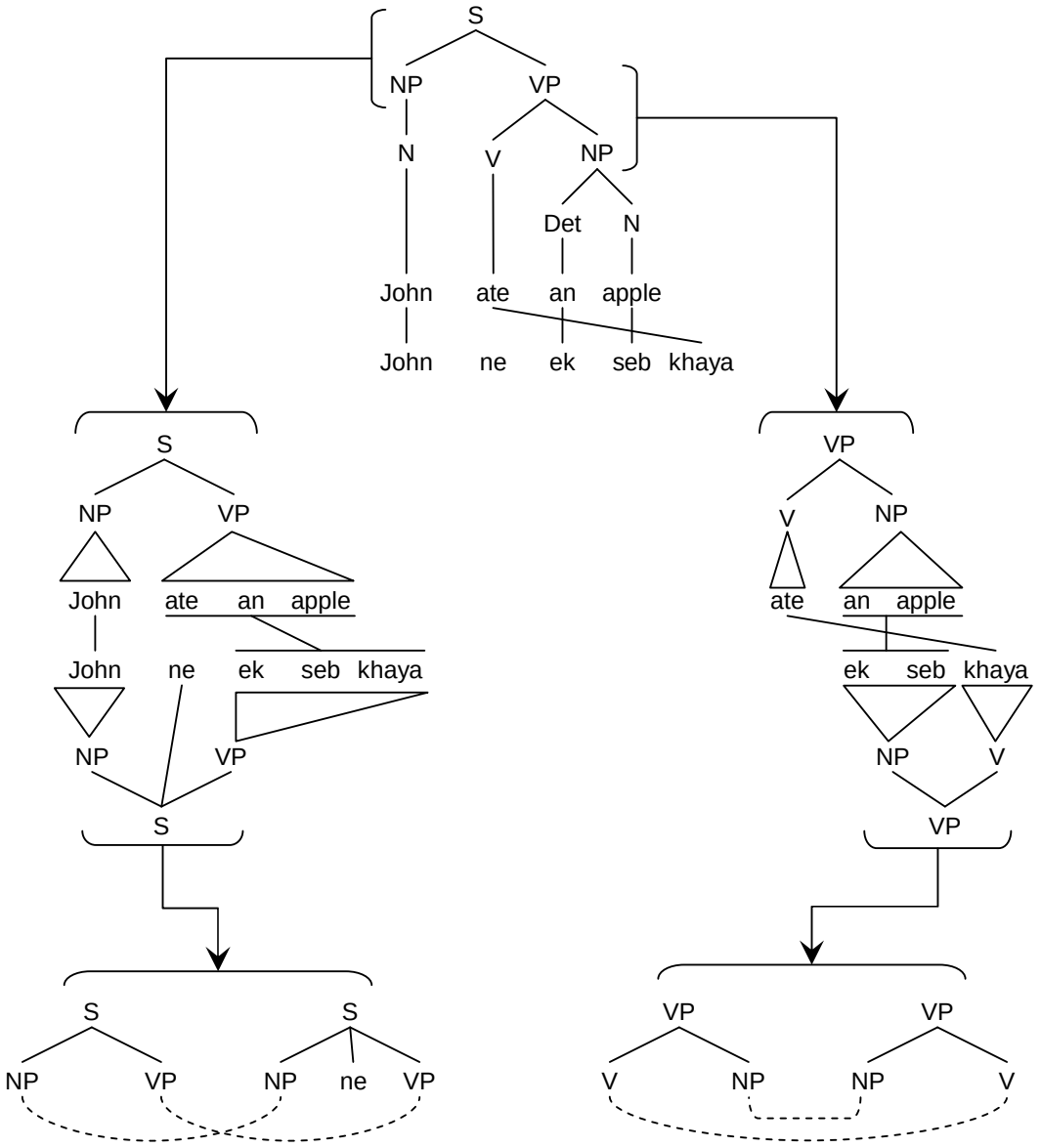


Figure 2: Sample Urdu to English syntax rules our projection induction algorithm would obtain. Structures like that at the top of this figure are input to our induction system. The input consists of parsed English that has been translated and aligned with a second language. Our algorithm extracts sub-trees from the input structure that are consistent with the word alignments. Two extracted sub-trees are pictured here, one headed by S, one by VP. The syntactic categories of English are then projected along word alignments to the second language. Finally, our algorithm produces synchronous context-free rules that capture word order differences between the aligned languages.

lation segment in the corresponding parallel target sentence. Third, we also extract the leaves in the source tree wherever they are one-to-one aligned with a word in the target sentence. The extracted leaves form a word level lexicon annotated with part-of-speech information.

Application of Rule Induction to Various Languages

We have applied our algorithm to several language pairs including English-Urdu and English-Telugu. Although both of these are languages with millions of speakers, these languages have relative little readily available machine readable text and therefore constitute resource-poor languages for machine translation. Bilingual speakers translated our elicitation corpus from English into Urdu, and, separately, into Telugu. From the three thousand translated sentence pairs from each language we extracted syntactic transfer rules. Table 2 shows some details of the rules extracted.

