International Journal of Information Movement *Vol.I Issue III (July 2016)*
ISSN: 2456-0553 (online) Pages 18-29

MACHINE AIDED LANGUAGE TRANSLATION TOOL FOR THE RURAL STUDENTS

M. Rajasekar
Research Scholar
Hindustan University, Chennai
Email: sekarca07@gmail.com
Dr. N. Rajasekharan Nair
Professor
Hindustan University, Chennai
Email: nrajasekharan@hindustanuniv.ac.in
Dr. A. Udhayakumar
Professor, Hindustan University, Chennai
Email: acoe@hindustanuniv.ac.in

Abstract:

The technology development makes all the human activities as more fast and systematic. We have to compete this development by getting familiar new technological things to our daily life. Likewise we are going to introduce a new method into our school studies. For this we have developed a new Machine Aided Language Translation Tool for the Rural Students, to help the students to translate their ideas, thought and learned things from the mother language (Tamil) to the second language (English). We have developed a Tamil to English translation tool, which translate a word or a sentence into second language (English). We are using the statistical – dictionary based machine translation method to translate.

Keywords: Artificial Intelligence, Machine Translation, Parts of Speech Tagging, Morphological Analysis.

1.0 INTRODUCTION

Technology is a basic thing to live in our present world. Mainly the knowledge of utilization of technology is most important. In all the areas of our life the technology involved to show the efficiency of production. Now consider our school education system is based on the text book oriented. The teacher must have sufficient knowledge on the subject. Actually education is needed to get knowledge and to build new thoughts by using their gathered knowledge on a particular subject. But the aim of our school education system is to prepare the students to get high scores in the examinations. Basically the marks are important. But the way of teaching and learning is to get some difference to get high scores and to get knowledge to promote the student's to achieve new things. For that the basic learning method should be changed in our schools.

Consider our rural school students; they are very clever in all subjects. But most of the students are getting fear on English subject. We could not find where is the fault, because the students are learning English as a second language, and the teachers are also teaching the English subject by the mother tongue only. They can be able to explain the content of the English subject by their mother tongue of students. Then only the students can understand the content of the subject. But that method also should be changed by using technology. Let us consider in a class the teacher will teach the English subject to the student in their own mother tongue. He/ She will explain the content of the lesson in their mother tongue. All the students have to get understand what the teacher is teaching. If the students got the idea of that lesson by their mother tongue, then they can express learned lesson by their mother tongue only. But in the cycle test and the public exam in the English subject the students have to write in English only.

For this reason, if the students got the idea about that learned lesson in their own mother tongue, they have to mockup whole lesson in English and write (vomit) in the test or examinations. By this mockup the students can be

able to write the learned content easily, but they could not memorize for long time, they can forgot after the exam. There is no usefulness of this type of teaching learning method.
A Successful student has to know a local meaning of every single word in the lesson. Then he should know how to express that learned lesson into second language without mistake. This is not an easy work for every student.
Our work is to aid these students to learn and express their learned lessons in second language.

2.0 OBJECTIVES

The objective to development of machine aided language translation tool for the rural students is to attain the good vocabulary in second language (English) learning, and to get familiar in generating the sentences in second language. The objectives are as follows:

- To computerize the concept of translation of Tamil - English
- To include most of the words in Tamil with equivalents in English
- To develop a user friendly application can be used by all type of users
- To develop sub tools
  - ✓ Morphological Processor
  - ✓ POS Tagger for Tamil
  - ✓ Dictionary for Classical Tamil - English
- To develop a larger corpus for Tamil – English Machine translation
- To be implemented in Hand held devices also (Laptop, Mobile, Tab)

2.1. A Overview of Artificial Intelligence

AI is a branch of Science which deals with helping machines find solutions to complex problems in a more human-like fashion.

2.2. Applications of AI

- Expert Systems
- Natural Language Processing (NLP)
- Speech recognition
- Computer vision
- Robotics

2.3. Natural Language Processing

Natural Language Processing is concerned with the communication/ interaction between human and computers in a natural (human) languages such as, English, Hindi, Tamil, Malayalam, etc., Figure 1.1. shows that as it is an inter-disciplinary subject.

