MACHINE TRANSLATION IN SOUTH-EAST ASIAN MINORITY LANGUAGES

Stephen Beale
Thammasat University / Summer Institute of Linguistics
Language Research and Development Project
Thammasat University Bangkok, Thailand

1.0 Minority Language Translation Needs.

Why minority languages??? Most people recognize the validity of a machine translation (MT) system between two large language groups which share a large volume of textual resources (i.e. technical manuals, information services, scholastic research, etc), but how can one justify the large expenditure of resources necessary to establish a translation system that is aimed at a relatively small population with relatively small translation needs? Is a system like this practical?

No. In one sense it is not. It is certainly not practical in a business sense, at least not in the short term outlook. To implement a comprehensive system for an entire country, typically linking together four or five main language groups along with all the related dialects, would require many years of linguistic research. And the benefits are not completely obvious.

But the minority language issue is a complex one in any country. There are many reasons why an automatic translation system targeted at minority languages could be very useful - both for the national government and the local people. This paper will seek to outline some of the positive results a policy of translation for the minority languages might have, and to present two MT systems that might be used together to implement such a policy.

1.1 An Alternative to Integration.

An effective translation program between the national language and the various minority languages could potentially offer many advantages. In general terms, it could produce many of the benefits of linguistic "integration" or "absorption" without creating the unwanted side-effects involved. What

exactly is meant by "integration," and what are the benefits and side-effects associated with it?

Integration, in the context of this paper, would produce a situation in the country where every citizen used the national language with equal facility and expertise as any other citizen (discounting the normal variations in intelligence and schooling found in "normal" communities). Every citizen would have equal access; no citizen would be at a disadvantage because of the choice of national language.

Integration is occurring in many countries around the world. In some places it is an official policy vigorously enforced. In other areas it is a naturally occurring process brought on by socio-linguistic factors. Wherever it occurs, the ultimate result will be that the entire population of the country will speak the national language fluently (and often exclusively).

The benefits of such a situation are obvious. Communication and written resources could flow freely to all parts of the country. No group would be isolated from any other groups. The mistrust that is often present between the national culture and the minority groups would be greatly lessened as communication is facilitated and misunderstandings reduced. Because of these and other results, the national government would not have to allocate special resources to the control and aid of the minority groups.

The side-effects of such an integration are, however, substantial. From a cultural and anthropological point of view, the suppression of the minority languages along with their distinctive cultures would be a great loss to the country. It would be naive to believe that integration could be achieved without the eventual collapse of the minority cultures. Perhaps a more pressing problem would be the unrest and revolt a policy of forced integration would bring. Minority groups (indeed any group, large or small) will not willingly accept the destruction of their language and culture. Many of the world's greatest conflicts (past and present) concern the relationships between national cultures and minority groups (note the recent problems in Yugoslavia and the Soviet Republics).

Of course, a system of gradual integration could be implemented, but such a system would need many decades to take hold. In the meantime, all the problems of a multi-lingual society must be dealt with. And whether gradual or rapid, the process of integrating a people group into the national language is, practically speaking, a very hard goal to achieve.

1.2 Some Benefits of Translation.

A multi-lingual translation system could, then, offer many of the benefits of integration without producing the drawbacks. Politically, communication between the government and the minority groups would be dramatically improved. As for the minority groups themselves, they would have much greater access to the resources of the national language. As these resources become available, many of the problems faced by these groups could be overcome. As a result, the status of the minority groups (in their own eyes as well as those of the country) would increase, making them more productive and important members of the country.

As noble and exciting as these results might seem, they might not even be the most important. The psychological effects on the population might eclipse them. A full scale development program of this sort by a national government would send the following message to groups all over the country: "we care about you, and we are willing to work at making you an important part of our country."

1.3 Some Useful Translatable Materials.

What types of translated materials would be useful for minority groups? The four areas of health, education, employment and politics summarize the important possibilities. Health problems often plague minority groups. Literature promoting and teaching good health practices would be invaluable. Information ranging from waste disposal to disease treatment could eliminate many of a group's problems. A book such as **Where There Is No Doctor** [Werner 1977] (various Asian editions are available) could be used.

