An Assessment of Language Elicitation without the Supervision of a Linguist

Alison Alvarez, Lori Levin, Robert Frederking {[nosila|lsl|ref]@cs.cmu.edu}
Jill Lehman {jill@kidaccess.com}
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, Pennsylvania

ABSTRACT

The AVENUE machine translation system is designed for resource poor scenarios in which parallel corpora are not available. In this situation, parallel corpora are created by bilingual consultants who translate an *elicitation corpus* into their languages. We have described the elicitation corpus in other publications. This paper is concerned with evaluation of the elicitation corpus: is it suitably designed so that a bilingual consultant can produce reliable data without the supervision of a linguist? We evaluated two translations of the English elicitation corpus, one into Thai and one into Bengali. Two types of evaluation were conducted: an error analysis of the translations produced by the Thai and Bengali consultants, and a comparison of Example Based MT trained on the original translations and on corrected translations.

1 INTRODUCTION

MT systems can be learned from large parallel corpora or they can be produced by humans writing rules. A few researchers have investigated whether, in the absence of human rule writers and corpora, an MT system can be learned from linguistically naïve human consultants (McShane and Nirenburg, 2003, McShane et al. 2002; Probst, 2005). Two approaches have been taken. The Boas system (McShane et al, 2002) trains the consultants in linguistic terminology and then asks them whether their language has, for example, nominative case or dual number. Our work relies on having the consultant translate a list of sentences, or "elicitation corpus", that is like a fieldworker's questionnaire. Each sentence is designed to elicit a specific morphosyntactic property of the language. For example, we compare the translation of *A tree fell* and *Two trees fell* to see if verbs agree with subjects in number.

Our approach relies on the consultant getting the point of each example, with minimal use of linguistic terminology (see below). But this approach can easily fail to produce data that is useful for training an MT system. For example, the consultant may speak a language that does not normally use articles, but may feel compelled to translate the English words *the* and *a*, resulting in a corpus and that translation may not accurately reflect the normal syntax of his or her language.

As part of a U.S. government project called REFLEX, we produced an elicitation corpus of 3124 English sentences, which the Linguistic Data Consortium (LDC) is translating into a number of languages, beginning with Thai and Bengali.

This paper is concerned with an evaluation of our elicitation corpus. Two types of evaluation are provided. First, we provide an error analysis of two human translations of the elicitation corpus. Second, we compare an Example Based MT (EBMT) system trained on original human-produced translations and on

srcsent: We baked cookies.
context: We = 5 men;

((actor ((np-function fn-actor) (np-general-type pronoun-type)(np-person person-first)
(np-identifiability identifiable) (np-pronoun-exclusivity inclusivity-neutral)
np-number num-pl) (np-biological-gender bio-gender-male)(np-animacy anim-human)
(np-specificity specific)(np-pronoun-antecedent antecedent-not-specified) (np-distance
distance-neutral)))
(undergoer ((np-function fn-undergoer)(np-person person-third)(np-identifiability unidentifiable)
(np-number num-pl)(np-specificity non-specific)(np-animacy anim-inanimate)
(np-biological-gender bio-gender-n/a)(np-general-type common-noun-type)
(np-pronoun-exclusivity inclusivity-n/a)(np-pronoun-antecedent antecedent-n/a)
(np-distance distance-neutral)))
(c-polarity polarity-positive) (c-v-absolute-tense past) (c-v-lexical-aspect activity-accomplishment)(c-general-type declarative-clause)(c-my-causer-intentionality intentionality-n/a)(c-comparison-type comparison-n/a)…

Figure 1: A source language sentence, its context field and its abridged feature structure.

corrected translations in order to see the extent to which the errors of a linguistically naïve translator affect translation quality. We will conclude by discussing the implications of using linguistically naïve consultants as a resource for building MT systems.

