EtranS- A Complete Framework for English To Sanskrit Machine Translation

Promila Bahadur
Amity University
Noida, India

A.K.Jain
Indian Institute of Technology
Kanpur, India

D.S.Chauhan
Uttrakhand Technical University
Dehradoon , India

Abstract— **Machine Translation has been a topic of research from the past many years. Many methods and techniques have been proposed and developed. However, quality of translation has always been a matter of concern. In this paper, we outline a target language generation mechanism with the help of language English-Sanskrit language pair using rule based machine translation technique [1]. Rule Based Machine Translation provides high quality translation and requires in depth knowledge of the language apart from real world knowledge and the differences in cultural background and conceptual divisions. A string of English sentence can be translated into string of Sanskrit ones. The methodology for design and development is implemented in the form of software named as "EtranS".**

Keywords- Analysis, Machine translation, translation theory, Interlingua, language divergence, Sanskrit, natural language processing.

I. Introduction

English is a widely spoken language across the globe and most official communication and documentation is being done in this language. In India, there exist several regional languages including Hindi, where a lot of documentation exists in this language. The Sanskrit is considered to be mother of all Indian languages and is one of the oldest synthetic language in which a lot of ancient literature exists. Since English is modern day "global language", it has always been a challenge before natural language processing community to find efficient mechanism for this translation pair [2, 3, 4].

We compare and analyze differences between the two languages which are pre-requisite before getting into translation technique. There are four major parameters namely, essence, tense, number and translational equivalence, that are needed to be considered for the translation of this language pair. The essence of English is that it is evolved, therefore, it is a natural language. Sanskrit is formulated by sages like Panini hence it is an Artificial or Synthetic language. The English language has twelve tenses in all primarily Past, Present and Future. All three have a Perfect, Indefinite, Continuous and Perfect Continuous and it makes twelve forms of tenses. Sanskrit has primarily six tenses, Present, Past, Future, Order, Blessing and Inspiration. The English have two numbers i.e., Singular and Plural whereas, Sanskrit has three numbers Singular, Dual and Plural [6,9,10,11]. In general, we can state that the model consists of array of translation rules to translate from source to target sentence, which is the frame of Rule based Machine Translation System. The approach is simple and effective. The rules are framed, keeping in view the grammar of the source and the target language (Translational Equivalence) [2].

We have discussed different types of sentences considered for translation, sources from which sentences are taken, formation of rules pattern, provisions for extension of rules and lexicon and features of the software developed. We have also discussed on robustness of the rules and limitation of the software. We begin with simple statements; subsequently translation of compound statements is done.

II. The Reasearch Approch

The translation model from English to Sanskrit is primarily based on formulation of Synchronous Context Free Grammar (SCFG), a sub set of Context Free Grammar (CFG). SCFG helps in linguistics representation of the syntax. A CFG (N, T, and P) consists of set of non-terminals N, terminals T and Productions P [7].

$$P= \{N \rightarrow \{N \cup T\}^*\}$$

The productions in the right hand side are replaced by that of left hand side sequence of terminal and non-terminals. Non terminals are written recursively until only terminal symbols remain [8]. In our case terminal symbols are words and non-terminal symbols are syntactic categories, i.e., a sentence will have a start symbol S which searches through different routes to rewrite the symbols until all the possibilities have been explored or input sentence is generated. The tree representation is shown in Figure 1.

III. The Process Engine

The functional approach to translation is developed on the basis of the process engine shown in Figure 2, further for implementation of the same "The Two Way Translation Model"[1] shown in Figure 3 is developed. This model states that for the translation from source to target language first Top Down and then Bottom UP approach is adopted.

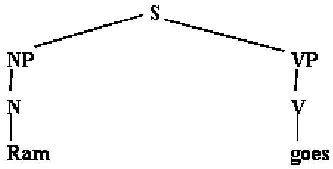


Figure1. Tree representation of "Ram goes"

It presents a simple technique for translation. It has two phases, the first phase follows, the Top Down approach. Here, we begin with syntax analysis, followed by semantic analysis and then mapping of tokens is done, which are generated during syntax analysis. The second phase, does Bottom to Top analysis. It begins with intermediate process of mapping, felicitated by first phase, which is followed by morphological analysis and finally target language is generated.

