Design and Development of a Malayalam to English Translator- A Transfer Based Approach

Latha R Nair *latha5074@gmail.com*
Assistant Professor
School of Engineering
Cochin University of Science and Technology
Kochi,Kerala,682022, India

David Peter S *davidpeter@cusat.ac.in*
Professor
School of Engineering
Cochin University of Science and Technology
Kochi,Kerala,682022,India

Renjith P Ravindran
School of Engineering *renjithforever@gmail.com*
Cochin University of Science and Technology
Kochi,Kerala,682022,India

Abstract

This paper describes a transfer based scheme for translating Malayalam, a Dravidian language, to English. The input to the system is a Malayalam sentence and he ouput isits equivalent English sentence. The system comprises of a preprocessor for splitting the compound words, a morphological parser for context disambiguation and chunking, a syntactic structure transfer module and a bilingual dictionary. All the modules are morpheme based to reduce dictionary size. The system does not rely on a stochastic approach and it is based on a rule-based architecture along with various linguistic knowledge components of both Malayalam and English. The system uses two sets of rules: rules for Malayalam morphology and rules for syntactic structure transfer from Malayalam to English. The system is designed using artificial intelligence techniques and can easily be modified to build translation systems for other language pairs.

Keywords: Malayalam Language, Transfer Based Approach, Machine Translation, Morphological Parser.

1. INTRODUCTION

Work in the area of Machine translation in India has been going on for several decades. Promising translation technology began to emerge by 1970 with the developments in the field of artificial intelligence and computational linguistics. Machine Translation Systems in certain well-defined domains have been successfully developed. Translation of gazette notifications, office memorandums,and circulars has been done successfully by Mantra system developed by centre for development for advanced computing(CDAC), Pune. Most of the systems developed are for Hindi, the official language of India. This paper describes a translator for translating sentences in Malayalam a Dravidian Language to English developed on a rule based architecture combined with linguistic knowledge components of both Malayalam and English. The system has a preprocessor for splitting the compound words, morphological parser for context disambiguation and chunking and a bilingual dictionary. A set of rules for Malayalam morphology and rules for syntactic structure transfer from Malayalam to English have been incorporated in the system.

Some of the organizations which are involved in the development of translation systems are: Indian Institute of Technology (Kanpur), Center for Development of Advanced Computing (CDAC) (Mumbai), CDAC(Pune), Indian Institute of Information Technology (Hyderabad) . They are

engaged in development of MT systems under projects sponsored by Department of Electronics, state governments etc. since 1990[1,2]. Research on MT systems between Indian and foreign languages and also between Indian languages are going on in these institutions.

The two major goals in any translation system development wrk are accuracy of translation and speed. Accuracy-wise, smart tools for handling transfer grammar and translation standards including equivalent words, expressions, phrases and styles in the target language are to be developed. The grammar should be optimized with a view to obtaining a single correct parse and hence a single translated output. Speed-wise, innovative use of corpus analysis, efficient parsing algorithm, design of efficient Data Structure and run-time frequency-based rearrangement of the grammar which substantially reduces the parsing and generation time are required [3]. A fully automatic Machine translation system should have different modules such as morphological analyzer, Part of speech tagger, chunker, Named entity recognizer, word sense disambiguator, syntactic transfer module and target word generator [3]. The different techniques used for translation differs in the number of modules used and also the way these modules are implemented. Both rule based and statistical approaches have been tried in the implementation of each of these modules.

The various approaches used in the MT systems for Indian languages are: Direct machine translation systems, Rule based systems and Corpus based systems. Rule based systems do not use any intermediate representation. This is done on a word by word translation using a bilingual dictionary usually followed by some syntactic arrangement. [4, 5,6] 2) Rule based translation which produces an intermediate representation, which may be a parse tree or some abstract representation. The target language text is generated from the intermediate representation. Of the two rule based methods, Interlingua and transfer based approach, transfer based systems are more flexible and it can be extended to language pairs in a multilingual environment. The Interlingua based systems can be used for multilingual translation [7]. The amount of analysis needed in Interlingua approach is more than that in a transfer based approach. The universal networking language has been proposed as the Interlingua by the United Nations University for overcoming the language barrier[8]. Corpus based MT is fully automatic and requires less human labour than rule based approaches. The disadvantage is that they need sentence aligned parallel text for each language pair and this method can not be employed where these corpora are not available [9, 10].