2 Background

The AVENUE project has two related foci: building MT systems in low-resource scenarios, and making robust, hybrid MT systems using combinations of deep linguistic knowledge and statistical techniques. The hybrid system is a *statistical transfer* system (Lavie et al. 2004), which makes use of transfer rules as well as a statistical decoder. The rules can be written by hand, or learned automatically (Probst 2005). The AVENUE system also includes an EBMT system (Brown 1996), in order to use any pre-existing parallel texts that do happen to be available.

One hypothesis of the AVENUE work for low-resource scenarios is that MT systems can be learned from small amounts of data if the data is highly structured (Lavie et al. 2003). The elicitation corpus is therefore designed to produce highly structured data. Each sentence is designed to elicit a specific morphosyntactic property of the language, and sentences are organized into minimal pairs (e.g., *A tree is falling* and *A tree fell*) to compare the effects of changing one grammatical feature at a time. Probst (2005) describes automatic rule learning from elicited data.

A small sample of elicitation sentences is included in the list below. A more detailed description of the elicitation corpus can be found in Alvarez et al, (2006).

- Mary is writing a book for John.
- Who let him eat the sandwich?
- Who had the machine crush the car?
- They did not make the policeman run.
- Our brothers did not destroy files.
- He said that there is not a manual.
- The teacher who wrote a textbook left.
- The policeman chased the man who was a thief.
- Mary began to work.

Each sentence in the elicitation corpus is associated with a set of feature-value pairs, which represent the meaning elements that may be reflected in the morphosyntax of the language. Figure 1 shows an example of an elicitation sentence and its feature structure.

As mentioned above, the elicitation corpus was translated into Thai and Bengali. The structural differences between Thai and Bengali make them excellent choices for our first elicitation corpus assessment. Bengali is a synthetic Indo-European language spoken in India and Bangladesh. It has rich system of tense and aspect. Thai is a highly analytic language with a complex pragmatic system and gender marking. It is the national language of Thailand and is a member of the Tai-Kadai language family.

In our analysis of corpus translations, we found 1064 elicitation errors in the Thai Corpus and 359 in the Bengali corpus. An elicitation error is any translation mistake that would lead to an incorrect characterization of a language. A discussion of these types of mistakes can be found in section 4.

We also wanted to see to what degree these translation errors in the corpus would harm an MT system learned from the data. For a variety of reasons, it was not practical to train our statistical transfer system on this data. We therefore assessed the impact of these elicitation errors by training two EBMT systems on our Thai data. One trained on our original unsupervised corpus and the other trained on a corpus corrected of elicitation errors. This evaluation is described in section 6.

3 Related Work

Two other projects that we know of formulate grammars based on elicited data. In addition to the Boas system mentioned above, which attempts to train naïve informants to provide linguistic information, the Grammar Matrix (Bender and Flickinger, 2005) collects facts like the existence of subject-verb agreement from a field worker and then automatically produces an HPSG grammar for the language. Both of these use knowledge that a trained human has put into technical linguistic form. In contrast, our approach analyzes translations of elicitation corpus sentences, and the underlying feature structures they represent, to derive the linguistic facts about the language automatically.

3 The Corpus and Support Materials

Our elicitation corpus is a monolingual corpus of 3124 English sentences. We designed it to be translated into any human language. Each sentence in the untranslated corpus is made of three main components. First, we start with a feature structure that represents the elements of meaning that will be in the elicitation sentence. This structure has separate fields each representing head-bearing phrases. Each field contains a list of features and values that represent the pieces of meaning underlying the source language sentence. By features we mean morphosyntactic phenomena, for example, person, number or tense (Alvarez et al 2006).

Next, we annotated each feature structure with an English sentence that would represent the features and values in its underlying structure. Because our feature structures are intended to cover the majority of morphosyntactic features that exist in human language, our English sentence may not adequately represent all of the features in the feature structure. For example, given the sentence “We baked cookies”, some languages would translate it differently based on whether the actor was dual, plural, male or female.

