English to Sanskrit Machine Translation-EtranS System

Dr. Promila Bahadur
Asst Professor
Rashtriya Sanskrit Sansthan
Lucknow India.

ABSTRACT
In this paper, we describe the functional aspect to English to Sanskrit Machine Translator. The system developed, has been named EtranS. This paper also discusses the results obtained after testing the system on variety of sentences gathered from various sources.

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

1. INTRODUCTION
In the present paper, we have presented English to Sanskrit rule based machine translation system. The system explores grammatical characteristics of source language and target language. The findings include the agreement and disagreement between the two languages in terms of parts of speech, verb, tense, aspect and number. This catered, in development of the EtranS software in .Net framework. The system accepts input in the text format and generates output in Unicode Devnagri format. The result obtained is based on different types of sentences considered for translation, sources from where 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.

2. BACKGROUND
Machine Translation of natural language is a widely discussed and challenging topic. It has attracted linguistic as well as researchers, from, the mid of 20th century. This century witnessed birth and growth of machine translation. The progress can be recorded decade-wise, since 50's and prominent developments took place in each decade [3]. The era can be broadly categorized into five progressive generations [5,6,7].

First Generation- During the initial years, inputs from government and intelligence agencies, showed the need of machine translation. It was heavily funded by agencies due to its high potential, high speed quality translation being visualized. This concept grew with the time.
Second Generation- By its beginning, it was realized that automated translation is not achievable target as per the outcomes of the first generation. A report from National Academy of Sciences condemned the field as well as its workers. Though, the report was criticized as narrow and short sighted but the recommendations were adopted [3].
Third Generation- The situation worsened by its advancement. The government funded projects were functioning only and most of the projects elsewhere stopped [8]. Gradually, later in the era, private companies floated machine translation projects. Though, high expectations resulted in disrepute of machine translation.

Fourth Generation- Machine translation gained momentum and attracted attention from various walks of society e.g., government, business and industry. Now machine translation and Machine Aided Translation (MAT) system were in use [9]. Private and Government sectors started funding machine translation projects as the expectations with machine translation were more realistic. It came into light that machine translation has great potential.

This generation witnessed, the popularity of machine translation, among developing countries, like India, Japan etc. Researchers [4] suggested demand for technical translation. It was reported that worldwide work on machine translation is going on that includes array of projects. The research included various categories of translation like rough translation, full translation using machine translation, Value Added Network (VAN) service based on machine translation etc.

Fifth Generation- New dimensions of the machine translation were explored which included Statistical Machine Translation (SMT). It was introduced by IBM researchers in a workshop sponsored by the US National Science Foundation and Johns Hopkins University's Center for Language and Speech Processing [10,11].

By this time machine translation had became a household name and many projects were undertaken to carry out translation from one language to another [12,13]. As an important issue, the percentage of accuracy of the translation persisted. Many international and national projects like SYSTRAN, METEO, LOGOS [4], Anusaarka, AksharBharati [15] gained momentum. Individual efforts also surfaced in the project named Transliteration [16].

Translation on spoken language projects such as C_Star, ART, Eutrans,TC-Star, PF-Star etc [5] were also attempted. The written language translation was also attempted. For researchers international languages like Spanish, Portuguese, Japanese, German, Chinese, Arabic, Hebrew, Sinhalese etc., became desirable languages for translation. In Indian-Sub continent widely attempted languages for translation were Bengali, Hindi, Gurumukhi, Telegu, Kannad, Oriya etc.