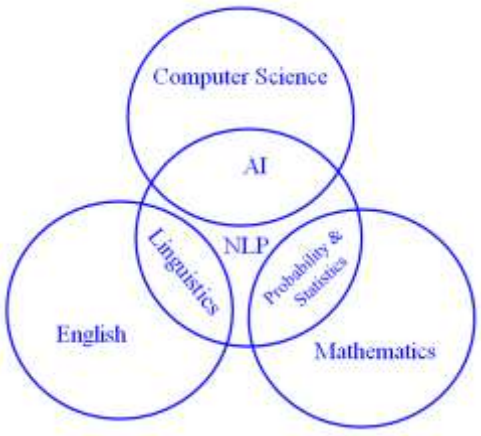


Figure 1.1. NLP

3. MACHINE TRANSLATION IN NLP

Natural Language Processing (NLP) is the field of computer science devoted to the development of models and technologies enabling computers to use human languages both as input and output. The ultimate goal of NLP is to build computational models that equal human performance in the task of reading, writing, learning, speaking and understanding. Computational models are useful to explore the nature of linguistic communication as well as for enabling effective human-machine interaction. Jurafsky and Martin (2005) describe Natural Language Processing as "computational techniques that process spoken and written human language as language". According to the Microsoft researchers, the goal of the Natural Language Processing (NLP) is "to design and build software that will analyze, understand and generate languages that humans use naturally, so that eventually one will be able to address their computer like addressing another person". Machine Translation is used for translating texts for assimilation purpose which
aids bilingual or cross-lingual communication and also for searching, accessing and understanding foreign language information from databases and web-pages. In the field of information retrieval a lot of research is going on in Cross-Language Information Retrieval (CLIR), i.e. information retrieval systems capable of searching databases in many different languages. Construction of robust systems for speech-to-speech translation to facilitate "cross lingual" oral communication has been the dream of speech and natural language researchers for decades. Machine translation is an important module in speech translation systems. Currently, computer assisted learning plays a major role in academic environment. The use of Machine Translation in language learning has not yet got enough attention because of poor quality of automatic translation output. Using good automatic translation system, students can improve their translation and writing skills. Such system can break the language barriers of students and language learners.

4. METHODOLOGY

The Translation tool is developed by using statistical machine translation method. It uses the words which are already available and already using in regular life. Tamil is a morphologically rich language with free word order of Subject-Object Verb pattern. The baseline machine translation system would not perform well for the languages with different word order and disparate morphological structure. For resolving this, binary tree traversal model is introduced in Machine aided language translation system. We can see the basics of the binary tree traversal method.

4.1 Binary Tree

In computer science, a binary tree is a tree data structure in which each node has at most two children, which are referred to as the left child and the right child.
Ex:

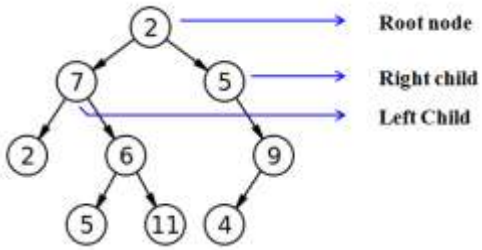


Figure. 4.1 B-Tree

4.2. Traversal methods on binary Tree

In-order - The ordering is
Left child node → Parent node → Right child node
So, 1 → 2 → 3

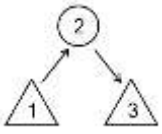


Figure 4.2. In-order

Pre-order
The ordering is
Parent node → Left child node → Right child node
So, 2 → 1 → 3

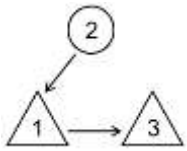


Figure 4.3. Pre-order

Post-order
The ordering is
Left child node → Right child node → Parent node
So, 1 → 3 → 2

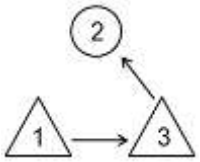


Figure 4.4. Post-order

We are implementing the in-order tree traversal for our machine translation system. The procedure and examples are as follows:

4.3. Implementation rules

- Every words in a given sentence are stored in every node of a Binary tree
- Verbs are stored in parent nodes only
- Nouns are stored in child nodes only
- A verb will combine at least 2 nouns

If more than 2 nouns then, a empty parent node will occur.