If a linguist were to administer this corpus it would be possible for the language consultant to ask clarification questions. However, for the REFLEX project, the LDC administered the translation of our corpus with a single translator per language and with no supervision from our team. We had no contact with the translators during translation of the elicitation corpus and were not present to answer questions. To clear up confusion about how we wanted the corpus sentences to be translated we used "context fields". The context field supplements our English elicitation sentences with information not easily represented in the English sentence itself, but represented in the feature structure.

Our feature structures by themselves are complicated and would be difficult for someone without linguistic training to understand. However, a context field and a source sentence together embody all of the information in their corresponding feature structure. Thus, we were able to hide the feature structure and give the translators just the elicitation sentence and context.

2a.
Sentence: You wrote.
Context: You = five men
Translation: antum katabtum

2b.
Sentence: You wrote.
Context: You = two men
Translation: antumaa katabtumaa

2c.
Sentence: You wrote.
Context: You = five men
Translation: escribieron

2d.
Sentence: You wrote.
Context: You = two men
Translation: escribieron

Figure 2: Context information isn't always incorporated into target language translations. The two sentences translated into Modern Standard Arabic (2a and 2b) are translated differently based on the number of people 'You' represents. However, the Spanish translations remain the same in 2c and 2d. This example and further ones can be found in our translator guide (Alvarez et al. 2007).

For further clarification, we wrote a translator guide with examples and explanations to steer the native speakers toward translations that would reveal the language features of the target language.

When we talk about revealing language features, we mean the morphosyntactic characterization of a language. That is, we want to be able to learn how language features are grammaticalized in a target language or if they are manifested at all. In our case, we strove to get the most natural sounding translation that would let us learn about the features of a language. This means that not every feature will be translated into our target elicitation language. This is an acceptable outcome as it is just as important to know what features are *not* grammaticalized in a language as those that are. For example, a Spanish speaker would translate the plural second person pronoun the same whether 'you' represented 2 or 5 people. However, in Modern Standard Arabic the two sentences would translate differently depending on whether the pronoun represented 2 or 5 people. Thus, the context field may play into the translation of one language, but not into another. Because we designed our corpus to be used with any language a translator may be faced with, context fields will contain information that that may or may not be able to be utilized by the language consultant. One of the tasks of our translator guide was to help the translator learn where to draw this line. The next section will examine the extent to which the guide achieved this goal and the extent to which we were able to acquire successful translations.

4 Elicitation Corpus Translation Assessment

We assessed our translations using methods similar to those used by field linguists (Longacre 1964). That is, we analyzed sentences by comparing them to one another in order to pick out translation patterns. However, the consequences of unsupervised translation cut both ways for us. Thus, while the translator was unable to get clarification directly from us, we were unable to get clarification directly from the translator. A linguist in the field would be able to ask the language

Thai Elicitation Errors		
Source Sentence Over-Translation	845	79.41%
Context Over-Translation	57	5.35%
Under-translation	88	8.48%
Mistranslation	68	6.39%
Grammar Mistakes	6	0.19%
Total	1064	100%

Bengali Elicitation Errors		
Source Sentence Over-Translation	0	0.0%
Context Over-Translation	24	6.68%
Under-translation	5	1.39%
Mistranslation	76	21.17%
Grammar and Spelling Mistakes	254	70.75%
Total	359	100%

Figure 3: Total elicitation errors for the Thai and Bengali translations of the elicitation corpus.

consultant about the meaning of individual words and morphemes, but without this resource we were forced to compensate with dictionaries, grammars and language learning materials in order to confirm correct translations. In cases where we were unable to account for every in a sentence we consulted with local native speakers to assess the meaning of unknown phenomena.

Based on this analysis, we were able to assess all of our Thai and Bengali translations and keep track of elicitation errors. By our standards, most sentences were translated in a way that would make them useful as a resource for learning about a target language. However, some sentences contained constructions that diminished the utility of the translation and would provide spurious information about the grammaticalization of the target language. Below you will find a classification of these errors and their consequences. For full results of these error types for Bengali and Thai see the tables in figure 3.