2. PREVIOUS WORK

English to Hindi MT system Mantra, developed by Applied Artificial Intelligence (AAI) group of CDAC, Bangalore, in 1999 uses transfer based approach. The system translates domain specific documents in the field of personal administration; specifically gazette notifications, office orders, office memorandums and circulars. It is based on lexicalized tree adjoining grammar (LTAG) to represent English and Hindi grammar which are used to parse source English sentences and for structural transfer from English to Hindi [2]. This system also works well on other language pairs such as English-Bengali, English-Telugu, English-Gujarati , Hindi-English etc and also between Indian language pairs such as Hindi-Bengali and Hindi-Punjabi. The Mantra approach is general but the lexicon and grammar have been limited to the specific domain of personal Administration. It uses preprocessing tools like phrase marker, named entity recognizer, spell and grammatical checker. It uses Earley's style bottom up parsing algorithm for parsing. The system provides online addition of grammar rule. The system produces multiple translation results in the case of multiple correct parses.

English to Kannada MT system has been developed at Resource centre for Indian Language Technology Solutions (RC_ILTS), University of Hyderabad by Dr. K. Narayan Murthy [2]. This also uses a transfer based approach and it can be applied to the domain of government circulars. The project is funded by Karnataka government. This system uses Universal Clause Structure Grammar (UCSG) formalism [15]. The technique is applied to English_ Telugu translation as well.

Other systems developed using this approach are : Matra- English to Hindi MTS developed by CDAC, Pune, Sakti- English to Marathi, Hindi and Telugu developed by IISc Bangalore and IIIT Hyderabad, Anubaad- English to Bengali developed by CDAC, Kolkata, English to Malayalam MTS developed by Amrita Institute of Technology.

It is found that translation between structurally similar languages like Hindi and Punjabi can be developed easily than translation systems between Indian languages and English which differ in the syntactic structure. The proposed translation system translates Malayalam sentences to English sentences. Since there is a wide difference in English and Malayalam sentences the system needs an additional modules for parsing and syntactic reordering.

3. DEVELOPMENT AND IMPLEMENTATION OF TRANSFER BASED MACHINE TRANSLATION SYSTEM

A transfer based MT system has been developed with the following system modules 1. A preprocessor for splitting the compound words [13] 2. a morphological parser for context disambiguation and chunking 3. A transfer module which transfers the source language structure representation to a target language representation. 4. A generation module which generates target language text using target language structure. Block diagram of the same is shown in fig 1. The grammar rules for Malayalam and some of the transfer rules for transferring source parse tree to target parse tree are stored in two separate files. Some of the transfer rules are embedded in the source code. The sentences stored in a source file are read one by one by the input module and given to the preprocessor module. The final translated output is stored in another file.