4.4. Motivations

People from rural
- Not able to communicate in Second Language (SL), but educated
- Not able to express their thoughts in SL
- Struggling to start higher studies in Engineering & Technology
- Not able to know the news & technologies which are in SL
- Not able to get knowledge in SL

4.5 Scope

It is very useful for
- all people who want to express their thoughts in English
- all rural schools & college students
- legal document preparation
- to teach Tamil for foreign students

5. LITERATURE SURVEY

5.1. Machine Translation Systems for Tamil

- Mr. Prashanth Balajapally developed English to Hindi, Kannada, and Tamil and Kannada to Tamil language-pair example based machine translation. This system is based on a bilingual dictionary comprising of sentence-dictionary, phrases-dictionary, words dictionary and phonetic-dictionary is used for the machine translation. Each of the above dictionaries contains parallel corpora of sentence, phrases and words, and phonetic mappings of words in their respective files.

- Prof. C.N. Krishnan at AU-KBC Research Centre, Anna University Chennai said that the machine-aided translation system to translate Tamil to Hindi. This system is based on Anusaaraka machine translation system. It uses a lexical level translation and has 80-85% coverage. Stand-alone, API, and Web-based on-line versions are developed. Tamil morphological analyzer and Tamil-Hindi bilingual dictionary are the by-products of this system. They also developed a prototype of English-Tamil MAT system. It includes exhaustive syntactical analysis. At present it has limited vocabulary and small set of transfer rules. Telugu-Tamil machine translation system is also being developed at CALTS. This system uses the Telugu morphological analyser and Tamil generator developed at CALTS. The backbone of the system is Telugu-Tamil dictionary.

- Mr. Ruvan Weerasinghe developed a SMT system for Sinhala to Tamil Language on 2004. In this the corpora were utilized from newspaper of Sri Lanka which publishing in both languages. He has also collected corpora from a website that contains translations of English articles into Sinhala and Tamil. These resources are formed a small trilingual parallel corpus for this research. This corpus consists of news items and articles related to politics and culture in Sri Lanka. The fundamental task of sentence boundary detection was performed employing a semi-automatic approach. In this scheme, a basic heuristic was first applied to identify sentence boundaries and those situations that were exceptions to the heuristic identified. Sentences are aligned in manual way. All language processing done used raw words and were based on statistical information.

- From Amrita School of Engineering, Coimbatore, the Computational Engineering and Networking research centre proposed a English – Tamil translation memory system. The system is based on phrase based approach by incorporating concept labeling using translation memory of parallel corpus. This system consists of 50,000 English – Tamil parallel sentences, 5000 proverbs, and 1000 idioms and phrases, with a dictionary containing more than 2,00,000 technical words and 1,00,000 general words.

- Mr. Saravanan developed a Rule based Machine translation system for English to Tamil on 2010. He used statistical machine translation approach; Google developed a web based machine translation engine for English to Tamil language. This system is also having the facility to identify the source language automatically.

5.2. Difference from other works

- Mainly focused on Tamil to English Translation
- Usage of huge amount of corpus
- Developing Huge Dictionary for Classical Tamil
- Use of Binary Tree Traversal method is new for Machine Translation in Tamil

6. OVERVIEW

This tool is a combination of rules based and statistical machine translation. On the rules based MT method translates every word of a sentence and rearranges it according to the second language. On the statistical based MT checks and debugs the errors of the translated sentence.
The tool has the following process to achieve the target translated sentence.