4.1 Context Over-translation

The elicitation corpus's context fields are designed to provide additional information that may or may not be used as clarification when translating a sentence. Referring back to figure 2, the distinction between dual and plural pronouns causes a difference in translation for the Arabic translation, but not for the Spanish. The information in the context field is not incorporated because the Spanish translations would be the same whether 'You' referred to two, five or a hundred people. The distinction between dual and plural pronouns is Spanish is not grammaticalized. However, if the translator is determined to use the information in the context field it is possible for them to translate the sentences into the Spanish equivalent of 'You two wrote' or 'You five wrote', or even 'You two men wrote' and 'You five men wrote'. While grammatical, the excess information does not clarify the translation, and furthermore, it adds information not found in the source sentence. Thus, if the over-translated source and target sentences were to be fed to a word alignment system or a statistical machine translation system we would see 'You wrote' aligned with the Spanish equivalent of 'You two wrote'. This increases the chance of generating incorrect translations and will reduce the quality of the translation system.

Furthermore, this error type can lead to translations that are awkward. The goal of our corpus is to elicit translations as they exist in their target language naturally.

An example of this elicitation error can be found in (a) in figure 4. The Bengali instance over-translates the distant past tense. In Bengali, the simple past

a. Context Over-translation

Bengali target: বিজয়া কয়েক সপ্তাহ আগে বংকীমকে বইগুলি দিচ্ছিল.

transliteraton:	BAiJAYYAaa	KAYYAeKA	SAPAVIRTAaaHA	AAGAe
	BAANUKAiiMAKAe	BAIGAuLAi	DAiCAVIRCHAiLA.	
gloss:	Bijoya	a-few	moment-plural	before
	Bankim-acc	books-plural	give/third-person/progressive	

source: Bijoya was giving Bankim books.
context: Translate this sentence as if the incident it refers to happened minutes ago.

b. Source Sentence Over-translation

Thai target: ผู้ชาย คน นั้น ได้ มี ความสุข

transliteration:	pôo chaai	kon	nán	mee	kwaam sòok
gloss:	man	person	that	is	happy

srcsent: The man was happy.
context:

c. Under-translation

Thai target: ผู้ชาย คน นั้น จะ ตำหนิ เด็กผู้หญิง คน นั้น

Transliteration:	pôo chaai	kon	nán	jà	dtam-nì dèk	pôo ying	kon	nán
gloss:	man	person	that	will	reprimand	girl	person	that

srcsent: The man will criticize the girl.
context: Translate this as if the speaker heard this information from a rumor.

d. Mistranslation

Thai target: รั้ว รอบ ทุ่งหญ้า พังทลาย ลง

Transliteration:	rúa	rôp	tôong yâa	pang tá-laai	long
gloss:	fence	around	pasture	fall	down

srcsent: The fence around the pasture collapsed.
context:

e. Spelling and Grammar Mistakes

Bengali target: মহিলাটি যে গুদামে নয় কথা বলিতেছে.

Transliteration:	MAHiLaaTTi	Ye	GAuDAaaMAe	NAYYA	KATHAaa
	BALAiTAeCHAe.				
gloss:	woman-def	what	store	negative	statement
	talk/third-person/progressive				

srcsent: The woman who is not in the store is talking.
context:

Figure 4: This figure catalogs examples of our five types of elicitation errors. They are discussed in the text.

tense of an action remains the same whether it occurred seconds, days or years ago. The Bengali translation for sentence (a) now means 'Bijoya was giving Bankim books a few moments before.' if translated back into English. This translation does not match the meaning of the source sentence or its feature structure.

