# Russian to Tamil Machine Translation System at Tamil University

**Dr. KC Chellamuthu,** <muthu@netapp.com>
San Jose, CA, USA

## 1.0 Objective:

The computers as a tool to analyze, interpret, translate and disseminate information from one language to another is of tantamount importance in a multi-lingual and multi-cultural society like India. Besides being cheaper and faster, computer aided language tools like Machine Translation (MT) promotes the understanding of different languages and cultures. The main objective of this paper is to outline the role of Machine translation in information dissemination, a brief history of MT, MT strategies, various components and functions of an early MT system developed in Tamil University, Tanjore for Russian to Tamil translation. The paper also explores the complexities of Natural Languages(NL) with regard to MT and the possible solutions to resolve them using AI based techniques.

## 1.1 Introduction:

Information plays a vital role in the growth of science and technology. Similarly the growth and development of a language, its culture and literature depends very much on the dissemination of information. Hence acquiring, transferring, disseminating and diffusing information is crucial to the development of a human society. Many of the decision making processes and the research programmes are based on the most valuable information. Translation is one of the strategies that plays an important role in disseminating the information contents from one language to the other.

Although MT may not provide a complete solution to the problems of translation due to the unique and complex nature of natural languages, but it can be an efficient tool in translation of text, atleast in a restricted knowledge domain. In addition, MT can help the human translators to improve the speed and productivity of translation. The specified knowledge domain for MT lies mostly in the area of science and technology. The literature in science and technology can be classified as technical reports, laboratory procedures, equipment manuals and patent forms etc. Irrespective of the methodology adopted, a MT system should have the necessary components such as pre-processor, parser, lexical analyzer, verb phrase analyzer, translator and target language generator. This paper discusses a brief history of MT, MT strategies, the structure and functions of a MT system for Russian to Tamil and the complexities of a NL with respect to MT and the AI tools to resolve them.

## 1.2 A brief History of Machine Translation System

What is a Machine Translation System?. A translation is a process involving the transfer of the contents from one language called Source Language(SL) to the TL without altering the concepts or thought processes. A translation system either assisted by a computer or performed directly with the assistance of a human can be termed as Machine Translation (MT). A practical and useable MT system remained as an elusive goal for long due to the

inherent complexities involved in different Natural Languages (NL) and the limitation of hardware and software to tackle such a complex problem. But today with the availability of a powerful hardware and the varieties of software tools, modeling the complexities of syntactic and semantic aspects of a NL has become a reality.

MT has become a potential research area for researchers all over the world. However the researchers in MT are well aware of the success and failures of the MT research programme. Considering the history of MT, one would discern the fact that there are three different generations of MT as categorized by the factors as 'information acquirement', 'information dissemination', and 'information diffusion'. Today there are several MT systems in different forms available for various languages. These MT systems differ in their functional structure and the methodology of formulation taking into consideration the nature and complexities of languages involved in the process. These MT systems acquire significant practical importance due to the explosive growth and usage of internet in the areas of online business, research, education, communication and in the government. Some of these MT systems provide a faster and cheaper translation besides assisting the human translators to improve their productivity and efficiency in translation.

Referring to the history of MT research, a number of operational systems have been designed between various languages viz., English, French, German, Russian, Chinese and Vietnamese in European countries. In the early 1950's and 1960's, attempts were made to develop MT systems to translate texts in a restricted knowledge domain such as technical reports, laboratory procedures, equipment manuals, millitary records and weather reports etc. In 1950 Warren Weaver issued a memorandum proposing the use of computers for translation at the Massachusetts Institute of Technology. Later on in 1954, it was followed up by a bigger project for MT in Georgetown University in collaboration with IBM for Russian to English translation. In 1956 a French to Russian MT programme was demonstrated at Moscow by O.S Kulagina and I.A.Mel'chuk. Later Peter Toma developed a commercial system called SYSTRAN for Russian to English translation. This MT system was extensively used by US airforce translating Russian text to English and about 24,000 pages of Russian text has been translated in 1974 alone. In 1971 Y-Bar-Hillel at the university of Texas proposed certain new concepts for the progress of MT programme. In 1973 LOGOS development corporation released a English to Vietnamese MT system A chinese MT system called CULT (Chinese University Language Translator ) for translating Chinese newspapers into English was developed by Hong Kong university.