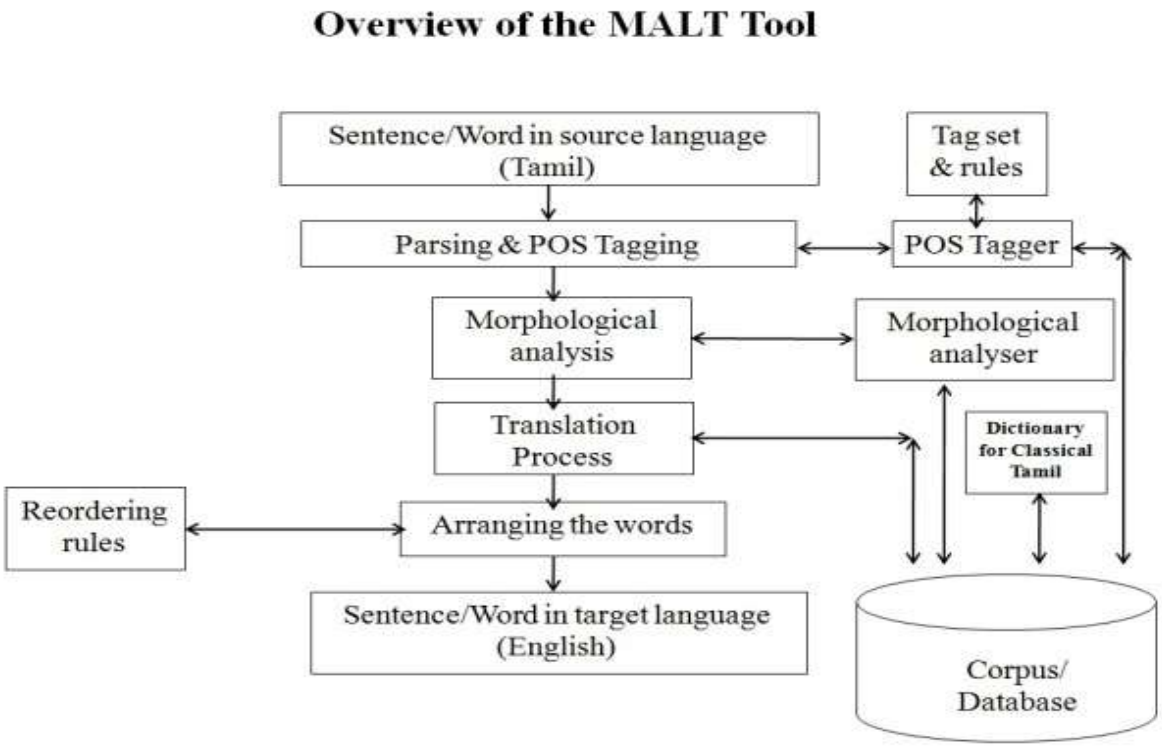


Figure 6.1

6.1. Sentence in source language

At first the user can enter the sentence in our own language (source language) Tamil. The user can use the virtual key board to enter the sentences in the appropriate place. The keyboard have all Tamil letters with its combinations also.