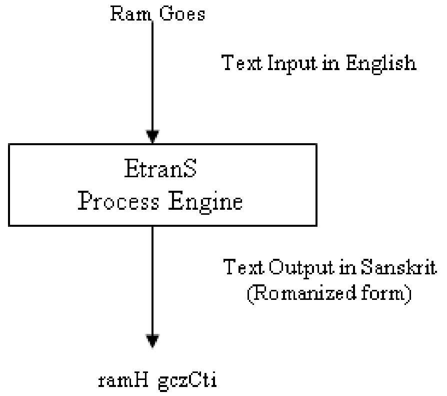


Figure 2. EtranS Process Engine

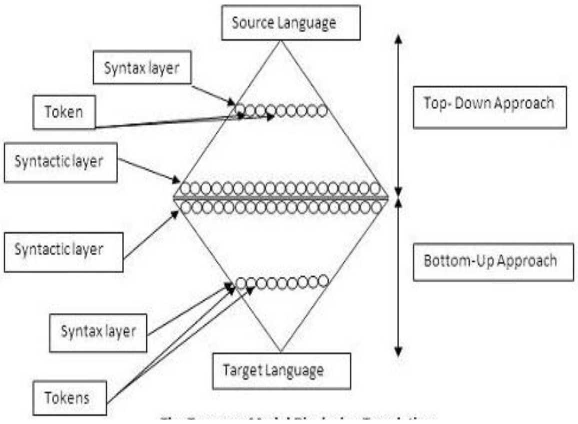


Figure 3. Two Way Model

The flow chart shown Figure 4 is based on the on the Two-way model. The model can be adopted in general for the translation from L1 to L2 language.

IV. GRAMMARS AND SENTENCE STRUCTURES

In this section, we have outline, how a sentence can be broken into major subparts and end up in forming a tree. Here each node represents phrases, such as, Noun Phrase or Verb Phrase and leaves represent parts of speech like noun, verb, adjective etc. Top down parsing is started from the sentence of source language and ends up to symbol list. A lexicon is used to store possible categories of words. Simple Top-Down parsing algorithm is used to generate possibilities list. The first element is a current state, which consist of symbol list and a word position in the sentence and the remaining elements are the back-up states, e.g., consider sentence Ram goes. As shown in Table 1.

Ram: N

Goes:V

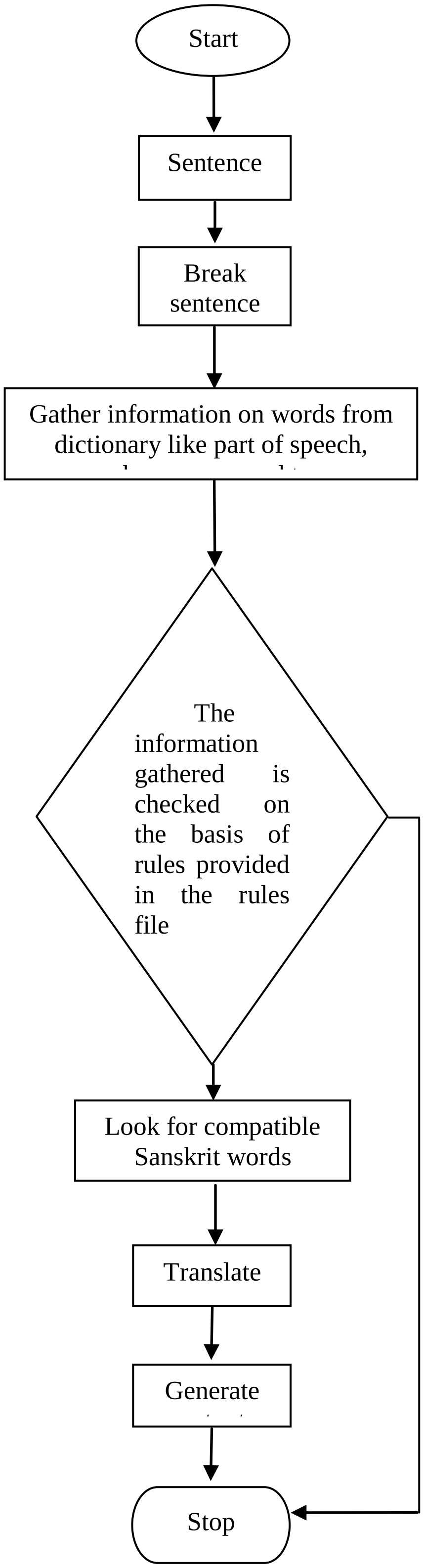

Start
Sentence
Break sentence
Gather information on words from dictionary like part of speech,
The information gathered is checked on the basis of rules provided in the rules file
Look for compatible Sanskrit words
Translate
Generate
Stop