3. THE SYSTEM
The system comprises of user interface developed using .NET framework and the lexicon using MS-Access 2007. The user interface which has various modules, which would be discussed below, are responsible for taking input and generate the output. The system heavily depends on the on the database for generating output and the programming is done in the interface to extract the information based on the logic developed. The software comprises of following modules [1][2], as shown in Figure 1. and Figure 2.:

i. Parse Module
ii. Generator Module

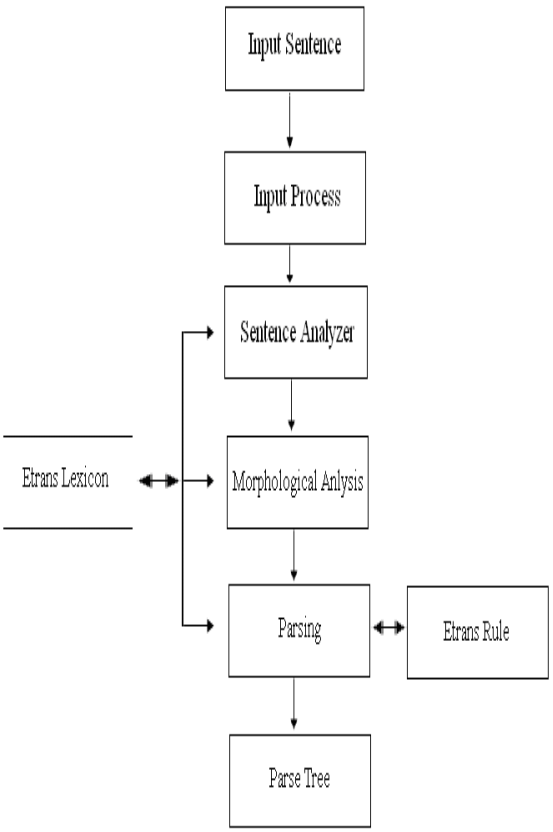

Figure 1: Parse Module

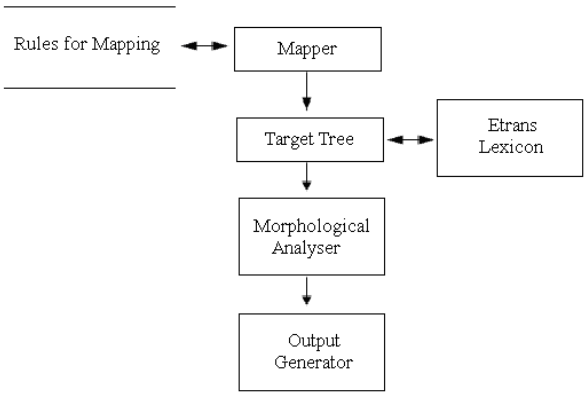

Figure 2: Generator Module

4. PARSE MODULE

This module takes sentence as input and performs top to bottom analysis. Following are the sub-modules needed to perform the analysis.

i. Sentence Analyzer Module
ii. Morphological Analysis Module
iii. Parsing Module

4.1 Sentence Analyzer Module

This module divides the input sentence into tokens and states the category of the input sentence after analyzing its size, as shown in Table1. The state of a sentence is stored into the database for further reference. We can see the illustration of same in Table 2.

Table 1: Category of Sentences

Category of Sentence	Size
Small	>=3 & <= 5 phrases
Large	>5 & <= 8 phrases
Extra Large	>8 & <= 20 phrases

Table 2: Analysis of sentence

Sentence	Number of Tokens	Category	Comments
The ant moved towards the leaf and climbed up there.	10	Large	-do-
She carried a large jug of milk on top of head.	11	Small	-do-

4.2 Morphological Analysis Process

This module takes the input token, generated by Sentence Analyzer Module to produce grammatical characteristic, e.g., 'Hari reads a book', 'Hari' is extracted as noun of singular form, third person, 'reads' as verb of singular form and 'book' as a noun. These characteristics are stored in the database, as shown in Table3.

Table 3: Morphological Analysis

Sentence	Token	Semantic Information	Comments
The ant moved towards the leaf and climbed up there.	the ant moved towards the leaf and climbed up there.	article noun, singular, neutral gender verb, past ,singular preposition article noun, singular, neutral gender conjunction verb, past ,singular	Tokens extracted and semantic Information gathered

		preposition preposition	
She carried a large jug of milk on top of head.	she carried a large jug of milk on top of head.	Preposition verb, past ,singular article adjective noun, singular, neutral gender preposition noun, singular, neutral gender preposition preposition preposition noun, singular, neutral gender	-do-

4.3 Parsing Module

The parsing module picks the characteristics obtained from morphological analysis and checks for syntax with the help of EtranS rule bank, as shown in Table 4, Figure 3. and Figure 4.