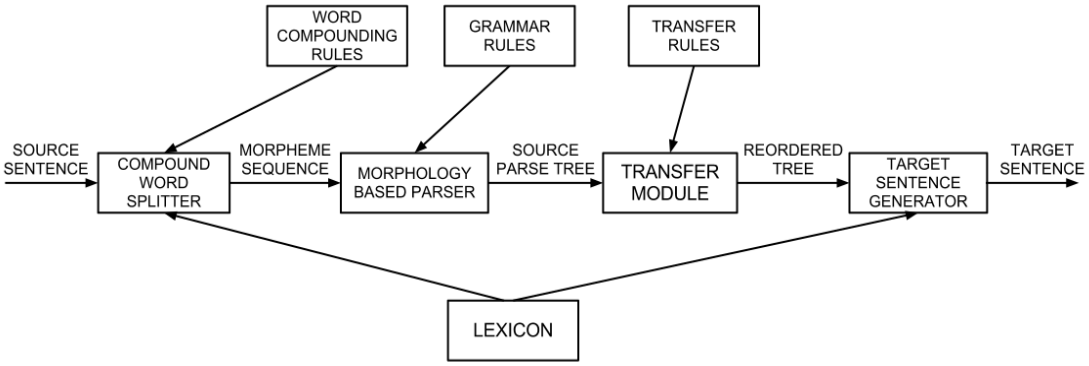


FIGURE 1: Block diagram of a transfer based system

3.1 Compound Word Splitter Module

Morphological variations for words occur in Malayalam due to inflections, derivations and word compounding. Malayalam is an agglutinative language where words of different syntactic categories are combined to form a single word. Formation of new words by combining a noun and a noun, noun and adjective, verb and noun, adverb and verb, adjective and noun and in some cases all the words of an entire sentence to reflect the semantics of the sentence are very common. The complexity of compounding in Malayalam language can be understood from the following example.

സീതയുടെപൂച്ചയൊരെലിയെത്തിന്നു- (1)

The English version being Seetha's cat ate a rat

The constituent words in 1 are to be separated before any further processing. Splitting has been done at morpheme level to reduce dictionary space. The above sentence gets split as shown in 2

സീത ഉടെ പൂച്ച ഒരു എലി എ തിന്നു--(2)

morpheme by morpheme translation for the sentence at 2 is :

Seetha 's cat a mouse (null) ate

The morpheme sequence will be as in 2 above. The sequence of morphemes is given to the parser for chunking and word sense disambiguation. The set of inflectional suffixes for nouns and verbs and derivational suffix for adjectives are based on previous works [11, 12]. Due to the ambiguity in the splitting rules the system generates multiple splits for the same input sentence and the split with least number of constituents is fed to parser.

3.2 Parser Module

Parser takes input from the splitter and does the following tasks. It groups the input sequence of morphemes into chunks [14, 15] and performs word sense disambiguation based on morpheme tags [16]. The chunking process finds the basic units for tree reordering. The word sense disambiguation is required as a morpheme can have multiple tags. The parser uses a depth first approach with backtracking [17]. The output of the parser is a parse tree for the next module. The parser uses the syntax rules for the morpheme sequences in Malayalam sentences in the regular expression form. A set syntax rules in the regular expression form are shown below:

1. S-> NP*VP
2. NP-> ADJ*NP | N NA
3. VP ->ADV* V VA| V VA

Rule 1 implies that a simple sentence is a sequence of noun chunks followed by a verb chunk. Based on the second rule, a noun chunk consists of a set of adjectives followed by a noun and suffixes like case, gender and number for nouns. According to the third rule a verb chunk consist of a sequence of adverbs followed by a verb. Only a subset of such rules derived is shown above. The chunks selected form groups for structural transfer to form target language structure.

A sample sentence and the parse tree generated for the sentence using the grammar rules are shown below:

Input sentence:

മാല മോഷ്ടിച്ച കള്ളന്മാർ രാത്രിയിൽ കാട്ടിലേക്ക് പോയെന്ന് പോലീസ് വിചാരിച്ചു.

English version: The police thought that the thieves who stole the chain went into forest in the night.

Output of the splitter:

മാല മോഷ്ടിച്ച കള്ളൻ മാർ രാത്രി ഇല് കാട് ലേക്ക് പോയി എന്ന് പോലീസ് വിചാരിച്ച

English version:chain stole theif 's night in forest to went that police thought

Output of the parser:

CS(NC(S(NG(ADJC(S(N (മാല) V(മോഷ്ടിച്ച) RP) NG(N(കള്ളൻ) PL(മാർ))) NG(N(രാത്രി) NA(ഇൽ)) NG(N(കാട്)

NA(ലേക്ക്)) V(പോയി)) NCA(എന്ന്)) S(N(പോലീസ്) V(വിചാരിച്ച)))

