3

An Approach to Tamil - Malayalam Machine Translation

Dr. Radha Chellappan,

Professor & Head, Dept. of Tamil Studies,
Bharathidasan University, Tiruchirappalli, Tamilnadu, India

The term Machine translation covers two features. One is machine translation and the other is machine-aided translation. In the machine translation the source text is translated fully automatically though the output can be passed on to a translator for post-editing either fully or partially. In the machine aided translation system translation rely on the human intervention at various stages of translation. According to Lehrberger and Bourbeau (1988:45) different types of Computer aided translation or Machine translation (MAHT - Machine Aided Human Transla-tion, HAMT-Human Aided Machine translation, FAMT - Fully Automatic Machine Translation, MAT - Machine Assisted Translation) take as their principal criteria the degree of automation that is the relative contribution of the machine and the human translator to the translation process.

Machine translation continues to be a challenging task for the past several years. As far as Tamil Language is concerned, Tamil University in Thanjavur ventured in Russian -Tamil Translation during 80s. In this Tamil text in the Roman script is used in the form of transliteration instead of Tamil alphabets. It is a fact that translation of a text from any two languages of the same family will be comparatively easier when compared to languages of different families. Languages belonging to the same family will naturally have similar morphological and syntactical structure. The cultural aspects of the speakers of both the languages will be either similar or easy for the translator to understand. With this hypothesis, the translation of Malayalam Tamil translation is thought off. In this article several features for developing a software tool for the Tamil-Malayalam are discussed.

For this project software is being designed especially for functioning as part of translators work station. The software will be developed with the inference that

- Tamil and Malayalam Languages, belonging to the same family of Languages, have similar morphological, syntactical structures.
- The transfer rules to convert the morphological and syntactical structures will be much less and more accurate in Tamil and Malayalam than other European Languages like English, French, German etc.
- The literary dialect is aimed at for the purpose of translation. By literary dialect, the dialect normally used in the language of schools, better books and by the more educated people is meant. The literary variety chosen is the creative literature in the fields of Modern literature like Novel, Drama and short stories.

- The formation of transfer rules and the conditional rules will work well with this pair of languages which share common conceptual or semantic structures.

Functional Components

Machine readable source text is to be processed with the computerized tools. Different markers wherever necessary will be used to delimit word boundaries. The intention is to do away with the translation in Tamil and Malayalam Scripts instead of Roman Script. The available word processing editors in Tamil and Malayalam will be used for the purpose.

Contrastive analysis will be used for the source and target languages where the differences between the two languages are identified. Here the structures of the languages are linked up structurally as they came with the family relationships. Generally Contrastive Analysis is done for language teaching and this analysis can be very well extended to translation also. In language teaching and learning the learner is made to understand the structure of the language to be learned by analogy. In this process it also should be bourn in mind that the structural similarity in all the linguistic levels may not be reliable in the sense a particular grammatical stretch in one language may be a requirement but in another it may be one choice amongst several.

Machine translation programme can be done by using a translation editor and several sets of grammatical categories of the two languages.

- **Lexical database**

  The basic and important thing in Machine translation is the lexical data base of the languages. Here the dictionary of the two languages in consideration, Tamil and Malayalam is prepared. This will be a bilingual dictionary of root words. All the noun roots and verb roots are collected. This will contain synonyms also with contextual interpretations if any. Here the inflectional forms of the pronouns also find a place. So *நான், என், நீ, உன்* (nAn, en, nI, un) in Tamil and corresponding *ഞാൻ, എൻ, നീ, നിൻ* (njAn, en, nI, nin) in Malayalam. The word dictionary plays an important role and provides the word equivalent between Tamil and Malayalam. Parallel list of collocations will also find a place in the data base. During the process of translation if the system comes across with the new word which is not found in the data base, then will interact with the human agent to add the new word with the translation equivalent.

- **Suffix database**

  Inflectional suffixes, derivative suffixes, plural markers, tense markers, sariyai, case suffixes, relative participle markers, verbal participle markers etc will be compiled as parallel modules so that the programme can find equivalents for all the above categories.

Tamil	Malayalam	
ai	a	2nd Case Marker
otu	nre kuute	3rd Case Marker
ku	a	4th Case Marker
in/ vita	kaal	5th Case Marker
atu	nre	6th Case Marker

The finite verb in Tamil denotes personal endings and number markers. But in Malayalam it is not so. In the examples avan vanthan, aval vantaal the finite verbs show masculine singular and feminine singular respectively and the corresponding Malayalam words avan vannu and aval vannu where no separate markers for gender and number are used. So nu will be the equivalent of all the personal endings in Tamil. For translating the sentence avan vantaan in Malayalam the T2M system first searches in the lexical data to identify the equivalent word in Malayalam. So it translates it as avan in Malayalam Script. The next word is vantaan. It searches it into the lexical data base again and this will not find a place there. So it will be given to the morphological analyzer where the word is split into va+ nt+aan.

- Morphological Analyzer

  A Morphological analyzer is to be designed to analyze the constituents of the words. It will help to segment the words into stems, inflectional markers. Now in vantaan the stem va and tense marker nt are replaced by va and nt respectively and personal ending aan will be replaced by u in Malayalam. The part of speech of the word will be identified. The analyzer starts to work on the word from the end position of the word. It splits the word, if necessary, and proceeds to find the translation equivalent. In forming the morphological rules exceptions should be eliminated.

- Syntactic Analyzer

  The syntactic analyzer will find the syntactic category like Verbal Phrase, Noun Phrase, Participle Phrase etc. This will analyze the sentences in the source text. For eg. if the analyzer finds the sentence as the noun Phrase then it will be made to enter the noun phrase analyzer where the phrase will be split accordingly. It is also likely that the structural ambiguity will be a problem of translation. But the experts in the field of Machine translation are of the opinion that it is possible to ignore ambiguities with the hope that the same ambiguity will carry over in translation also. For eg, the sentence ராமன் தம்பி வீட்டுக்குப் போனான் (rAman tampi vITTukup pOnAn) there can be two interpretations. One is Raman went to his brother's house and another is Raman's brother went to his house. This structural ambiguity in the Tamil Sentence will be reflected in the translated text also.

Machine translation can be done in two ways. One is viewing the source text and the target text in the same screen in two different windows simultaneously. Another is overwriting the source text and thus by creating the translation. The T2M will be designed so as to view the source text and target text in the same screen but in two windows. The programme has to be so designed that the textual elements that could not be translated will be marked in different colour, say in red. So during the process of post editing, the human translator can easily identify the untranslated words and do the translation. The translation strategy adopted will be mainly through transfer rules for the translation of root word, stems, suffixes and the like. For the system to understand the selection of certain suffixes, conditional rules will also have to be framed.

Reference Books

1. Chellamuthu et al (1984) Tamil University Machine Translation System (TUMTS), Tamil University, Thanjavur

2 . Lehrberger, J. & L. Bourbeau (1988) Machine translation. Linguistic Characteristics of MT systems and General Methodology of Evaluation. John Benjamin

3. Newton, John (ed) (1992) Computers in Translation: A Practical Appraisal, London and New York: Routledge.

4. Nida, E.A (1964) Toward a Science of Translating . Leiden: E.J. Brill.

5. Sager, Juan C. (1990) Language Engineering and Translation, Consequence of Automation, John Benjamin.