The French institute in France developed a MT system called TITUS for translating simplified sentences between German, Spanish, French and English. A weather report translation system called TAUM was initiated by the university of Montreal for translation from English to French in 1975. During 1972-73 a center for automatic translation for English to Russian was established in USSR and in 1976-77 the USSR financed a central office for translation to initiate a multi-lingual translation project. A MT group called GETA at Grenoble University has developed a MT system for Russian to French translation. It was a joint venture of scientists from USA, Canada, Japan and France. In May 1986 Japan has commercially released MT systems for Japanese into English and vice versa. Microelectronic giants such as Nippon, Fujitsu, Atlas, Sharp have developed these MT systems on a 32 bit microcomputers with a dictionary containing about 50,000 lexical entries. These MT system were used for translating laboratory records, equipment manuals and patent forms etc.

Reviewing the current activities in MT, the notable among them are MT research projects between several world languages at the Center for Machine Translation at the School of Computer Science at Carnegie Melon University. Various research projects at this center provides MT for text-to-text, speech to speech and text-to-speech and vice versa. The other centers are EAMT- European Association for MT, University of Maryland, Brigham Young university, University of Essex, IBM projects, a German-Danish system called METAL, an English-Russian-English system called PARS, and many other commercially available PC/unix/Linux based software for translating text in specific application domains, MT systems for translating web contents, email and URL etc., are also widely used today.

Currently notable MT and Natural Language Processing -NLP research activities in India exploiting the revolution of hardware and software advancements are, IIT, Kanpur, IIIT Hyderabad, Central Institute of Indian Languages in Mysore and Anna University in Madras.

### 1.3 Types of Translation Procedures

Translation plays an important role in information dissemination. The process of translation is not just a mere process of translating words but concepts, not syntagms but ideas. It is a thinking process and is not just matching of words nor a function of arithmetic or geometric progressions of a probable word combination occurrences. Translation can be briefly defined as follows: "It is an art of handling and manipulating the lexical power of linguistic elements of a language. It is an intellectual exercise involving the expertise of a language and its grammar. It involves comprehension, analysis and conceptualization of thoughts and of thought patterns in the source language and then expressing those grammatically and idiomatically in the target language". A good human translator should normally possess sensitivity with intelligence, creativity with good organization and inventiveness with good discipline. He should also have the ability to handle and manipulate the lexical power of linguistic elements of a language. It involves the expertise and the ability to tackle even the most difficult situations pertaining to syntax, semantics and styles etc., of a NL during the translation.
A translation procedure can be classified into the following categories:

1. Human Translation – HT
2. Machine Aided Human Translation – MAHT
3. Human Aided Machine Translation – HAMT
4. Machine Translation – MT

In the case of a human translation, the translation is achieved completely with the knowledge to handle the lexical power of linguistic elements of a language whereas in Machine Aided Human Translation-MAHT, the translation is aided by the computer providing a suitable word equivalents from the already stored glossary of words. In certain technical and scientific area conveying the basic meaning is considered to be sufficient through word-to-word translation by MAHT. HAMT involves the help of a human to pre-edit the SL text before being fed to the computer for translating into a target language. Due to the information explosion and the necessity of faster and cheaper translation, HAMT is very suitable for translating texts in science and technology, weather reports, equipment manuals, millitary records, laboratory procedures, articles and research monographs etc. Since human intervention is required in some form either pre-editing a SL text or feeding the input or specifying certain parsing/lexical rules, most of the currently available MT systems belong to HAMT. Completely automated translation called MT would perhaps become a reality one

day when the entire domain knowledge with complete and appropriate translation rules are modeled to resolve the syntactic and semantic ambiguities of a NL. The extensive research in the area of NLP and the various tools of AI would certainly lead the way to achive this goal in the near future atleast in the field of science and technology.

### 1.4 Machine Translation Strategies

Most MT systems currently developed are capable of translating scientific and technical documents. Translation of a literary text through MT involves more complexity as regard to syntax and semantics compared to technical documents. It is due to the fact that literary text normally uses a language which is entirely different from that of other knowledge domains. The literary language will have a place of expression for joy, emotions and sentiments with much of rhetoric and metaphors. Literary works normally involves a great deal of imagination and creativity with poetic talents and idiosyncrasies of an author. Hence this necessitates a scholarly talent to interpret the various literary intricacies of a literary language to produce a meaningful translation. However there are other areas where MT can play a vital role to disseminate the information at a much faster and cheaper rate than a manual translation.