Our rule induction algorithm is generic and can be applied to not only resource-poor languages, but also re-

source-rich languages like French, German and Chinese, which each have extensive parallel corpora with English. The challenging part of expanding to resource-rich languages from resource-poor languages is the computational complexity of extracting rules which number in the millions. Furthermore, the extraction algorithm must be adapted to account for the extra linguistic knowledge, including syntactic parses, that may be available for the second language. Since this is out of the scope of the current paper, we skip detailed discussion, but provide rule extraction statistics for some languages in Table 2.

	# of Sentences	Structural Rules	Rules with freq > 2	Phrasal lexicon
English-Urdu	3126	6824	640	12456
English-Telugu	3126	7543	721	13500
English-German	300K	183K	16K	680K
English-French	1200K	1.3M	45K	4.2M

Table 2: Statistics of rule induction from MT in resource-rich to resource-poor language scenarios with limited corpora and in resource-rich to resource-rich language scenarios with large corpora

Syntactic Refinement

We are developing algorithms to automatically expand and improve translation rules that have been previously written or induced. Bilingual speakers who are not linguists or MT experts may be the only source of knowledge readily available for resource-poor languages. Hence, we have designed and implemented a user-friendly online graphical interface called the Translation Correction Tool (TCTool), shown in Figure 3.

The TCTool allows non-experts to detect and remediate errors in MT output. The tool graphically presents the source language sentence and a target language automatic translation that needs correction. For example, given the Mapudungun sentence: *pu püchükeche awkantuy kiñe awkantun* (*children played a game*), our prototype hand-written translation grammar for our MT system outputs the (incorrect) Spanish translation: **niños jugaron un juego* (left snapshot in Figure 3). To make this translation acceptable in Spanish, a bilingual speaker clicked on the [New Word] button on the top right corner of the TCTool and typed in the missing determiner (*los*). The bilingual speaker then dragged the newly inserted word *los* into the correct position in the translation, the beginning of the sentence, as shown in the right snapshot of Figure 3. The resulting corrected translation is thus: *los niños jugaron un juego*. New syntactic structures that result from automatic refinements are correction-driven, and thus, rather than guaranteeing linguistic perfection, this automatic process guarantees wider coverage based on usage. Alignment information determines whether changes are done at the lexical level (collocations) or at the grammar level. ARR parameters can be set to only execute refinements to the translation rules when sufficient information is available to reliably modify the grammar and lexicon.

We conducted a user study to measure the accuracy with which bilingual speakers identify and correct machine translations. Ten bilingual speakers of Spanish and English corrected the translations of 32 sentences. A small hand-written English to Spanish grammar produced the sentence translations using our rule-based MT system. These ten non-expert bilingual speakers reliably corrected MT errors 90% of the time.

The TCTool outputs correction instances, such as that shown in Figure 4. Correction instances are fed to the next stage of syntactic refinement, an automatic rule refiner (ARR). The ARR modifies the original grammar to account for each correction instance. The ARR can automatically add missing lexical entries, perform structural modifications of existing grammar rules, and fix incomplete or incorrect rules that applied during the generation of MT output.

We developed and tested our rule refinement approach on English to Spanish MT. And we have successfully

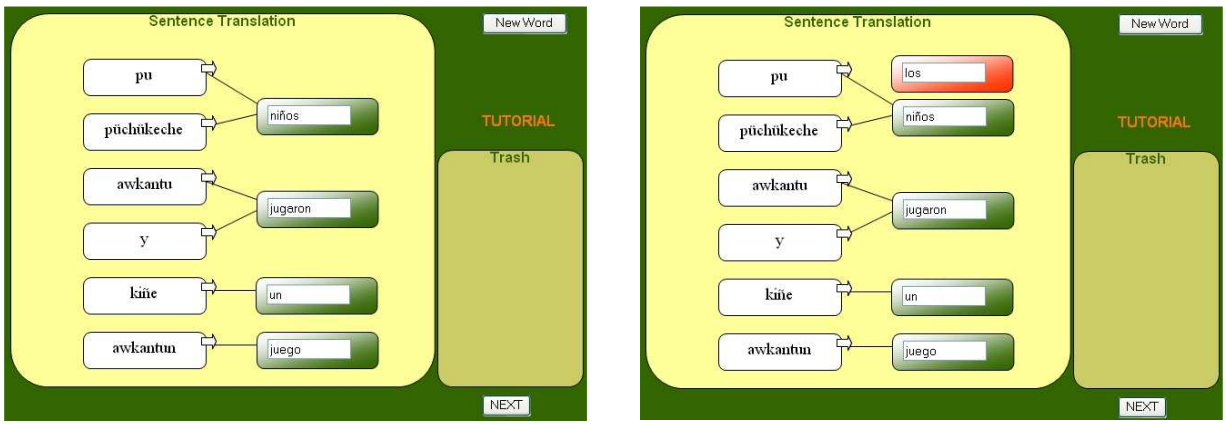


Figure 3: Snapshots of the Translation Correction Tool, before and after correcting a Spanish sentence that was automatically translated into Mapudungun.