English version: CS(NC(S(NG(ADJC(S(N (chain) V(stole) RP) NG(N(theif) PL('s))) NG(N(night) NA(in)) NG(N(forest) NA(to)) V(went)) NCA(that)) S(N(police) V(thought)))

The corresponding parse tree generated is shown in Fig.2

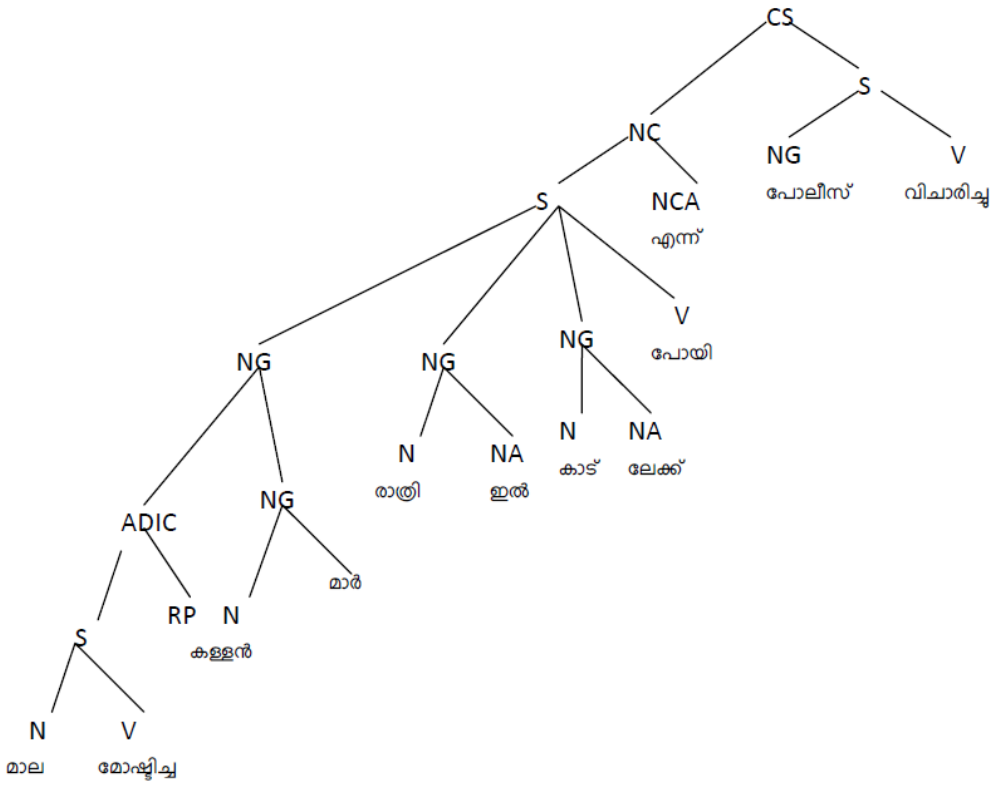
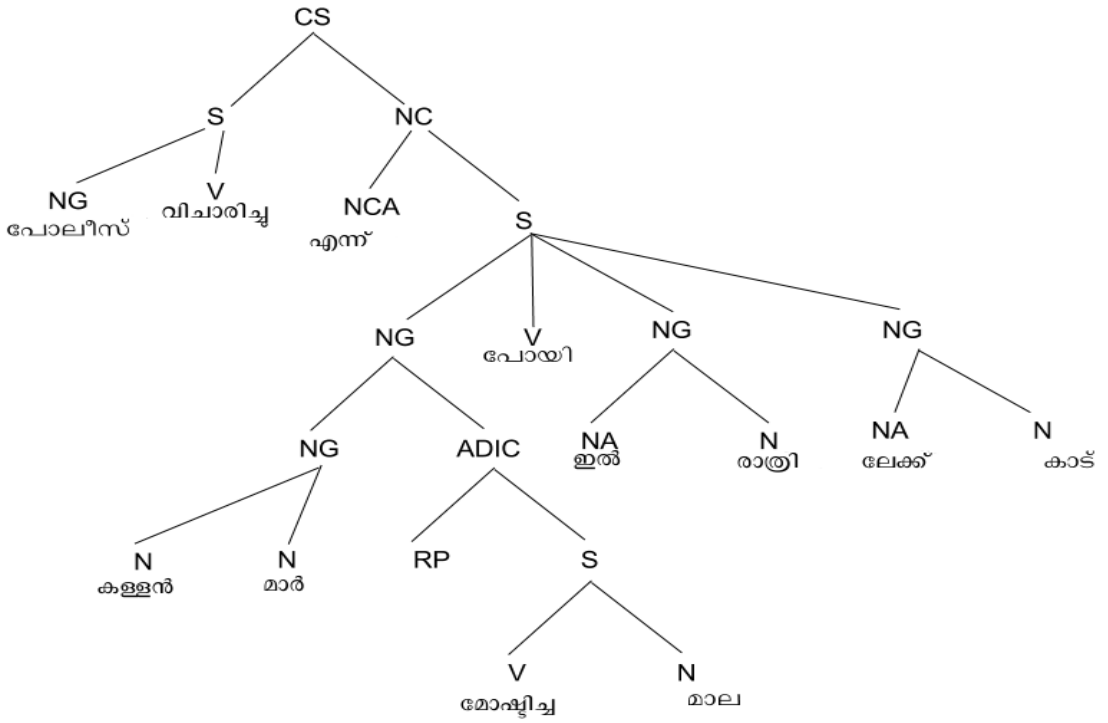