Table 4: Parsing of sentences

Sentence	Token	Number	Rule id	Comments
She carried a large jug of milk on top of head.	She Carried a large jug of milk	804 101 701 601 66 406 3	a29	Numbers are generated and rules are matched
	on top of head.	405 403 406 3		
Suddenly a fish appeared and picked him up	suddenly a fish appeared and picked him up	201 701 99 101 501 101 813 411	A35	-do-

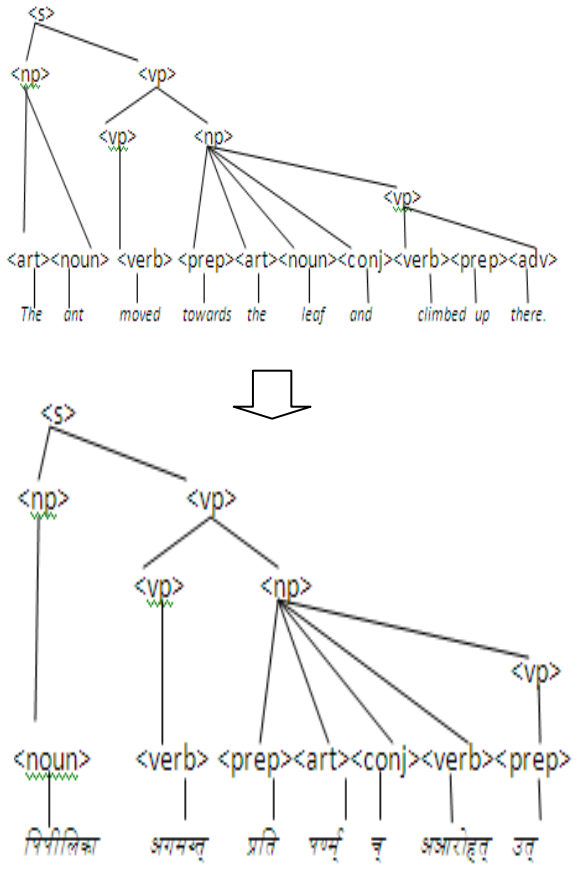

Figure 3. Mapping Tree

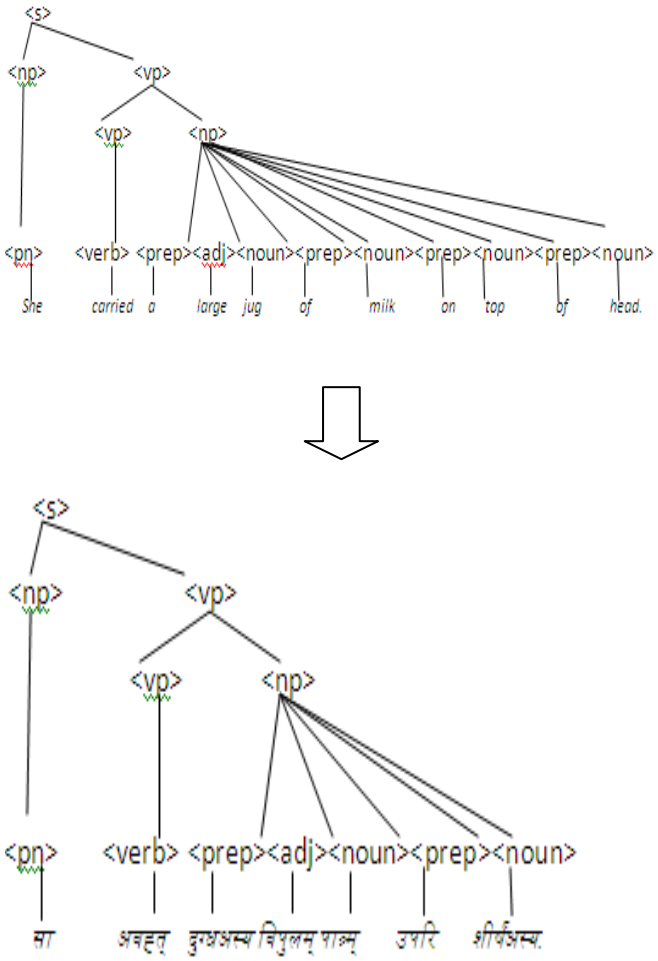

Figure 4. Mapping Tree

5. GENERATOR MODULE

The Generator Module performs the second phase of translation process, which is responsible for generation of the translated text after going through mapping and morphological modules. The generator module has following sub-modules:

i. Mapping Module
ii. Morphological Module
iii. Output Module

5.1 Mapping Module

This module maps the source tokens with the target tokens based on the rules and the information provided in the lexicon. It is responsible for gathering initial information of the source sentence and maps it with that of target sentence, e.g., "Ram Reads". The following information is obtained:

Length of the sentence – 2
ram-noun, singular, third.
reads - verb, singular, present.
Mapping would be done with
rAmH(jke%) - noun, singular, first.
pThati(ikB□□) - verb,singular,present.

The information gathered further helps in mapping process with the target language, as shown in Table 5.

Table 5: Mapping of sentences

Sentence	Tokens	Translation	Semantic Information	Root word	Comments