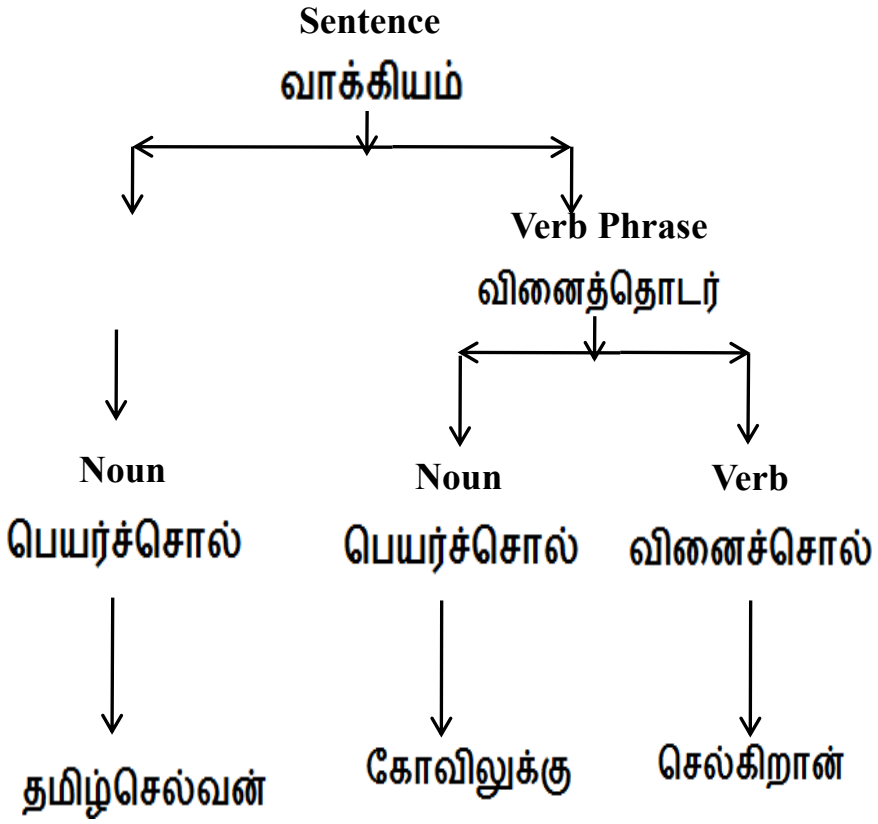
Figure 6.2Tamil Typing method

6.2. Pre-process

Sentence/Word in source language

⟶ தமிழ்செல்வன் கோவிலுக்கு செல்கிறான்

Parsing & POS Tagging



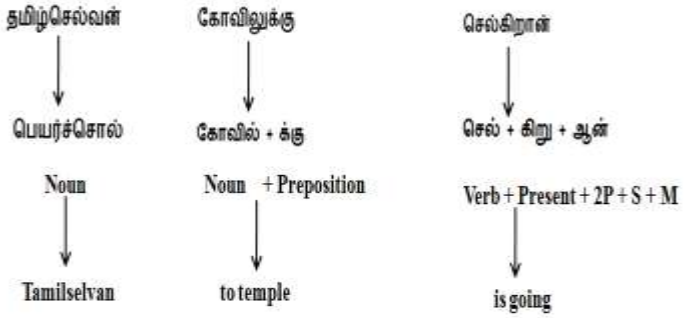
S. No	Word	Tag Set
1.	தமிழ்செல்வன்	பெயர்ச்சொல் (N)
2.	கோவிலுக்கு	பெயர்ச்சொல் (N)
3.	செல்கிறான்	வினைச்சொல்(V)

Stores in an array W[]= { தமிழ்செல்வன் , கோவிலுக்கு , செல்கிறான் }
and a tag array T[]={ பெயர்ச்சொல் , பெயர்ச்சொல் , வினைச்சொல் }

6.3. Morphological analysis.

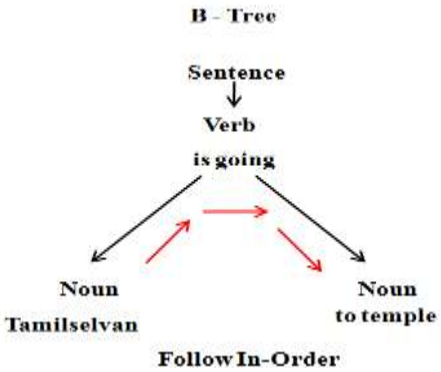
W1	W2	W3
தமிழ்செல்வன்	கோவிலுக்கு	செல்கிறான்
தமிழ்செல்வன்	கோவில்	செல்
தமிழ்செல்வனை	கோவிலினை	செல்கிறேன்
தமிழ்செல்வனோடு	கோவிலை	சென்றேன்
தமிழ்செல்வனால்	கோவிலோடு	செல்வேன்
தமிழ்செல்வனுக்கு	கோவிலினால்	செல்கிறார்
தமிழ்செல்வனின்	கோவிலால்	சென்றார்
தமிழ்செல்வனது	கோவிலுக்கு	செல்வார்
தமிழ்செல்வனுகாக	கோவிலிற்கு	செல்கிறான்
தமிழ்செல்வனின்கண்	கோவிலின்	சென்றான்
தமிழ்செல்வனாலான	கோவிலது	செல்வான்
தமிழ்செல்வனுடைய	கோவிலுக்காக	செல்கிறாள்
தமிழ்செல்வனில்	கோவிலின்கண்	சென்றாள்
தமிழ்செல்வனுடன்	கோவிலாலான	செல்வாள்
	கோவிலுடைய	செல்கிறாய்
	கோவிலால்	சென்றாய்
	கோவிலில்	செல்வாய்
	கோவிலுடன்	செல்கிறோம்
		சென்றோம்
		செல்வோம்
		செல்கிறார்கள்
		சென்றார்கள்
		செல்வார்கள்
		செல்கிறது
		சென்றது
		செல்லும்
		செல்கின்றன
		சென்றன