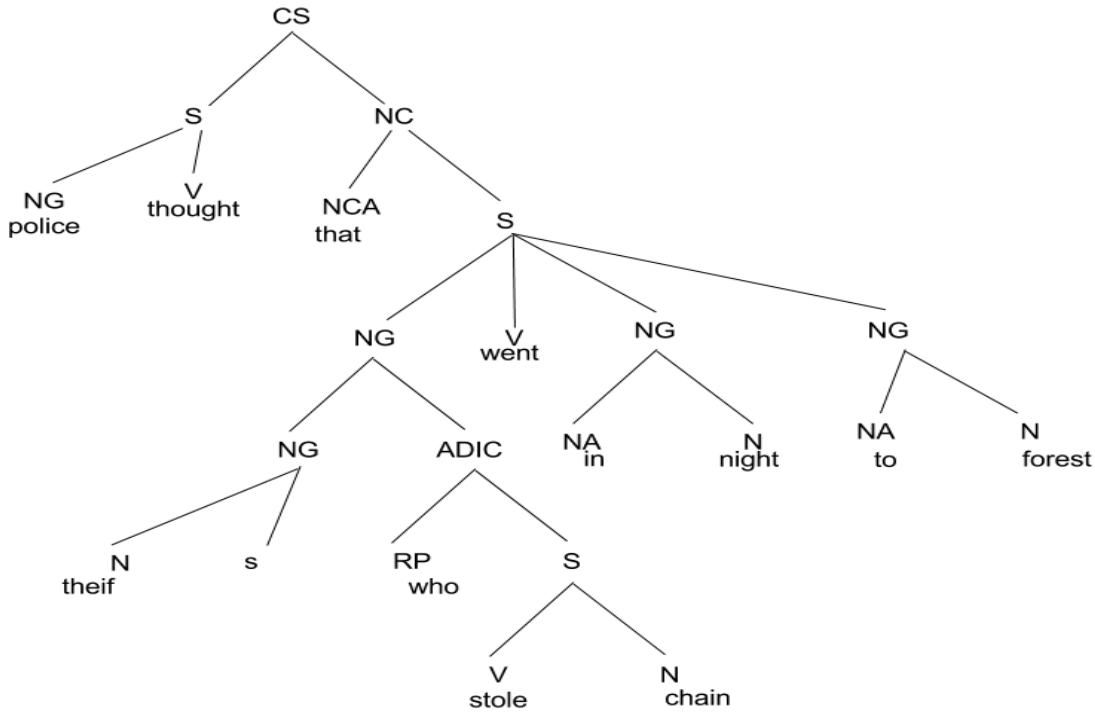
FIGURE 2: Generated Parse tree

3.3 Syntactic Structure Transfer Module

The transfer module transfers the source language structure representation to a target language representation. This module needs the sub tree rearrangement rules by which the source