4.2 Source Sentence Over-translation

Source sentence over-translations occur when the translator over-specifies the translation in order to match the source sentence at the sacrifice of fluency or natural sounding translations. For example, in example b. found in figure 4 the Thai translator attempted to add definiteness to his/her translation by including the Thai demonstrative 'nán', which translates as 'that' in English.

There are two problems that arise

with this elicitation error. First, Thai doesn't mark definiteness explicitly, and certainly not with a demonstrative word. Secondly, the source and target language sentences have slightly different meanings. The original source sentence is '*The* man was happy,' but the translation means '*That* man was happy'. A more appropriate translation would have been 'pôo chaai kon mee kwaam sòok' or 'Man is happy'. While the ideal translation leaves the definiteness as ambiguous, it gives us a natural, reasonable translation, and, more importantly, gives us information about what features in the source sentence remain unmarked in the translation sentence.

Source sentence over-translation differs from context over-translation in one key way. In the case of source over-translation there is no information included in the target sentence that is not found in the source sentence. However, with context over-translation the target sentence includes information found in the source sentence that should remain unspecified in the translation. So, source sentence over-translations include too many features from the source and context over-translation includes too many from the context.

For the Thai elicitation corpus, source sentence over-translation was the most prevalent elicitation error found, but it is relatively rare in the Bengali corpus. This can be explained by how closely each language is related to English. Like English, Bengali is an Indo-European language. In addition it marks definiteness and number just as English does. However, Thai leaves both of these features unmarked morphosyntactically. In fact, out of the 845 Thai over-translation errors over 578 were made over specifying definiteness, identical mistakes that were repeated over and over again. This feature couldn't be over-translated in Bengali because it is marked morposyntactically just as in English. This explains the total of zero source sentence over-translations for Bengali.

4.3 Under-translation

Under-translation occurs when information from the context or source sentence is not translated into the target sentence. Thus, under-translation is an elicitation error caused by leaving something out. For example, substituting the word for 'person' for that of 'woman' or 'man' eliminates the feature of gender that would otherwise be evident in a sentence.

However, most under-translations are not that obvious. Under-translations can be difficult to find compared to over-translation. In our case, we discovered over-translations just by glossing sentences and double-checking those we discovered with a native speaker. In addition, we relied on language grammars and language typology charts (comparative tables indicating the morphosyntactic characteristics of many languages) to help discover this error.

The only under-translations we found were related to source marking. According to Iwasaki and Ingkaphirom (2005), evidentiality is marked in Thai analytically, especially in cases of hearsay. Our Thai translator, however, made no distinction between sentences describing events directly observed by the speaker and those heard from a rumor or gathered from evidence. Each sentence is translated grammatically, but omitting a key word that would give us insight into the categorization of information sources.

This elicitation error is rare, but having translators look at sentences within a narrative might mitigate this error, especially with regard to evidentiality.

4.4 Mistranslation

Mistranslations occur when the target sentence means something different from the source sentence. This means that the feature structure representing the meaning of the first sentence would be different than that of the target sentence feature structure.

For example, one of the most common mistranslations involves mistaking the aspect represented by the source sentence. For example, a habitual source sentence might be translated as present progressive. Another example would be the Thai translation (d) in figure 4. A past tense English sentence was translated as a present tense Thai sentence. Thus the Thai translation would be translated back into English as 'The fence around the pasture collapses.' There is a natural, fluent way to translate the Thai sentence in the past tense, thus it is likely that the translator made a mistake and translated using the wrong tense.

One reason for the occurrence of this error might be that some of our English source sentences appear to be too ambiguous or have overly subtle distinctions. This might leave the translator to interpret the sentence to the best of his/her abilities and that interpretation might not match up with what we expect to elicit. Compounding this is the fact that some of our sentences are awkward, unclear or absent of a narrative. Of course, some of this may be attributed to human error. Out of several thousand sentences some mistakes can be expected.