Figure 4. Translation Process

TABLE I. TOP DOWN PARSING OF THE SENTENCE "RAM GOES"

Step	**Current State**	**Backup state**	**Comment**
1	(S)1		
2	NP VP		S rewritten to NP VP
3	(N VP)1	((N VP)1)	NP rewritten
4	*((N VP)2)*	*((N V)1)*	*The back up state remains*
5	*(())*		*Success*

V. MORPHOLOGICAL PROCESSING

A word may consist of single morpheme (a smallest single meaningful unit) or may have root word comprising of an affix or suffix. Lexicon lists all forms of the word. An input sentence can be processed into sequence of morphemes. At times words may be ambiguous and have multiple decompositions into morphemes. E.g., "Ram brought pen", this sentence depicts past form of buy therefore the lexicon provides grammatical information related to it both in English and Sanskrit.

VI. THE FUNCTIONAL APPROACH

The process engine shown in Figure 2 can be divided into two major components and are the foundation of EtranS system, given below:

i. The parsing process

ii. The generator process

A. The Parsing Process

As shown in Figure 5, the process is responsible for the top to bottom analysis phase. It has following sub processes:

i. The Input Process
ii. Sentence Analyzer Process
iii. Morphological Analysis Process
iv. The EtranS Lexicon
v. The Parse Tree

1) The Input Process

Input process is the first small step towards translation process. It takes sentence as input in a text box developed as a part of GUI for the translation process.

2) Sentence Analyzer Process

Sentence Analyzer process does analyses the sentences taken by the Input process and state the category of the input sentence, i.e., whether the sentence is small, large or extra-large. The sentence is divided into tokens (e.g. Ram goes to the market.

For this sentence, tokens would be generated as ram, goes, to, the, market). Tokens can also have more than one word e.g. (take off, in case of). It also extracts the root words from the tokens. E.g., from "goes" "go" is extracted. The tokens are identified and morphological analysis is done.

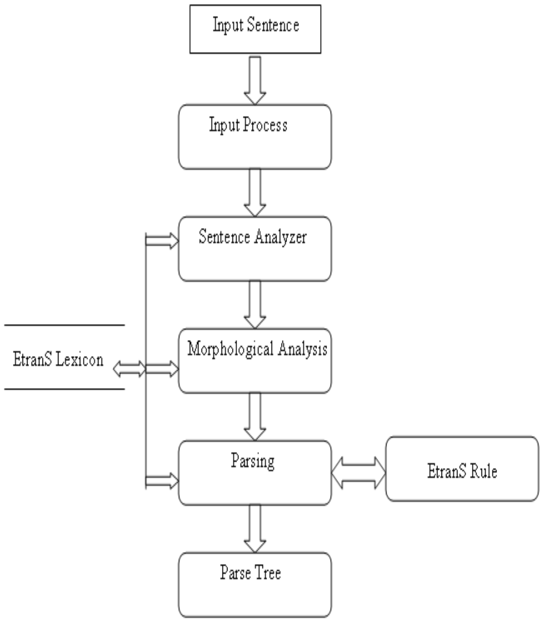


Figure 5. Parsing Process

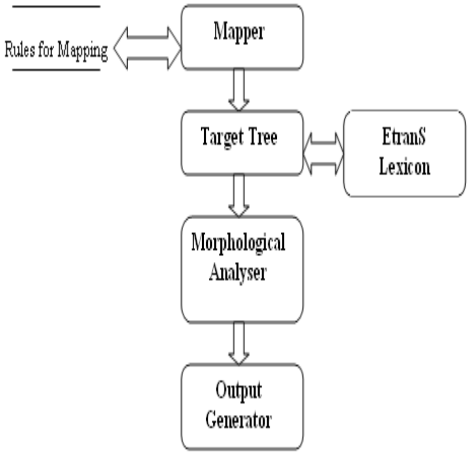


Figure 6. Generator Process

3) Morphological Analysis Process

The lexicon or the database developed plays a pivotal role for Morphological analysis. As it searches through the lexicon to gather above mentioned categories of the part of speech and its sub categories.

This process takes the tokens as input and gathers grammatical information on that. E.g., Morphological analysis for tokens like Ram, is, eating would provide following information is shown in Table 2.