a



b

FIGURE 3: a. Parse tree after structural transfer b. Corresponding parse tree in English

language sentence syntax tree can be transformed into target language sentence syntax tree. The system performs most of the commonly needed reordering for Malayalam to English translation. The tree after reordering for the above Malayalam sentence using the transfer grammar rules using the transfer rule identified is shown in fig3(a) and its corresponding English tree is shown in fig3(b).A set of transfer rules used by the system are shown in Table I.

	Malayalam structure	English structure
1	PP: NG P	PP: P NG
2	VG: ADV V	VG: V ADV

TABLE 1: A set of system transfer rules

According to the first rule the order of case suffix and noun chunk should be interchanged in a prepositional chunk. The Second rule accords that in verbal chunk the adverb and verb should be interchanged.

3.4 Target Sentence Generator Module

The generation module generates target language text using target language structure [18]. This uses inter chunk dependency rules and intra chunk dependency rules. It involves lexical transfer of verbs, transfer of auxiliary verb for tense, aspect and mood and transfer of gender, number and person information. A depth first traversal of the target parse tree generates the following English sentence

Input Malayalam sentence:

മാല മോഷ്ടിച്ച കള്ളന്മാർ രാത്രിയിൽ കാട്ടിലേക്ക് പോയെന്ന് പോലീസ് വിചാരിച്ചു.

Correct English translation: The police thought that the thieves who stole the chain went into the forest in the night

Sentence generated by the system:

The Police thought that thieves who stole the chain went in night to forest.

3.5 Cross lingual Dictionary

The dictionary includes most of the commonly occurring verbs, nouns, pronouns, adjectives, inflectional and derivational suffixes, clause suffixes etc. Each entry in the file has three fields: the root word (morpheme), the morpheme tag and its translation. The verbs in past tense have their root words stored along with them. Since the system works with morphemes, the space required for the dictionary is less.

Root word	Morpheme tag	Translation
പൂച്ച	Noun	cat
ഉടെ	Case suffix	's

TABLE 2: Lexicon

Presently the system works for sentences which contain upto two adverbial or adjectival clauses which is commonly found in Malayalam texts. The system can be modified to handle other sentences by adding appropriate grammar rules and transfer rules to the rule database. As the parser is a general parser, it can handle sentences of any depth.

3.6. Implementation and Testing of the System

The system was implemented in Python language and tested with a source file which contains 1000 sentences. The sentences which follow the grammar rules were translated. A group of results are tabulated in table 3.

4. RESULTS AND DISCUSSION

The system was tested with more than 1000 different kinds of sentences with and without subordinate clauses which follows the identified morpheme sequences. The system returned correct meaningful translations in most of the cases. A group of sample input sentences with the tabulated outputs are shown in table 3 to give a correct picture of the results obtained..In around 20% of sentences the system returned the exact English version of the input sentences. In balance translations the output sentences were meaningful but had small shortcomings due to the following reasons:

i) The positioning of articles is not considered.
ii) Many inter chunk and intra chunk dependencies are not considered.
iii) The lexicon stores only the common translation for polysemous words.

The system takes care of word sense disambiguation based on lexical category successfully. The compound nouns are also not handled by the system as the shallow parser cannot group them using the current set of rules. The system output can be enhanced including rules which can take care of the above shortcomings.