Educational materials could be translated into the minority languages. Even if eventual integration was a goal, a translation system could help the process by providing a solid education at the lower levels. It is generally accepted that education in a national language will proceed much better when supported by a base of "mother-tongue" education. School textbooks and other appropriate reading materials for children could greatly enhance the education of minority groups.

Employment helps could also be a potentially fruitful source of translation materials. Included here would be information about farming methods, as well as literature related to other professions common in minority groups. Vocational training could also be of benefit as these groups move into greater contact with the national society.

Political information would help keep the minority group in touch with the rest of the country. Election materials and laws, along with a host of other materials in their own language, would greatly benefit minority groups. Translating a daily newspaper might be one of the most significant contributions. A newspaper is a source of all sorts of information, including all four of the areas described above.

Many other kinds of literature, from religious to pleasure reading, could be translated too. And, as technology grows, more sophisticated systems could be implemented. In 50 years (or 100?), systems that input and output speech may be developed. Accessing information from all over the world may become as common and as easy as going to the library.

2.0 Machine Translation Methods for Minority Languages.

Two approaches to machine translation are viable for reaching minority populations. To service languages that are closely related dialects of the national language (or are closely related to some other minority language for which translation has already been done) a dialect adaptation system (i.e. CADA) will probably be sufficient. This type of system transfers the surface structure features of the source language (including morphology and syntax) to the corresponding surface forms of the target language. Basically, this system will make all the needed changes to the vocabulary, and will rearrange syntactic and morphological structures as necessary.

For more distantly related languages (for example Thai and a Mon-Khmer or Karen language), those types of operations on the surface level alone will not be sufficient. Semantic analysis of the source text will need to be done, with the resulting semantic representations being transferred to the target languages. This approach is to be implemented in the TU-SIL Translation System.

2.1 Computer-Aided Dialect Adaptation (CADA).

The CADA approach, as its name implies, is useful for translating between related dialects. This raises an important question: how close do two languages have to be to be related"? Would Thai and Northern Khmer be considered "related"? This question will need to wait for further experience. Finding out how far one can push CADA and still obtain acceptable results will be a priority of the upcoming research.

2.1.1 A Simple Example.

How does CADA work? The following simplified example gives a good introduction to the major features of the program.

LANG A	LANG B	
man	mun	(phonology)
woman	girl	(lexical)
hit-s	hit-ing	(morphology)
S=Sub+Verb+Obj	S=Sub+Obj+Verb	(syntax)
man hit-s woman	mun girl hit-ing	

There are four types of changes that the CADA program handles: phonological, lexical, morphological and syntactical. In the example above, assume all of the /a/'s in Language A are pronounced as /u/ by the Language B people. These are phonological changes. Not all phonological changes require a change in the way a word is written, but sometimes they do. Complete lexical changes can also be handled. In the example, the B people always use the word "girl" while the A people always use "woman". Morphological changes can involve re-ordering of morphemes within a word or substitution of one morpheme for another. Both cases can be handled by CADA. Above, the "present-tense" morpheme "-s" is replaced by "-ing" in Language B. And finally consistent syntactic reordering can be easily handled by CADA. So, as in this example, if one language is Subject-Verb-Object while the other is Subject-Object-Verb, the necessary changes can be made.

There are many language situations where the types of changes shown above occur. For instance, the differences between Thai and Isaan (the Northeastern dialect of Standard Thai) are mainly lexical, with a few syntactic reorderings (ignoring the phonological differences which do not affect the spelling in most cases). It is expected that languages within the same language family will be good candidates for CADA. Even though the problems mentioned below might be encountered, they will not be so frequent nor will they affect understanding so much.