TABLE 2. MORPHOLOGICAL ANALYSIS OF "RAM IS EATING"

Noun	Verb	Root Word
Ram- name of a person Proper noun , Masculine gender, Singular	Is auxiliary verb Eating-to eat transitive verb, animate, continuous word	Eat

4) The EtranS Lexicon

A bilingual lexicon is developed as a bridge between the source as well as the target language [1]. Structure of the lexicon contains various categories and sub categories pertaining to the source and the target language. The database consists of words, their grammatical characteristics like part of speech, tense, number and gender. The database in Sanskrit is stored in the same manner though emphasis is on phonemes like akarant[1], Akarant, ikarant, Ikarant, ukarant etc words and on the gender, e.g., we can take up akarant masculine words like ram[2], shyam, ghat etc, to make it vibhakti[3] ekvachan[4], we have to add "H" as suffix therefore the words can be like

Ram +H	ramH

The information on verb is grouped on the basis gan[5] which are ten in number and have further three types atmaypad[6],parasmaypad[7] and ubhaypad[8].These in turn, have different types of tenses for e.g., in case of lat lakar or present tense root word is taken and appropriate suffix is added to it, to obtain the desired result taking care of exception. For example,

Gam	gczC+ti=gczCti

Gam is a root word which forms gachti means to go.

5) The Parse Tree

This process checks whether the input sentence is grammatically correct with the help of EtranS rule bank, as shown in Figure 5. The information gathered from above mentioned process helps in analyzing the grammatical aspect of the sentence and on the basis of the rules assessment is done for e.g. for sentences "Ram is eating" as shown in Figure 7 and "Ram are eating" as shown in Figure 8 we have the following analysis respectively

i. This is a rule in Present Continuous tense, therefore the sentence stands to be true both at morphological as well as parser level.

ii. Hence, this sentence is correct.

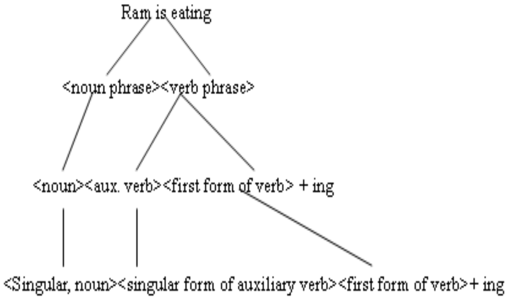

Ram is eating
<noun phrase><verb phrase>
<noun><aux. verb><first form of verb> + ing
<Singular, noun><singular form of auxiliary verb><first form of verb>+ ing

Figure 7.Tree Showing analysis of "*Ram is eating*"

.

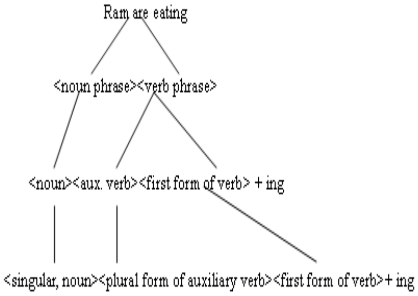

Ram are eating
<noun phrase><verb phrase>
<noun><aux. verb><first form of verb> + ing
<singular, noun><plural form of auxiliary verb><first form of verb>+ ing

Figure 8.Tree Showing analysis of "*Ram are eating*"

i. This is not matching rule meant for Present Continuous tense, therefore the sentence stands to be true at morphological level but false at parser level.

ii. Hence, this sentence is not correct.

B. THE GENERATOR PROCESS

As shown in Figure 6, the Generator process is the second phase of translation process which is reverse of the parsing process. It is composed of following processes:

i. Mapping Process

ii. Morphological Analysis

iii. Output Process

1) Mapping Process

Mapping process looks for grammatical compatibility between the source and the target language, as both have different grammatical approach[1], e.g., English have preposition and conjunction in separate, while in Sanskrit prepositions and conjunctions are in-built with the respective words itself.

In Figure 9(A), tree is showing parsing of the sentence "Ram goes to school" and Figure 9(B) displays its translation after parsing and mapping process.

[1] These are noun categories in Sanskrit.
[2] These are examples of noun.
[3] It represents number of noun.
[4] Is singular form of noun.
[5] Main group of verb
[6] Is a type within karak or verb.
[7] Is a type within karak or verb.
[8] Is a type within karak or verb.