4.5 Spelling and Grammar Mistakes

This elicitation error covers the spelling mistakes and grammar mistakes that happen within the corpus. Also included in this category are sentences that are faithful translations, but are ungrammatical in the target language. A certain degree of human error can be expected; the frequency of this type of mistake will depend on the education level of the translator.

However, large numbers of these elicitation errors could point to larger difficulties with translations. A portion of our Bengali elicitation corpus contains a number of recurring mistakes that are unlikely to have been made by a native speaker.

For example, the Bengali sentence (e) in figure 4 is an ungrammatical way to represent a relative clause in Bengali. In reality this sentence would have to be translated with two separate clauses which can be taken to mean the following as an English equivalent: 'The woman who is angry, she is talking'. It is possible that the translator was trying too hard to stick to the structure of the English translation, but the Bengali sentence as it stands is not correct Bengali in any dialect.

Further mistakes were made with regard to using inanimate markers on animate noun phrases and the use of classical Bengali in inappropriate contexts. The common Bengali name 'Bankim' was even spelled incorrectly for a portion of the corpus. Both of our native speaker consultants agreed that translations involving these mistakes were unlikely to have been made by a native speaker.

These mistakes were the most popular for the Bengali corpus and accounted for 254 total errors, or 70.75%. In comparison, the Thai corpus only contained a total of 6 spelling and grammar mistakes.

5 Suggestions for Improving the Elicitation Error Rate

The cause of these elicitation errors could come from three places.

First, our documentation may not be clear enough. It could be lacking in examples or be lacking in clarity. We were hindered because we were forced to use translation examples from an assortment of languages, none of which are the language of the translators, to illustrate our arguments. However, the translators seemed to have understood the documentation and followed its directions. They made few mistakes with regard to the context field and only over interpreted it in 57 out of 3124 sentences for Thai and 24 out of the same number for Bengali. Even the error of source over-translation, while widespread, did not occur 100% of the time in places where it could have appeared. For Thai, it seems that our Thai translator was torn between delivering natural translations and delivering ones that conformed as closely as possible to the English source sentence. In light of this, we will be adding further examples to the documentation to clarify this, the most prevalent translation error.

Secondly, it is possible that some of the elicitation corpus sentences are unwieldy and difficult to translate. Magnifying this awkwardness is the fact that our sentences are without discourse context. That is, the sentences might benefit from appearing as part of a larger narrative or a story. Other sentences, such as those exploring locative features might benefit from pictures or other visual aids to clarify the meaning of each locative construction. Field linguists often use pictures or stories to clarify their elicitation sentences, so it might be of benefit to us to do the same.

Lastly, it is possible that our corpus is *too* unsupervised. A short period of training for the translators would be a way to catch and correct common types of elicitation errors. Though the point of this corpus is to perform unsupervised elicitation, it could be beneficial to administer a short pre-test with detailed feedback. This strategy could be a way to catch the most common elicitation mistakes. Our most common elicitation errors were really one mistake repeated many times. As we said in section 4.2, our Thai translator over-translated definiteness 578 times. Eliminating just this mistake reduces the elicitation error by 68.4%. Caught early, these easily correctable mistakes could dramatically improve our chances of getting the translations we desire.

6 Elicitation Errors and Machine Translation

To further assess the impact of elicitation errors found within unsupervised elicitation corpora, we trained two EBMT systems (Brown, 1996) to compare the results between one trained on our unsupervised data and one trained on the same data cleaned of elicitation errors. This corrected corpus will represent an ideal corpus translated under the supervision of a linguist.

Of the two corpora available, we chose to work with Thai rather than Bengali. This is because the errors for the Bengali corpus were too extensive to be corrected by a non-native speaker. Additionally, the errors in the Thai corpus were repetitive and less resource intensive to correct. Furthermore, the lack of morphology and the stable orthography made Thai the clear choice for a machine translation system trained on such a small corpus without segmentation.