The ant moved towards the leaf and climbed up there.	the ant moved towards the leaf and climbed up there.	□□□ □□□ □□ □□□ □□□ □ □□□ □□ □□□ □□□ □ □□ □□□ □□□ □□	□□□□□□ □□□□□□□, एकवचन ,□□□□□□□□ □□□□ □□□□□ □□□□□, एकवचन एकवचन एकवचन □□□□□ □□□□□, एकवचन	□□□□ □□□□ गम □□□□ □ □□□□	According to the root words the information is mapped with the semantic information and text would be generated
She carried a large jug of milk on top of head.	she carried a large jug of milk on top of head	□□ □□□ □□□ □□□ □□□ □ □□□ □□□ □□ □□□ □□□ □□□ □□□ □ □□□ □□□ □□□ य	□□□□□□□□ □□□□□, एकवचन □□□□□ □□□□□□□ □□□□□ □□□□□□□	□□□	-do-

5.2 Morphological Module

This module searches the root word in the target language and

maps similar information to complete the translation process, e.g., "Hari reads a book" would be translated as "hariH pustakam paThati"(□□□□ □□□□□□□□□ □□□□).
Here hari would be searched as root word and H would be added as part of singular noun and pratham vibhakti to form 'hariH'(□□□□). Further, 'pustakm' (□□□□□□□□□) would be searched as it is neutral gender and singular in number and this would be followed by search of 'paT'(□□□), which is root word for read and 'ati' would be added as it is singular in number and in present tense. We can see the illustration of same in Table5.

5.3 Output Module

This module presents the output in the Sanskrit text with the help of Unicode after the Morphological Module has generated text. For example, as shown in Figure 5 "Hari reads a book". The Roman form would be 'hariH pustakam paThati' and Sanskrit form would be '□□□□ □□□□□□□□ □□□□' .We can see the illustration further in Figure 6 and Figure 7.
\

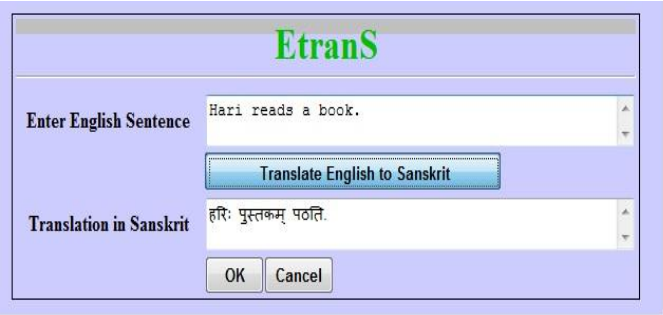

Figure 5.: Translation from English to Sanskrit

Figure 6.: Translation of "She carried a large jug of milk on top of head" is "सा दुग्धअस्य विपुलम् पात्रम् उपरि शीर्षअस्य अवहत्"