SL: pu püchükeche awkantu y kiñe awkantun
TL: niños jugaron un juego
AL: ((1,1),(2,1)),(3,2),(4,2),(5,3),(6,4))
Action 1: add (W_1=los)

C_TL: los niños jugaron un juego
CAL: ((1,2),(2,*2*)),(3,*3*),(4,*3*),(5,*4*),(6,*5*))

Figure 4: The correction instance extracted from a TCTool log file which corresponds to the user interaction shown in Figure 4.

applied our approach to Mapudungun-Spanish MT, showing generality. Experiments with our English-Spanish MT system have demonstrated statistically significant improvements on unseen data, as measured by standard MT evaluation metrics. For an in depth description of our automatic rule refinement approach, see Font Llitjós (2007).

Conclusions

The paired aids of linguistic structure and strategic use of bilingual informants have guided our research efforts to develop methods for machine translation for minor languages. By applying these two aids to such MT subproblems as morphological analysis, syntax induction, and syntax refinement, we have begun to overcome the significant challenges of resource scarcity that minor languages pose.

Acknowledgments

The research described in this paper was supported by NSF grants IIS-0121631 (AVENUE) and IIS-0534217 (LETRAS), with supplemental funding from NSF's Office of Polar Programs and Office of International Science and Education.

References

Alvarez, Alison, Lori Levin, Robert Frederking, Simon Fung, Donna Gates. *The MILE Corpus for Less Commonly Taught Languages*. HLT-NAACL. New York, New York, USA. 2006.

Creutz, Mathias. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition.* Ph.D. Thesis in Computer and Information Science, Report D13. Helsinki: University of Technology, Espoo, Finland, 2006.

Font Llitjós, Ariadna. *Automatic Improvement of Machine Translation Systems.* Ph.D. Thesis in Language and Information Technologies. Pittsburgh: Carnegie Mellon University, USA, 2007.

Font Llitjós, Ariadna, Lori Levin, and Roberto Aranovich. *Building Machine Translation Systems for Indigenous Languages.* Second Conference on the Indigenous Languages of Latin America (CILLA II). Texas, USA. 2005.

Goldsmith, John. *Unsupervised Learning of the Morphology of a Natural Language.* Computational Linguistics. 27.2, 2001, pp153-198.

Kurimo, Mikko, Mathias Creutz, and Matti Varjokallio. "Unsupervised Morpheme Analysis – Morpho Challenge 2007." March 26, 2007. <http://www.cis.hut.fi/-morphochallenge2007/>

Lavie, Alon, Erik Peterson, Katharina Probst, Shuly Wintner, and Yaniv Eytani. *Rapid prototyping of a transfer-based Hebrew-to-English Machine Translation system.* 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI). 2004.

Lavie, Alon, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitjós, Rachel Reynolds, Jaime Carbonell, and Richard Cohen. *Experiments with a Hindi-to-English Transfer-based MT System under a Miserly Data Scenario.* ACM Transactions on Asian Language Processing. Vol. 2, No. 2, June 2003, pp143-163.

Monson, Christian. *ParaMor: from Paradigm Structure to Natural Language Morphology Induction.* Ph.D. Thesis in Language and Information Technologies. Pittsburgh: Carnegie Mellon University, USA, 2008. In Press.

Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. *ParaMor: Finding Paradigms across Morphology*. Workshop of Morpho Challenge 2007. Budapest, Hungary. 2007a.

Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. *ParaMor: Minimally Supervised Induction of Paradigm Structure and Morphological Analysis.* Computing and Historical Phonology: Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology. 2007b.

Probst, Katharina. *Automatically Induced Syntactic Transfer Rules for Machine Translation under a Very Limited Data Scenario*. Ph.D. Thesis in Language and Information Technologies. Pittsburgh: Carnegie Mellon University, USA, 2005.

Probst, Katharina, Ralf Brown, Jaime Carbonell, Alon Lavie, Lori Levin, and Erik Peterson. *Design and Implementation of Controlled Elicitation for Machine Translation of Low-density Languages*. MT2010 Workshop at the 8th Machine Translation Summit 2001 (MT-Summit). 2001.

Smeets, Ineke. *A Mapuche Grammar*. PhD Dissertation. University of Leiden. 1989.

Snover, Matthew G. *An Unsupervised Knowledge Free Algorithm for the Learning of Morphology in Natural Languages*. M.S. Thesis in Computer Science. Saint Louis, Missouri: Sever Institute of Technology, Washington University. 2002.

Stump, Gregory T. Inflectional Morphology: A Theory of Paradigm Structure. Cambridge University Press. 2001.