2.1.2 Problems With CADA.

CADA is very helpful for languages that are related. But for the most part, one has to understand the text in the source language before translating it into another language. The following illustrates this point using a very common example (a possible CADA

translation for each sentence is given in parentheses):

Greeting (English)	
A. Hello! How are you?	(sawatdii - pen yangay)
B. I'm fine. And you?	(dii - khun la)
A. Fine, thanks.	(dii - khawp khun)
Greeting (Thai)	
A. Pay nay maa.	(Where have you been?)
B. Pay thiaw.	(Out having fun.)
A. Kin khaaw ru yang.	(Have you eaten rice yet?)

The main point to understand here is that semantically, these two greetings are equivalent - they constitute what two friends might say to each other upon meeting on the street. It is obvious, however, that there would be no easy formula to translate from one form to the other. The actual words and phrases used to communicate "greeting" are not really related at all between the two languages. A CADA approach, which simply substitutes words and rearranges sentence patterns would not produce correct translations.

This greeting example is a very simple one, but it illustrates what happens over and over again in languages. In this case, an English to Thai translation using CADA would probably still be understood (though it would sound foreign). A Thai to English CADA translation of this greeting would probably not be understood correctly (it would be very confusing for English readers). In general, two languages that are not closely related will communicate the same semantic content in dissimilar ways. Sometimes a CADA translation will be sufficient. Other times it will seem foreign. Still other times it will be completely misunderstood, communicating something entirely different than was intended. A translation system that first attempts to "understand" the source text and then translate meaning (instead of surface structures) will perform much better for languages that are not "CADA-able".

2.2 The TU-SIL Machine Translation System.

The TU-SIL Machine Translation System is an experimental system being developed to aid in the translation of materials for minority languages. It seeks to overcome the deficiencies of the CADA program by analyzing the meaning of the source text, and then transferring the meaning, not the surface structures, to the target language. In addition, it is called a "system" because, in reality, it is a complete

translation environment designed around the complexities involved in the minority language situation. Specifically, the system provides a method for analyzing, inputting and changing language rules, and for editing translated output. These functions and methods will be described briefly below.

2.2.1 Translation Methods.

The heart of the TU-SIL System consists of a semantic analyzer and a target language generator. It produces an intermediate semantic representation of the source text that can be used by text generators for multiple languages. This process is illustrated in the following example:

Text: Although John was fired because he was lazy, his wife still loved him.

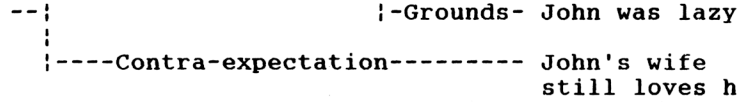
Conceptual Analysis:
John: Name of person, male, etc.
wife: kinship based on marriage - wife
lazy: moral/ethical quality, laziness/idleness
fired: takes agent, object
reasons: neg. quality of object, etc.
love: attitudes/emotions - love/affection
takes agent, object
etc.

Propositional Analysis:
Each of the concepts analyzed above will be expected to combine with other concept types. The acceptable combinations will be used to separate the propositions. The relations between concepts in the proposition (i.e. agent, location, etc.) will be part of the propositional representation.

(1) Boss (agent) ¦ fired ¦ John (object)
("boss" supplied by conceptual analysis)
(2) John (topic) ¦ was lazy (attribute)
(3) John's (kinship) wife (agent) ¦ loves ¦ him (John-obj)

Discourse Analysis:
Certain of the lexical items (in combination with the existence of the correct propositional types) signal different relations between whole propositions. For example, "although...still" can signal a "Concession-Contra expectation" sequence. Similarly, "...because" signals an "Event-Grounds"

construction. The example text is analyzed as follows:



These semantic analyses can then be used to translate into the target language. Starting at the discourse level, a template of the target text is made that will put propositions in the right place connected with the correct target language lexical items to produce the desired discourse features. Filling in this template, then, will be the text generated from the propositional and conceptual analyses.

Of course, this discussion is very simplified, but it gives an overview of the translation process. The system is explained in more detail in [Beale 1991]. The theory behind it is a combination of the ideas in [Schank 1975] (Conceptual Dependency Theory), [Nida 1975] (componential analysis of words) and [Beekman 1981] (componential analysis of propositions and discourse).