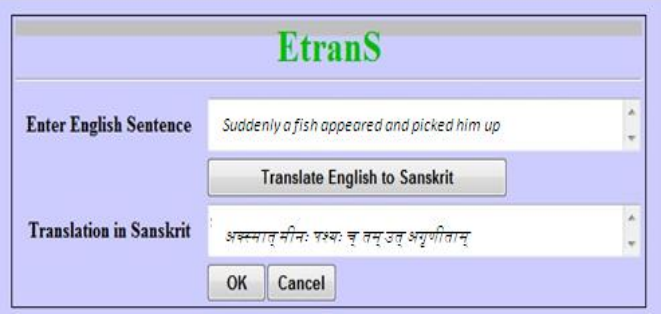

Figure 7.: Translation of "Suddenly a fish appeared and picked him up" is "अक्स्मात् मीनः पश्यः च् तम् उत् अगृणीताम्"

4.4 Database

The database consists of words, and its grammatical characteristics are part of speech, tense, number and gender. The database in Sanskrit is stored in the same manner. The emphasis is on phonemes like akarant[1], Akarant, ikarant, Ikarant, ukarant, 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 word can be as follows:

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, 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, e.g., 'Gam' is a root word which forms 'gachti' means to go.

gam	gczC+ati=gczCati

6. ALGORITHM

The algorithm for translation model discussed above is given below. The algorithm integrates findings related to the grammatical nature of the language and the computer science understanding. It applies knowledge to present translation from the source to target language.

The Algorithm

Step 1. Break the source sentence into tokens.

Step 2. Depending on the number of tokens generated, loop is generated to gather semantic and syntactic information of the source sentence.

Step 3. The information gathered from step 2, looks into the rule base for corresponding rule.

Step 4. The information gathered is checked by the rule base file.

Step 5. Mapping is done for the compatible words of target language.

Step 6. Generate output.

[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.

7. CATEGORISATION OF SENTENCE

The selection of sentences is categorized on the basis of their grammatical features. These sentences include simple and compound sentences of affirmative and imperative types in active voice. We have considered sentences from all the three tenses i.e., present tense, past tense and future tense.

7.1 Selection of Sentences

The samples of selected sentences are taken from various sources which include literary sources, grammar books and websites etc.

We have primarily referred grammar books for sentence sampling. They helped us in both in rules formation and sentence collection. Literary work as novels, short stories, essays etc. has been collected. Further, we have gathered samples from websites like that of BBC, Cambridge Press, and Oxford Press etc.

8. FORMATION OF RULES PATTERN

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

The rules are framed in ascending order of phrase size. The formation of rules begins with simple sentences and finish with compound sentences. The size of the sentences, vary from small to large.

8.1 Extension of the Rule Base and Lexicon

The rule base can be extended on requirement basis. The extension of lexicon can be done for all the groups present in Sanskrit and English language. Identification number has been provided to each part of speech (noun, verb, conjunction, or preposition) or in terms of Sanskrit language (‘vibhakti’ or ‘karak’). The new word can be added into the lexicon, based on grammatical characteristic of the system. For e.g., If we need to add a noun ‘bucket’ in the dictionary, the number assigned would be 3.

9. 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, assertive, negative and imperative types. In any of the three tenses i.e., present tense, past tense and future tense.

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

IV. The lexicon is enriched for framing rules. The 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 numbers.

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

VI. Correctness of the Software
The sentences are showing 79-81% correct results of translation for sentences with 3 to 9 phrase size.