Although the goal in a MT system is to transfer the meaning and contents of a SL into a TL in a faithful way, it differs based on the techniques employed. These strategies depend on the syntactic and semantic nature of SL and TL, methodology and the tools adopted in modeling the MT etc. Based on these factors, MT can be categorized into three major groups as follows:

1. Direct translation strategy.
2. Transfer Strategy
3. Interlingua Strategy.

The level of complexities of a MT system depends on the relative relationship in syntax levels and other linguistic aspects of source and target language(s). In a direct translation strategy a SL text is analyzed and is directly transferred to a TL through a series of stages of operations. It neither uses an intermediate language nor a parsing system. The output of this system depends on a codified dictionary and the pre-specified sentence patterns and also on the morphological analysis. The Georgetown MT system is one of the fine examples for a direct MT translation strategy. In the case of a transfer method the SL text is analyzed and transferred into an intermediate language called a meta-language with the help of a TL lexicon and then restructured before transforming the sentences according to the syntax of TL. A MT system developed by the group called GETA in Grenoble University, France falls under this category.

In the case of MT through interlingua strategy, an intermediate or universal language is used for the translation. In this method, Artificial Intelligence tools such as rules for knowledge representation schemes involving a high level structure and appropriate inference mechanism to resolve syntactic and semantic ambiguities and pragmatics are adopted. Interlingua method employs a universal language which is independent of a NL involved in the process. This intermediate language helps to resolve several problems of translation which otherwise could not be solved using the regular strategies. The various stages in an interlingua strategy are analyzing the text for conceptual representation, providing contextual world knowledge through inference mechanism, reproduction of the language free representation of the SL sentences into TL etc. This technique fully exploits the AI tools and the NLP analysis to achieve a meaningful translation.

## 1.5 Structure of a Russian to Tamil MT system

Although the NL is distinguished from a formal language with a help of a restricted grammar, certain characteristics of a NL plays a vital role in the syntax analysis and parsing during the process of MT. Though the translation is essentially creating a correspondence between SL and TL, conceptual understanding and analysis of SL sentences is essential for a meaningful translation.

In addition to several computer assisted language research projects emphasizing Tamil, MT was one of the main projects of Tamil University, Tanjore during 1980's. As a beginning a machine oriented translation involving Russian to Tamil was initiated during 1983-1984 under the leadership of the Vice-Chancellor Dr.V.I Subramaniam. It was taken up as an experimental project to study and compare Tamil with Russian in order to translate Russian scientific text into Tamil. Hence the goal was kept minimal and the scientific text belonging to a specific domain was used as SL input. A team consisting of a Linguist, a Russian language scholar and a computer scientist was formed to work on this project. During the preliminary survey, both Russian SL and Tamil were compared thoroughly for their style, syntax and the morphological level etc. The initial study helped to make the following inferences:

1. Russian is considered to be a highly inflectional language.

2. Word order is relatively free in both Tamil and Russian.

3. It is believed that the transfer rules which convert the surface syntactic structure of SL into the surface syntactic structure of the TL may be less when Russian is selected instead of other European languages.

The following are some of the objectives specified for the experimental programme:

1. To study and compare the syntax levels of Russian and Tamil.

2. To translate automatically scientific texts from Russian to Tamil.

3. To process the various attributes of Russian texts such as passive voice, present/past/past participles, cases, proper names, plurals, verb and noun phrases from the given stem, automatic pre-editing, handling of comparative and genitive phrases etc.

The Russian to Tamil MT system consists of various functional components such as a pre-processor, parser, lexical analyzer, bi-lingual dictionary, morphological analyzer, translation and generation modules. Depending upon the strategy adopted in a MT system, the functional organization may vary from system to system. In a MT system the primary task would be analyzing the input text, parsing the sentences, analyzing the words lexically and morphologically, conceptualizing the SL sentences, table look up using bi-lingual dictionary and translating the input word using the linguistic knowledge already defined in the system.