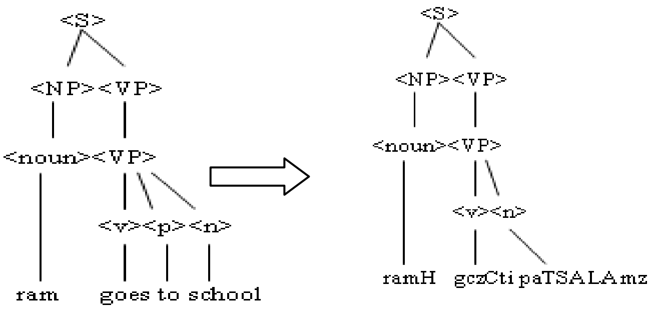


Figure 9(A)

Figure 9(B)

Trees showing translation (mapping) process from source to target language

2) Morphological Analysis

The Morphological Analysis looks for the root word in the target language and, on the basis of the above, mentioned mapping information, it maps with respective grammar to make the complete sense. As we can see in the Figure 9(A) that ram is a singular noun, therefore, in Figure 9(B) it is mapped with ramH (ram+H) [In Romanized form] which means root word + vibhakti [1].

3) Output Process

Output process gathers the information from the leaves of the trees generated after above processes. The leaves contains translated text which is finally taken as output to generate translated text.

VII. ETRANS SYSTEM

The system comprises of user interface developed using .NET framework and the lexicon using MS-Access 2007. The modules are developed on the basis of process engine. The user interface is responsible for taking input and generating the output. It is heavily dependent on the tables created in the database for generating output and the programming done to extract the information based on the logic developed.

The software comprises of following modules:

i. Parse Module

ii. Generator Module

A. PARSE MODULE

Parse module is responsible for taking sentences as input, analyzing and parsing; therefore a sentence has to go through the module to generate tokens, grammatical characteristics and syntax analysis :

i. Input Module

ii. Sentence Analyzer Module

iii. Morphological Analysis Module

iv. Parse Module

v. Parse Tree

1) Input Module

This module takes the sentences and submits it to the sentence analyzer module. It considers punctuation marks like comma (,) as it have a very prominent role in conveying meaning of sentence.

2) Sentence Analyzer Module

The sentences are divided into tokens and the category of the sentence is decided. The length of the sentence assists the software program in deciding the number of times loop have to be generated, for searching the grammatical information related to the tokens as shown in Table 3.

TABLE 3. ANALYSIS OF SENTENCE

Sentence	**Number of Tokens**	**Category**	**Comments**
Ram goes to school with his friend by car	9	Extra Large	Number of tokens extracted
Sita dances at annual function of the school with students	10	Extra Large	-do-

3) Morphological Module

Morphological analyzer generate token specific semantic information with the help of lexicon. The analysis of the above mentioned sentences is shown in Table 4.

TABLE 4. MORPHOLOGICAL ANALYSIS

Sentence	**Token**	**Semantic Information**	**Comments**
Ram goes to school with his friend by car	Ram goes to school with his friend by car	noun, third person, singular, masculine gender verb, present, singular preposition noun, singular, neutral gender preposition possessive pronoun noun, singular, third, neutral gender preposition noun, singular, neutral gender	Tokens extracted and semantic Information gathered
Sita dances at annual function of the school with students	Sita dances at annual function of the school with students	noun, third person, singular, feminine gender verb, present, singular preposition adjective noun, singular, third, neutral gender preposition article noun, singular, third, neutral gender preposition noun, plural, neutral gender	-do-

4) *Parse Module*

The semantic analysis is performed in this part of translation. The analysis is done on the basis of EtranS rule bank, which helps to ascertain whether a sentence is grammatically correct or not as shown in Table 5.

TABLE 5. PARSING OF SENTENCE

Sentence	Token	Number	Rule id	Comments
Ram goes to school with his friend by car	*Ram*	*99*	*a27*	*Numbers are generated and rules are matched*
	goes	*101*		
	to	*401*		
	school	*99*		
	with	*499*		
	his	*801*		
	friend	*3*		
	by	*403*		
	car	*3*		
Sita dances at annual function of the school with students	*Sita*	*99*	*a29*	*-do-*
	dances	*101*		
	at	*402*		
	annual	*601*		
	function	*3*		
	of	*406*		
	the	*703*		
	school	*3*		
	with	*499*		
	students	*3*		

5) *Parse Tree*

The parse tree is based on minimal attachment principal according to which the preference is given to syntactic analysis that creates the least number of nodes in the parse tree. Thus, based on the minimal attachment principal, following trees are prepared for the statements mentioned above. In particular a parsing tree on the left hand side depicts structure of the source and the corresponding target language tree is shown in the right hand side along with replacements done(for conjunction, preposition etc.) as show in Figure 10 and 11.