The error rate is nearly of 19-21% due to following reasons

a) wrong syntax

b) words not available in the lexicon

c) rules not defined for the sentence

d) due to the representation of words in English language, such as the word “long” which gives different meaning in context to the sentence with which it is used, e.g., ‘Monkeys have long tail’. ‘I have long way to go’. The translation would be ‘□□□□□□□□□□□□□□ □□□□□□ □□□□□□’ and ‘□□□□ □□□ □□□□□ □□□□□□□□ □□□□□’ respectively. As we can see that in Sanskrit there are different representations for “long”. This is a constraint for the software but linguist can decide where to use which word.

e) due to the appearance of the words that can be both noun and verb. It is a constraint for the system. It is a limitation of the software and may be advanced in further researches, e.g., “Leaves are falling from the tree”.“Train leaves at two p.m”. It also requires inclusion of pratyay (□□□□□□□), which is based on gender. The gender based translation for verb is not the concern of our research. It will be the part of further advanced research . However, we have included stari pratyay(□□□□□□□□□□) as a part of our research.

VII. Limitation

a) A word can support more than one part of speech,as a word can be verb and noun at the same time. The popular reference of the word has been considered. This kind of research will be taken care in future course.

b) The complex sentences have not been included. It can be added by enhancing the rule base.

c) Date and time translation is not available it also can be done in future course.

d) Due to code optimization, the same type of words in the common number, witness extra characters or half characters. It can be refined in future research, e.g., verbs ‘climbs’ and goes have been assigned same identification number. While translation for past tense climbed is (□□□□□□□) and for went is (□□□□□□□□) for first person in Sanskrit. Here, extra (अ) has been added to the spelling of ‘□□□□□□□’, while translation.

10. RESULT

The system has been tested on sentences from grammar books, online text, children short stories, tagged parts of speech etc. We have pool of eight hundred and fifty sentences approximately. The sentences are in active voice, divided into three tenses, simple and compound in nature and having affirmative, imperative, interrogative and negative types. The sentences are divided into three categories that are small, large and extra large on the basis of phrase size. The sentences are graded as correct termed as A, grammatically deficient termed as B and incorrect translation termed as C. The performance is as below:

Table 6: Categories used for result analysis

Category	Description	Remarks
A	Sentence is correct in terms of grammar and translation.	-Nil-
B	Sentence is deficient in terms of grammar.	due to the representation of words in English language, such as the word "long" which gives different meaning in context to the sentence with which it is used, e.g., 'Monkeys have long tail'. 'I have long way to go'. The translation would be '□□□□□□□□□□□□□□ □□□□□□ □□□□□□' and '□□□□ □□□ □□□□□ □□□□□□□□ □□□□□' respectively. As we can see that in Sanskrit there are different representations for "long". This is a constraint for the software but linguist can decide where to use which word. Due to the appearance of the words that can be both noun and verb. It is a constraint for the system, which is a limitation of the software and may be advanced in further researches, e.g., "Leaves are falling from the tree". "Train leaves at two p.m". It requires inclusion of pratyay (□□□□□□□), which is based on gender. The gender based translation for verb is not the concern of our research. It will be the part of further advanced research. However, we have included stari pratyay (□□□□□□□□□□) as a part of our research.
C	Incorrect Translation	This is reported due different ways of expressing natural language, which requires interpretation by the linguistic, e.g., duitiya vibhakti requires "to" same is with chaturthi therefore there is ambiguity while it represents.