6.4. Translation Process



Source words in an array W[]= {தமிழ்செல்வன் , கோவிலுக்கு ,செல்கிறான்}
Translated words in an array T[]={ Tamilselvan , to temple , is going }

6.5. Arranging the words



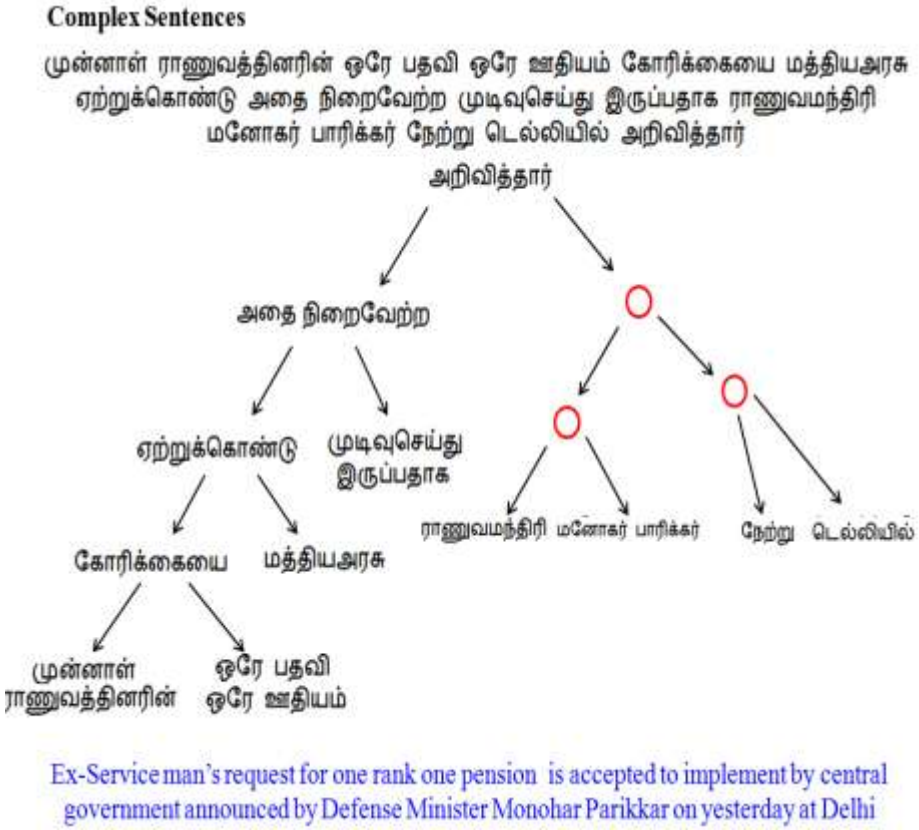
Before Tree Traversal : Tamilselvan to temple is going

After Tree Traversal : Tamilselvan is going to temple

Input for Translate → தமிழ்செல்வன் கோவிலுக்கு செல்கிறான்

Tamilselvan is going to temple

6.6. Process on Complex Sentence



6.7. Developing large set of Corpus

The basic process for the development of our tool is to develop large collection of corpus with local meaning and it's Unicode. In our corpus we are trying to collect above 6 Lakh words in English. For every English word we have to find out a suitable meaning in local language and the Unicode for that local meaning. For this large size of corpus we are collecting most of the English words, Indian places, Indian names, name of the Indian things, etc., The words in the corpus is stored with its POS Tag set also. If is it Tree (மரம்) means it would be stored with its POS Tag information also. Tree (மரம்) => N(T), Noun (Thing).