B. *GENERATOR MODULE*

Generator module takes semantic information from the morphological analysis module and does mapping followed by searching for the correct form of the words from the lexicon by considering root words to generate output of the source language.

1) *Mapping*

Mapping is done purely on the basis of the information passed from the morphological module. Since in Sanskrit a word contains conjunction, preposition and other information therefore this needs to be considered while mapping process as shown in Table 6

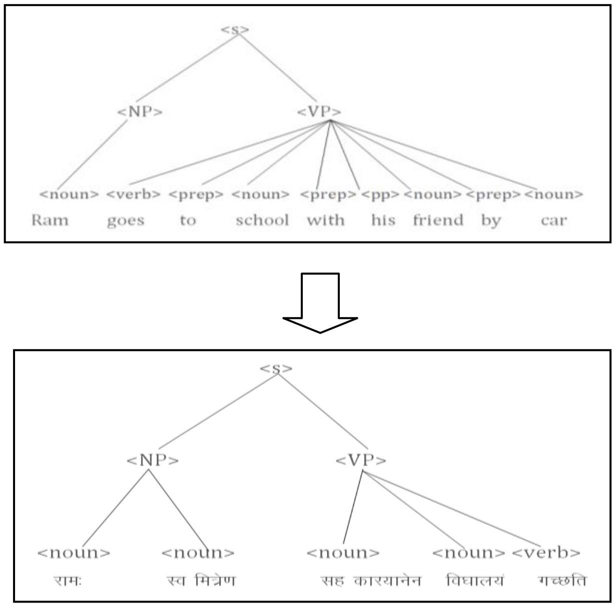


Figure 10[9] ParseTree of source and destination language

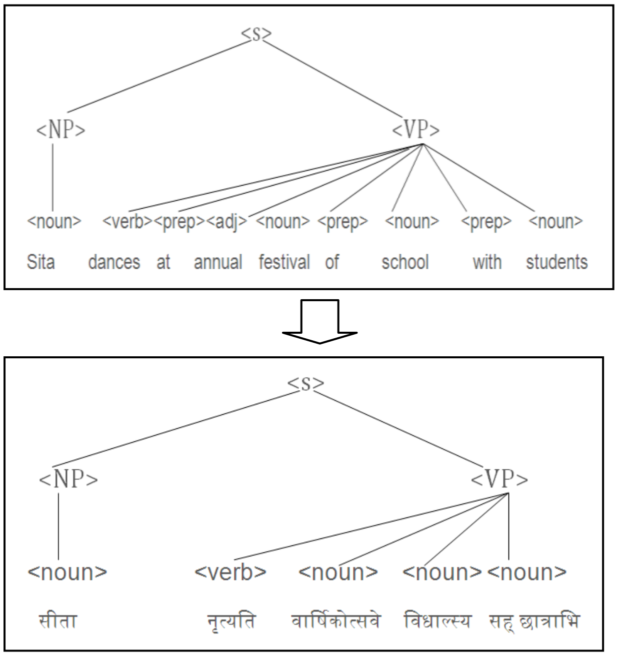


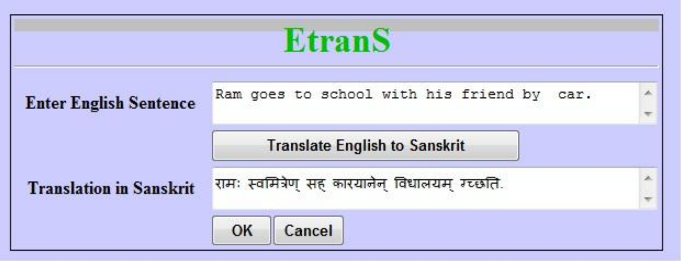
Figure 11.[10] ParseTree of source and destination language

[9] Figure.10 we can see that "his friend" is replaced by "Lo fe=s.k" , "by car" is replaced by "lg dkj;kusu" and "to school" is replaced by "fo?kky;a".

[10] Figure.11 we can see that "at annual festival" is replaced by "वार्षिकोत्सवे" and "of school" is replaced by "विधाल्स्य".