Based on the observations above, several experiments with EtranS were conducted. The analysis is based on simple, compound and the category of sentences. The sentences are taken in various moods as assertive, imperative and interrogative etc. The results of these experiments are summarized below:

Table 7: Analysis of the EtranS system based on simple sentences

Simple Sentence	A (%)	B (%)	C (%)
	52.81385	29.00433	18.1818

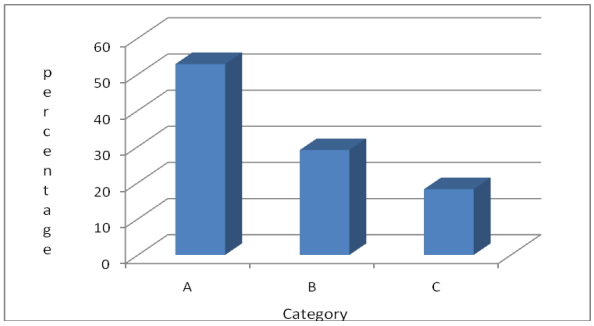

Figure 8.: Chart showing performance of the EtranS system based on simple sentences

Table 8: Analysis of the EtranS system based on simple sentences'phrase wise'

Phrase Size	A (%)	B (%)	C(%)
1	100	0	0
2	70	10	20
3	51.20	30.43	18.35
4	50	50	0
5	62.5	12.5	25

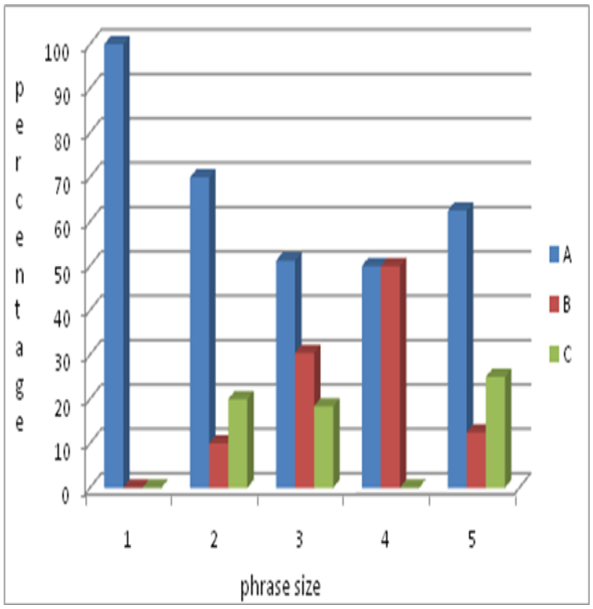

Figure 9: Chart showing performance of the EtranS system based on simple sentences 'phrase wise'

Table 9: Analysis of the EtranS system based on compound sentences'phrase wise'

Sentence	A (%)	B (%)	C (%)
	39.39394	39.39394	21.21212

Table 10: Analysis of the EtranS system based on compound sentences 'phrase wise'

Phrase Size	A(%)	B (%)	C(%)
4	100	0	0
5	50	33..33	16.66
6	21.42	50	28.57
7	50	50	0
8	0	50	50
9	0	100	0

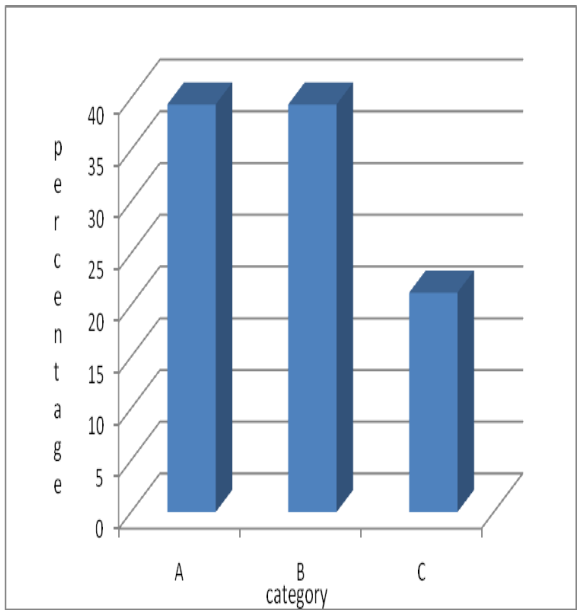

Figure 10.: Chart showing performance of the EtranS system based on compound sentences