அகக் கணக்காய்வு	internal audit
அகட்டல்	Dilatation
அகத்தூண்டுதல்	inspiration
அகநகர்	citadel
அகநய	economic
அகந்தை	courage
அகன்ற	broad
அகன்ற அலைவரிசை	broadband
அகன்ற பலமான பலகை	tabletop
அகப்படாமல் தப்பு	to evade
அகப்படு	acquire
அகப்பை	ladle
அகப்பை	dipper
அகமுக நோக்காளர்	introvert
அகமொழி	langue
அகம்பாவம்	effrontery
அகம்பாவம்	impudence
அகரமுதலி	Dictionary
அகற்றல்	adoption
அகற்றுதல்	to Ablation
அகலக்கோடு	latitude
அகலத் திற	to unlock
அகலத்திரை	widescreen
அகலப் பட்டை	wide band
அகலப்பாதை	Boulevard
அகலமாக்கு	broaden
அகல் தளப் படகு	barge

Figure 6.3. Corpus

6.8. Developing tag set and designing rules

The tag set is developed depending upon the grammatical rule of target language (English).It has all the basic and some compound tags with its correlated rules for tagging. The rule for Tamil sentence is பெயர் + இடை + உரி + வினை. So the sentence order in Tamil is SOV pattern. Then we consider the English word order in a sentence, it should be a SVO pattern. Likewise when the Tool reordering the source sentence into its corresponding sentence in target language it looks up the tag set and the rules to tag the corresponding tags.

6.8.1. Reordering into target language

This is the final process to assigning the correct meaning of the source language sentence to its target language sentence. It checks the meaning, morphological correction, sentence structure. This step results the sentence in target language with accurate meaning.

6.8.2. Developing front end

The front end is the user friendly page, with a Tamil keyboard to enter words and sentences in local language. The system will do the process as we discussed in the figure 6.1. It should display the actual meaning of the word or sentence in second language (English). For this conversion process the system will do some background process.

6.8.3. Background work

After retrieving an equal word in second language the tool lookup the next word in the source language field. The same work will be done for the next word also. At the end of the every word translation in the sentence, the tool can verify the order of every word in the sentence and it shows the meaning in second language and local language also. The user (student) has to find out the correct meaning of the sentence in the first language then the tool will convert it into the second language with appropriate meaning.

6.8.4. Trail results

By implementing the developed front end designing, and background work such as corpus, morphological analyzer, POS tagging we are getting the result of actual work what we have expected at first. Figure 6.3 show the trail translation from Tamil (source language) to English (second language).

Figure. 6.4.User Entry

Figure 6.5 Morphologically searching

Figure 6.6. Morphologically searching - I

7.0 Conclusion

We have successfully developed a trial tool that translates mother language (Tamil) to second language (English) used in rural students. But we have to do more improvement in the corpus and translation accuracy. We have implemented Binary tree traversal method to search words for its appropriate meaning from the corpus. We hope it will get good results from the users. We are in development of the full version of the tool.

8.0 References

1. De Marneffe, Marie-Catherine, and Christopher D. Manning. "Stanford typed dependencies manual." The Stanford Natural Language Processing Group. Last modified 2008. http://nlp.stanford.edu/software/dependencies_manual.pdf.
2. Dhanalakshmi, V., M. Anand Kumar, R.U Rekha, C. Arun Kumar, K.P. Soman, and S. Rajendran. "Morphological Analyzer for Agglutinative Languages Using Machine Learning Approaches", Advances in Recent Technologies in Communication and Computing." *International Conference on Advances in Recent Technologies in Communication and Computing*,, 2009.
3. Dorr, Bonnie J., Eduard H. Hovy, and Lori S. Levin. "Machine translation: Interlingual methods." (2004).
4. Girju, Roxana. "Introduction to Syntactic Parsing." *l2r. cs. uiuc. edu* (2004).
5. https://www.youtube.com/watch?v=bx7eqslOCI8
6. http://developeriq.in/articles/2010/aug/26/transliteration-using-java/
7. ww.w.iloveindia.com\languages of India
8. http://jsfiddle.net/R7ZjV/
9. http://www.sas.upenn.edu/~vasur/project.html
10. http://www.learnersdictionary.com/3000-words/alpha/d/7.
11. http://www.manythings.org/vocabulary/lists/l/
12. http://www.talkenglish.com/Vocabulary/english-vocabulary.aspx
13. http://www.onlineocr.net/
14. Vikram, T. N., and Shalini R. Urs. "Development of Prototype Morphological Analyzer for the South Indian Language of Kannada." *Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers*, 2007, 109-116.