We translated from Thai to English. The system trained only on about 2900 sentences from our elicitation corpus. The training sets used by our two EBMT systems used corresponding sentences for training data. This means that if a specific sentence from the uncorrected corpus were to be added to the training set, its corrected counterpart would be added to the set of training data for our corrected elicitation corpus.

Of the remaining 200 sentences, 100 were using for tuning the systems and 100 were used for testing. The test sentences in both cases were from the corrected corpus, since we want to test against gold standard translations. We also used a pre-trained English language model to aid in output generation.

Our results are displayed in the table below:

EBMT BLEU Results	
Uncorrected Thai	0.499
Corrected Thai	0.552

There is a 9.6% difference between the scores of the two systems. The Bleu scores are high due to the short sentences in our test set and the redundancy throughout our corpus.

Because we trained and tested only on the source and target sentences without their contexts there will be a number of sentences with duplicates in the corpus. Sentences that are found both in the training and target sets are assured perfect matches from the EBMT system and contributed to the high Bleu scores.

However, we are more interested in the difference between the two scores than in the performance of the systems themselves. The 9.6% difference is significant, but the uncorrected data system was still in a comparable range with the one trained on corrected data.

7 Conclusion

While there were numerous elicitation errors occurring with both the Thai and Bengali elicitation corpora, these errors were not so serious that they would render sentences useless for learning about a language, especially for human analyzers.

Elicitation errors also significantly affected the performance of the EBMT system. However, despite this, the Bleu score declined by less than 10%, providing some evidence that the uncorrected translations would still be able to train a usable system.

We will conduct further experiments to gauge the effect of elicitation errors on larger sets of training data. We will also investigate methods for recovering from noise in our training data, when it is not systematic.

8 Acknowledgements

We would like to thank our language consultants, Dipanjan Das, Satanjeev Bannerjee and Vorachai Tejapaibul. In addition, we would like to thank Aaron Phillips for his help training and testing our EBMT system.

9 References

Alvarez, Alison, Lori Levin, Robert Frederking. (2007) Elicitation Corpus Translator Guide. Technical Report. To appear.

Alvarez, Alison, Lori Levin, Robert Frederking, Simon Fung, Donna Gates and Jeff Good. (2006) “The MILE Corpus for Less Commonly Taught Languages”. *Proceedings of HLT- NAACL-2006*, New York.

Bender, Emily M. and Dan Flickinger. 2005. Rapid Prototyping of Scalable Grammars: Towards Modularity in Extensions to a Language-Independent Core. *Proceedings of IJCNLP-05 (Posters/Demos)*, Jeju Island, Korea.

Brown, Ralf D. (1996) "Example-Based Machine Translation in the Pangloss System". In *Proceedings of the 16th International Conference on Computational Linguistics* (COLING-96), Denmark.

Iwasaki, Shoichi, Preeya Ingkaphirom. (2005) *A Reference Grammar of Thai*. Cambridge University Press.

Lavie, Alon, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitjos, 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)*. June 2003. Pages 143-163.

Lavie, Alon, Shuly Wintner, Yaniv Eytani, Erik Peterson and Katharina Probst. "Rapid Prototyping of a Transfer-based Hebrew-to-English Machine Translation System". In Proceedings of TMI-2004, Baltimore, MD.

Longacre, Robert. (1964) *Grammar Discovery Procedures.* Mouton & Company, the Hague.

McShane, Marjorie, Sergei Nirenburg. (2003) Parameterizing and Eliciting “Text Elements across Languages for Use in Natural Language Processing Systems”. *Machine Translation 18(2):* 129-165

McShane, Marjorie, Sergei Nirenburg, James Cowie, and Ron Zacharski. (2002) “Embedding knowledge elicitation and MT systems within a single architecture.” *Machine Translation* 17(4).271–305.

Probst, Katharina. (2005) *Learning Transfer Rules for Machine Translation with Limited Data*. Ph D. Dissertation, Carnegie Mellon University.