The translation strategy adopted in the Russian to Tamil MT system is a transfer methodology involving an intermediate language. The Russian to Tamil MT system uses an intermediate language with a syntax more related to TL. In this system the given Russian

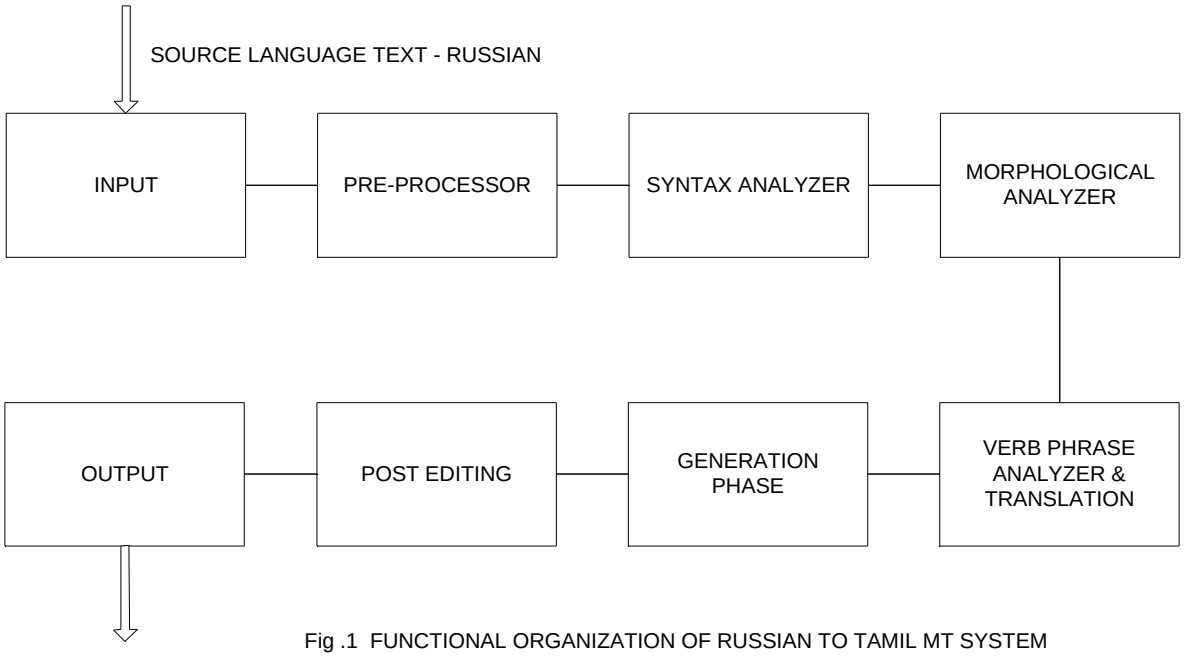


Fig .1 FUNCTIONAL ORGANIZATION OF RUSSIAN TO TAMIL MT SYSTEM

sentence is lexically analyzed and the syntax is transformed to the grammar of an intermediate language after carrying out the syntax and morphological analysis and word by word translation. The functional block diagram for Russian to Tamil translation system called- TUMTS is shown in Fig.1.

### 1.5.1 Pre-Processor

The main function of a pre-processor in the Russian to Tamil MT system is to perform syntactic analysis in order to verify the correct form of an input sentence. It also converts the complex sentences into simple forms. Apart from this, a pre-processor transforms the syntactic structure of an input text into a reference structure. The main task of a parser in the MT is to verify the syntactic structure of the input for its correctness in order to create a parsing structure. Thus the parser identifies grammatically correct and incorrect sentences by analyzing each one of the lexical elements. The lexical analyzer analyses each one of the lexical elements of a SL text by breaking the word into stem, inflection, and the markers etc. It provides the necessary information for the MT system to identify the particular lexical category of a word such as verb, noun, adjective, adverb, cases and tenses, gender etc.

### 1.5.2 Bi-Lingual Dictionary

A dictionary plays a vital role in a MT system and it provides the word equivalent between SL and TL. The bi-lingual dictionary between Russian and Tamil helps to provide equivalent words in Tamil for a Russian input text. This dictionary contained about 1200 vocabularies with certain lexical markers and attributes. The dictionary being a major database of a MT system, it must provide the equivalent meaning of all the probable words that would occur in a SL text. In this experiment, it was concluded that for a reasonable translation using MT, a bi-lingual dictionary must contain atleast a minimal lexical entry of about 50,000. Due to the memory constraints, it was possible to create only a small

dictionary. Apart from the main dictionary there are other tables of databases for easy access of words specifying lexical attributes and the marker equivalents etc.