2) Output Module

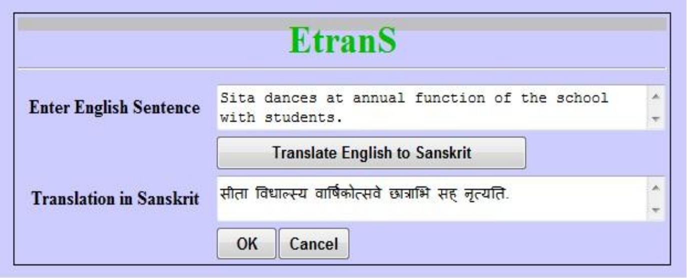
This module is provides the translated text produced after going through above mentioned modules, as shown in Figure 12 and 13.



Ram goes to school with his friend by car.

रामः स्वमित्रेण् सह् कारयानेन् विधालयम् ग्च्छति.

Figure 12.Translation of “Ram goes to school with his friend by car” is “रामः स्वमित्रेण् सह् कारयानेन् विधालयम् ग्च्छति”.



Sita dances at annual function of the school with students.

सीता विधाल्स्य वार्षिकोत्सवे छात्राभि सह् नृत्यति.

Figure 13.Translation of “Sita dances at annual festival of the school with students” is “सीता विधाल्स्य वार्षिकोत्सवे छात्राभि सह् नृत्यति”

VIII. FORMATION OF RULES PATTERN

The rules formation is the most challenging task of the machine translation as it is the back bone of the whole process. It covers the entire process of translation starting from syntax analysis ending up to translation; we also covered the mapping process from source to target language.

The rules are framed in ascending order. We have started formation of rules starting with simple sentences and ending up to compound sentences. The size of the sentences ranges from small to large.

A. Extension of rule base and lexicon

The rule base can be extended on requirement basis or if we want to add complex sentences also. The extension of lexicon is also a simple affair as for all the groups present in Sanskrit and English language ids have been provided. The new word can be added keeping in view the grammatical information of the word to be added, e.g. if we need to add a noun like bucket dictionary can be referred and the number assigned would be 3

IX. FEATURES OF THE SOFTWARE

The software developed has following features.

i. The system can translate Simple and compound sentences from English to Sanskrit.

ii. The sentences can be simple and compound with affirmative and imperative type or of active or passive voice having any of the three tenses i.e. Present, Past and Future.

iii. Rule base is easy to expand as we have divided sentence into three categories namely Subject, Verb and Object and have provided identification numbers accordingly. The Rule base looks for number combination for making new additions.

iv. The lexicon is capable for new framing and following features have been added to the lexicon

a. Identification number has been assigned to all the groups available in English and Sanskrit.

b. New words can be added to the database by identifying the identification number both in English and Sanskrit.

c. Sanskrit is a strong typed language therefore word order is not a matter of concern.

X. RESULT AND CONCLUSION

In this paper, the complete framework for Rule Based Translation is outlined. The chosen language pair is English and Sanskrit, as a source and target language. The system (EtranS) supports both English and Sanskrit grammar such as noun, verb adjective etc. To check robustness of the rules, EtranS system took samples of five hundred sentences of various types, as the sentences of simple and compound of affirmative, interrogative and imperative types in active and passive voice. We have considered sentences from all the three tenses i.e. present, past and future. It is our belief that this methodology can be adopted for translation of similar languages. The rule base can be extended to translate various types of literature in English to Sanskrit. Based on the observations above, several experiments with EtranS were conducted. For each run we considered a set of sentence type like simple, imperative etc. The results of these experiments are summarized below in the Table 8.

The sentences are divided into three categories that are small, large and extra-large to find out the accuracy of the EtranS system. The performance of translation is graded as three categories A, B and C given in the Table 7.

In the proposed approach we have obtained ninety-nine percent of correctness for the small sentences and ninety percent accuracy for the extra-large sentences. The result shows ninety percent of the sentences are correctly translated, however due to linguistic ambiguities two percent of the sentences have reported error.