2.2.2 System Features.

This system is drastically different from "normal" machine translation projects that work on large language pairs like "English to Japanese". In those projects, many hours are spent by native speaking linguists (on both sides) analyzing and writing grammars and dictionaries. Often these linguists have to be very familiar with the computer and the translation program, or have someone helping them who is. And the quality of translation put out by the system is expected to be very high.

A system for minority languages, on the other hand, cannot have such stringent requirements. In most cases, there will be no native speaking linguist. A huge investment of time preparing dictionaries and grammars will be unfeasible. It must be assumed that the linguists available to do the research will not be computer experts. With all these drawbacks, a high quality translation seems out of reach. The TU-SIL system seeks to alleviate these problems as much as possible. Features are added to help simplify, shorten and improve the language analysis and description time. In addition, a complete system for editing translated

outputs is included that will allow the editor to view the analysis and generation process, make changes at any points, and teach the system principles for use in further translation attempts.

TEACHING.

Picture the linguist out in the village. How does he/she start? For this system, it is expected that there will be an initial body of texts that will need to be translated. These texts will have been given to the computer and analyzed before the linguist goes out to the villages. Once there, the computer can then tell the linguist what it needs to know: "There is this construction, this logical formation, this concept used in this context." The linguist can then sit down, knowing what is initially needed, and teach the computer how to communicate these things in the language. In its essence, it is really the development system leading the linguist through the needed analysis, and then the linguist "teaching" the computer.

Think for a moment about the benefits of this. Does a 10,000 word dictionary need to be made? No. The concepts needed to translate the texts at hand are known. Only those concepts will need to be translated. The system can focus the linguist's attention on what is really important. In this way, the language learning and analysis time will be greatly shortened.

LEARNING.

The "learning" process is going to be directed by the reactions of the native speakers to the translated output: "no - that doesn't sound quite right", or "that word doesn't make sense there." After such remarks, the computer can draw on the source text and its analysis for the section in question, it can display the linguist's analysis of the needed structures and concepts, it can show the choices made in the translation process, and from all this, modifications to the analysis can be made on the spot. The changes can then be implemented right away, with the new translation displayed automatically, along with possible changes that may have occurred elsewhere. And, of course, the computer will remember what it has learned as it goes on with the translation process.

This will be, in addition to the method for solving analysis problems, the primary vehicle for adjusting the target language analysis when more source language texts are submitted for translation later on.

The argument for this type of approach is clear. The linguist is out in the village. He/she can't get the analysis right the first time. He/she won't have a computer programmer with them. The system needs to be able to go on from a basic start and learn from its mistakes. This is how children learn languages: when they come across something they don't understand - they ask. When they do something wrong - they are corrected. And they learn and modify their behavior accordingly.

EDITING.

This system realizes that it is not going to be perfect. And that at some point the linguist is going to get tired of fixing rules (or will be unable to). So the whole goal of the system is to produce output that is "edit-able".

What, then, would help an editor? What would help someone make changes to a computer's translation? As much information as can be given concerning the questions that arise. Maybe looking at the source text would be helpful. Maybe a dictionary look-up of some source text words. Maybe the semantic analysis of the source text. Maybe the different ways that type of semantic structure could be implemented according to the linguist's analysis. Maybe synonyms for some of the words. Maybe taking a look at the choices the computer made in the translation process. Maybe being able to change some of those choices and see what the results are. Maybe the simple ability to just ignore the computer's translation and input one of the editor's choosing. Anything that would help the editor create a polished translation. And, again, in a user-friendly way that will be helpful and not cumbersome.

Some of these features will only be practical for a trained linguist to use. But even twenty years down the road when the developing linguist is long gone, the system intends to present a helpful environment for editing the translation. Given the limitations of the minority language situation, this feature is one of the most important of the whole system.

3.0 Conclusion.

This paper has presented some of the attractive possibilities a nationwide program of translation for minority languages might provide. It has presented two options for implementing such a strategy: a dialect adaptation system (CADA) and a complete MT system for minority languages (the TU-SIL Translation System).