### 1.5.3 Translation Phase

The translator of a Russian to Tamil system performs the task of word to word translation. The syntax and the morphological analyzer analyses the word constituents of Russian input sentence and splits the word into stem, inflection and lexical markers etc. The delimiters in the pre-processed text helps to identify the sentences and words in order to locate a particular word or a phrase. Input sentence is identified either as a verbal or non-verbal category. If it is a verbal category it is passed on to a verb phrase analyzer otherwise the translator with the help of a morphological analyzer gets the equivalent meaning in Tamil by referring the bi-lingual dictionary.

### 1.5.4 Verb Phrase Analyzer

The lexical analyzer with the help of a delimiter identifies the various grammatical attributes of the word elements in the input at the pre-editing stage. The pre-processor after identifying the correct syntax of the sentence, classifies the input sentence as verbal or nonverbal, the verb and also the adverb if any. The lexical details are derived from the markers attached to the verb phrase. The presence of adjectives, voice, plural and cases are identified by means of special markers.

### 1.5.5 Generation Phase

The translation phase after referring to the bi-lingual dictionary provides corresponding equivalents to the input Russian word. Similarly the equivalents for identifiers, inflections and case endings in the TL are also generated in this phase. The text translated in this process is in the syntax of an intermediate language and so they may not provide semantically acceptable phrases. Hence generation phase concatenates the string phrase and transforms the syntax into the syntax of Tamil.

## 1.6 Complexities of Natural Language Structures with regard to MT

The translation of a NL is not just matching of words but is rather a conceptual than a syntactical transfer. In order to design an efficient and usable MT system, it is imperative to analyze, interpret and to understand the complex syntactic and semantic aspects of a NL. The major problems encountered during MT process is regarding semantics rather than syntactic. It arises mostly due to the inadequate details of semantic representation and inefficient techniques adopted to represent the ambiguous situations and contextual variations. Due to these reasons the future and current challenges for MT research would be to adopt new techniques such as AI tools and knowledge base engineering to achieve more meaningful translations. The most complex NL problems as related to MT are syntactic ambiguity, Lexical and semantic ambiguities and Idiomatic expressions, pragmatics or language in context and anaphoric references.

### 1.6.1 Resolving Ambiguities and Pragmatics in MT

The ambiguity occurs in a NL either explicitly or implicitly. It is imperative that the ambiguity must be eliminated while creating a knowledge representation structure,

otherwise the interpretation will be meaningless or provides altogether a different meaning. Hence for a computer to understand a NL and to translate from SL to a specified TL, the ambiguity at the input must be eliminated.

#### 1.6.1.1 Syntactic or Grammatical Ambiguity

The ambiguities may be syntactic, semantic, lexical and pragmatic. Ambiguous situations arising out of the grammatical incorrection of a SL sentence is called as syntactic or grammatical ambiguity. Let us consider a SL sentence which is syntactically ambiguous and see as to how the AI techniques help to solve the MT problem.

1. Gopal to went school the

In the above sentence the word sequence is not in conformity with the syntax of English. A sentence is a structure which relates the words to each other in a proper sequence. If the sequence of words in a structure does not follow the rule, then the parsing system while parsing (ATN) the SL sentence will reject the sentences which are grammatically incorrect. Since the ATN parsing scheme focuses more on the concept and knowledge of the input sentence, the incorrect sequence of words conveying the wrong meaning is identified and rejected by the parser.

#### 1.6.1.2 Lexical and Semantic Ambiguity

The syntax of a NL sentence may not help to infer the meaning. In order to comprehend a sentence, it is necessary to understand the meaning of each one of the words in the sentence. Hence putting the words together in a right sequence form a structure that represents the meaning of the entire sentence. A process to identify the correct meaning of each one of the words is called as word sense disambiguation or a lexical disambiguation. The lexical or semantic ambiguity in a NL sentence arises when a single word has more than one word sense or meaning (homonym). The parsed structure provides a mapping between the syntactic structures and objects. When there is no mapping, then the structure will be rejected. Logical constraints of a text and the immediate sentence and the associated connections between the word sense and the context will help to solve the word sense ambiguity. One way of solving the semantic or lexical ambiguity is to associate with each word in the lexicon information regarding the contexts in which each of the word's senses may appear. Let us consider the following sentence.