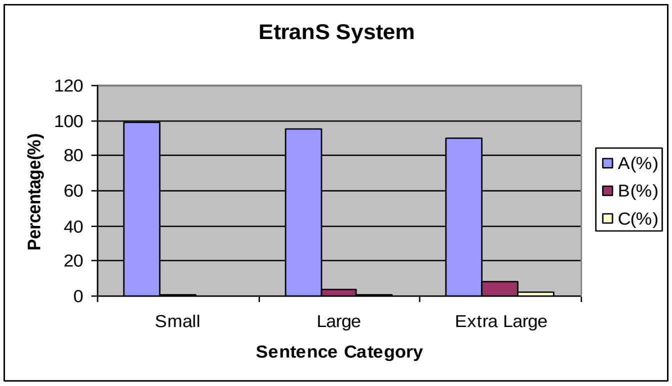


Figure 14 . Chart showing performance of the EtranS system

TABLE 6.MAPPING OF SENTENCE

Sentence	**Tokens**	**Translation**	**Semantic Information**	**Root word**	**Comments**
Ram goes to school with his friend by car	Ram goes to school with his friend by car	रामः ग्च्छति स्व विधालयम् मित्रेण् सह् कारयानेन्	प्राथमा विभक्ति , एकवचन ,पुल्लिंग शब्द प्रथम पुरुष, एकवचन (लट् लकार). अपव्व्य द्वितिया विभक्ति, एकवचन तृतीया विभक्ति, एकवचन. अपव्व्य तृतीया विभक्ति, एकवचन	राम गम् विधालय् मित्र कार	According to the root words the information is mapped with the semantic information and text would be generated
Sita dances at annual function of the school with students	Sita dances at annual function of the school with students	सीता नृत्यति वार्षिकोत्सवे विधाल्स्य सह् छात्राभि	प्राथमा विभक्ति , एकवचन , स्त्रीलिंग शब्द प्रथम पुरुष, एकवचन (लट् लकार-) सम्मी विभक्ति, एकवचन षष्ठीविभक्ति ,एकवचन अपव्व्य तृतीया विभक्ति ,बहुवचन.	सीता नृत्य विधालय् छात्रा	-do-

TABLE 7. CATEGORIES USED FOR RESULT ANALYSIS

Category	**Description**	**Remarks**
A	Sentence is correct in terms of grammar and translation.	-NIL-
B	Sentence is correct in terms of grammar but translation is not correct.	Due to the linguistic representations, few words in English may have multiple roles to play, e.g., the word became is used as a multipurpose word in English, e.g., He became king. She became sad. In Sanskrit, there are different representation for became in the above example. This is a constraint for the software but a linguist can decide where to use which word.
C	Sentence is ambiguous, i.e., it failed at the parsing level.	Few words in English may be used as both noun and verb. This generates ambiguity for the system. For example "The can can have water". A further line of work is required in this area to understand these anomalies.

TABLE 8. ANALYSIS OF THE ETRANS SYSTEM

Sentence	**Size**	**A (%)**	**B (%)**	**C (%)**
Small	>=3 & <= 5 words	99	1	0
Large	>5 & <= 8 words	95	4	1
Extra Large	>8 & <= 11 words	90	8	2

REFERENCES

[1] Promila Bahadur,A.K Jain and D.S Chauhan,"English to Sanskrit Machine Translation, ICWET 2011, Bombay, ACM 2011

[2]Shachi Dave etal, "Interlingua-based English-Hindi Machine Translation and Language Divergence",Machine Translation, Vol. 16, Issue 4, pp: 251 - 304, 2001.

[3] James Allen, "Natural Language Processing", Pearson Educations, 2005

[4] Anurag Seetha etal, "Improving performance of English- Hindi CLIR System using Linguistic Tools and Techniques", Pages: 261 HCI 2009.

[5] Alka Choudhary, Manjeet Singh, "GB Theory Based English to Hindi Machine Translation System, and IEEE 2005

[6] Promila Bahadur, A.K Jain,D.S Chauhan, "EtranS-English to Sanskrit Machine Translation" ICWET 2012, Bombay, ACM 2012

[7] Adam Lopez, "Statistical Machine Translation", ACM Computing Surveys", Vol. 40, No. 3, 2008.

[8] Aho,Ullman," The theory of Parsing ,Translation and Compiling", Pearson Educations, 2003

[9]English to Sanskrit machine translation semantic mapper, , International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5313-5318

[10]English to Sanskrit Machine Translator Lexical ParserAnd Semantic Mapper, Vaishali Barkade, National Conference On "Information and Communication Technology",NCICT-IOJ,2010.

[11]Approach of English to Sanskrit machine translation based on case-based reasoning, artificial neural networks and translation rules, Vimal Mishra, International Journal of Knowledge Engineering and Soft Data Paradigms, 2010-12-01

AUTHORS PROFILE

1. Ms Promila Bahadur is presently working as Assiatant Professor at Amity University, Noida since 2007. Her research is focused of Natural Language Processing.
2. Prof. Ajai Jain is Professor at Department of Computer Science, IIT, Kanpur.
3. Prof. D.S.Chauhan is working as Vice Chancellor at Uttrakhand Technical University, Dehradun