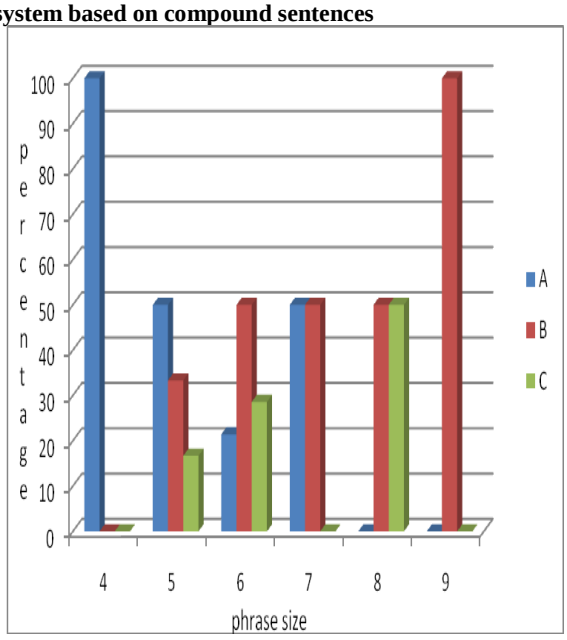

Figure 11: Chart showing performance of the EtranS

system based on compound sentences 'phrase wise'

REFERENCES

[1] English to Sanskrit Machine Translation; Promila Bahadur, A.Jain, D.S Chauhan, ACM Digital Library, 2011

[2] EtranS-English to Sanskrit Machine Translation; Promila Bahadur, A.Jain, D.S Chauhan, ACM Digital Library, 2012

[3] A Survey of machine translation –it's history, current status and future prospects, Jonathan Sloculn Microelectronics and Computer Technology Corporation Austin, Texas Computational Linguistics, Volume 11, Number 1, January-March 1985

[4] James M Lufkin Currentg Trands in Technical Translation 1989 IEEE

[5] Integration of Speech Recognition and Machine Translation in Computer-Assisted Translation Shahram Khadivi and Hermann Ney, Senior Member, IEEE, IEEE Transactions on Audio Speech and Language Processing, VOL. 16, NO. 8, NOVEMBER 2008

[6] An experiment on the machine translation of Languages carried on BESM, I. S Mukhin, Digital Computer Techniques, Paper no 2144, Nov 1956

[7] An experiment on the machine translation of languages carried on BESM, I. S Mukhin, Digital Computer Techniques, Paper no 2144, Nov 1956

[8] Computer translation —is it worth while? by W. L. PRICE, B B.Eng., Ph.D. Electronics & Power September 1967

[9] Benefits of automating data translation, Sandra A Mamra , Julie Barnes, 07407459/93/370C/0082/ C IEEE July 1 9 9 3

[10] Statistical Language Approach Translates into Success Steven J. Vaughan-Nichols, Techonology News, November 2003

[11] An Online Relevant Set Algorithm for Statistical Machine Translation Christoph Tillmann and Tong Zhang, IEEE Transactions on Audio Spech and Language Processing , Vol. 16, NO. 7, September 2008

[12] Toward Human Level Machine Intelligence—Is It Achievable? The Need for a Paradigm Shift Lotfi A. Zadeh University of California, USA, August 2008 | IEEE Computational Intelligence Magazine, Digital Object Identifier 10.1109/MCI.2008.926583

[13] Machine Translation Inching toward Human Quality Jan Krikke, 1541-1672/06/ © 2006 IEEE Intelligent Systems, Published by the IEEE Computer Society

[14] Milam W. Aiken, Zachary Wong,Spanish-to-English Translation Using the Web

[15] Sudhir Kumar Mishra, Sanskrit Karaka Analyzer for Machine Translation,PhD thesis submitted at Jawaharlal Nehru University 2007

[16] B.Hettige,Karunananda ,Transliteration System for English To Sinhala Machine Translation,Second International Conference on Industrial and Information Systems,ICIIS 2007,8-11 August 2007 ,Sri Lanka.