2. They are flying Planes.

There are various interpretations possible for this sentence when tracing the different paths. It needs more of an expensive backtracking and duplicated processing for tracing different paths. Hence it is advisable to make a single plausible interpretation. In the above sentence 'they' should be understood as 'planes' not as 'pilots' in order to make the statement unambiguous. The lexical ambiguity could be solved during translation only when the machine is able to fix the most appropriate meaning of a word by context and the linguistic interrelation of other components of a sentence.

3. Leela hates cold

Here it may be inferred that the verb 'hate' can come only along with an animate subject and

hence the object 'cold' is understood not as soft drink but as a feeling of a person. The following are some of the English language sentences in which semantic ambiguities occur.

4. Gopal caught a cold and Raju caught a ball.
5. Time flies like an arrow.

### 1.6.2 Idiomatic Expressions

Translating an idiomatic expression from a SL directly to a TL will provide a meaningless translation. The syntactic structure, semantics and pragmatics and the referential relation between various word constituents may not solve the problem of translation, when there is an idiomatic expression in the input text. It necessitates to capture the semantic knowledge as conveyed by the sentence by adopting suitable scheme of knowledge representation. Hence to translate an idiomatic expression, it is necessary to recognize and interpret the entire concept of SL sentence and representing them with equivalent concept in TL.

6. The spirit is willing but the flesh is weak.

When we attempt to translate the above sentence directly into Russian the output would be:

The Vodka is good but the meat is rotten.
i.e The direct translation in this case becomes meaningless.

### 1.6.3 Pragmatics or Language in Context

Pragmatics is a problem concerned with the use of language in context. Pragmatics deals more with the interpretation of a language with respect to the context and usage rather than syntax or semantics. In the pragmatics the interpretation depends on the pronouns, definite references and the speakers intentions and responds to a highly elliptical or even illformed inputs. Apart from understanding the semantic knowledge as conveyed by the input structure based on the context and other references, proper interpretation must be made by relating the meanings of word senses with the adjacent and previous components of a sentence.

7. Gopu went to the supermarket. He picked up some provisions from the rack. He paid for them and left.

In the above statement, the referrent 'them' of the last sentence must be interpreted as 'provisions' as represented in the preceeding sentence with an assumption that one in the supermarket will have to pick up the provision. Thus a noun phrase reference is made here.

8. The passengers are near the terminal.

9. The software Engineer is near the terminal.

In the above sentence (8) and (9) the 'terminal' in the first sentence means it is an airplane terminal, whereas in the second sentence the word 'terminal' refers to a computer terminal. Thus the associative connections between the word sense and the context helps to make a proper interpretation.

### 1.6.4 Anaphoric Reference

It is one of the major difficulties in language understanding and so it will have to be resolved in SL text itself before attempting a translation. In a continuous text with more dialogues, the anaphoric references will be common. These references may be pertained to either of the participants of the dialogue. In order to solve this problem, a knowledge structure which could relate the different sentences and make a multiple sentence understanding is necessary. A pronoun in the SL may stand for many things, but at the same time there may be only a unique word in the TL and it can be resolved only when we are sure of the proper references made.

10. Gopal bought a scooter. Ramesh wanted to drive it.

The word 'it' in the second sentence must be identified as the 'scooter' and this kind of references are called anaphoric references.

11. Rama saw a blue ball in a shiny red wagon. He wanted it.

Here 'it' in the second sentence is ambiguous. We are not sure whether 'it' refers to the ball or to the wagon. If the sentence is specified as,

Rama saw a blue ball in a shiny red wagon. He wanted the ball.

Then the understanding would have been better without any ambiguity.

### 1.7 Conclusions

This paper discussed the concept and history of MT in the first part. The functions of various components of a Russian to Tamil MT system has been explained in the second part. The major problems normally encountered in the NL during MT has been listed. The NL understanding for MT involves solving of syntactic, semantic and pragmatic ambiguities and anaphoric references. The tools of AI and certain strategies have been suggested to resolve the syntactic, semantic and pragmatic ambiguities during MT with appropriate examples in the third part. It is concluded that the NL understanding would solve most of the major problems of MT and it in turn depends upon adopting a suitable AI technique.

### References

1. K.C. Chellamuthu et al. (Dec 1984) "Tamil University Machine Translation System (TUMTS)", Thanjavur: Tamil University, India.

2. Walter Sedelow A. and Sally Yeates Sedelow. (1979), "Trends in Linguistics Studies and Monographs 5" – Computers in Language Research, Germany: Mouton Publisher.