5. CONCLUSION

Various MT groups have used different formalisms best suited to their applications. Of them transfer based systems are more flexible and it can be extended to language pairs in a multilingual environment. A transfer based MT system has been developed for Malayalam, a Dravidian Language which comprises of a preprocessor for splitting the compound words, a morphological parser for context disambiguation and chunking, a syntactic structure transfer module and a bilingual dictionary. The system was tested successfully for more than 1000 different types of sentences wherein the system returned true results for sentences which contain two subordinate clauses. Even for sentences with more than two subordinate clauses the system

System Input and Output	Remarks
1. Input: ഒരു നല്ലവനായ രാജാവ് ഒരിടത്തൊരിടത്ത് ഉണ്ടായിരുന്നു. **Word to word translation:** a kind king in a place was. **English version of the sentence:** There was a kind king in a place.	System output is in line with the English version except for the positioning of the article .Meaningfully correct sentence
System Output: A kind king was in a place.	
2.Input:. രാജാവിന്റെ മകൾ അതിസുന്ദരി ആയിരുന്നു. **Word to word translation:** King's daughter very beautiful was **English version:** The king's daughter was very beautiful.	System output is exactly in line with the English version except for the positioning of article to the noun king since same word is not there in input language. Meaningful correct translation
System Output: King's daughter was very beautiful.	
3.Input: ആ രാജകുമാരിക്ക് സൂര്യനെപ്പോലെ തിളങ്ങുന്ന ഒരു സ്വർണ്ണപന്തുണ്ടായിരുന്നു. **Word to word translation:** That princess sun like shining is a goldenball. English **version** : The princess had a golden ball which was shining like the sun.	System output gives translation without positioning of article for the nouns. The variations in translations for ആ have not been considered. System output is meaningful correcttranslation
System Output: That princess had a ball which is shining like sun.	
4.Input: രാജകുമാരിക്ക് പൂന്തോട്ടത്തിൽ പന്തു കളിക്കാൻ ഇഷ്ടമായിരുന്നു. **Word to word translation:** Princess garden in ball play to liked **English Version** : The princess liked to play in the garden	Output gives translation without positioning of preposition since same word is not there in input language Meaningful correct translation
System Output: Princess liked to play in garden.	
5.Input: രാജകുമാരി ഒരു ദിവസം കളിച്ചുകൊണ്ടിരുന്നപ്പോൾ സ്വർണ്ണപന്ത് അടുത്തുള്ള കിണറ്റിൽ വീണു. **Word to word translation:** princess a day play was when golden ball nearby well in fell. **English Version** : When the princess was playing one day the golden ball fell into a nearby well.	Output translation is without positioning of presposition since same word is not there in input language. Multiple translations of ഒരു has not been considered. Meaningful correct translation
System Output: When princess was playing a day golden ball fell in nearby well.	
6.Input: രാജകുമാരി കരയാൻ തുടങ്ങി. **Word to word translation** :Princess cry to started **English version** : The princess started to cry	System output gives exact English version of input sentence
System Output: Princess started to cry.	
7.Input: കിണറ്റുവക്കിലിരുന്ന ഒരു തവള ഇതെല്ലാം കാണുന്നുണ്ടായിരുന്നു **Word to word translation:** well side in sat a frog all these see was **English version:** A frog sitting beside the well was seeing all these.	Meaningful correct translation which slightly varies from the English version.
System Output: frog which was sitting in side of well was seeing all these.	
8.Input: തവള രാജകുമാരിയുടെ അടുത്തേക്ക് ചാടി ചെന്നു. **Word to word translation:** frog princess's side jumped went **English version** : The frog jumped to the princess.	Meaningful correct translation which slightly varies from the English version.
System Output: frog went jumping to princess's side.	

TABLE 3: Group of Tabulated results

returned translated output sentences which could give basic understanding of the input sentences. More rules can be added to make the system to give exact translation of input sentences in all cases. Additional modules like finding and replacing collocations, finding and replacing named entities can also be added to the basic translator. The results obtained are encouraging. The work can be extended to create a full fledged machine translator from any Dravidian language to English since they all exhibit structural homogeneity.

6. REFERENCES

[1] P Dubey et al. "Overcoming the Digital Divide through Machine Translation". *TranslationJournal.,* Vol.15, 2011, http://translationjournal.net/journal/55mt_india.htm [Dec 12, 2011].

[2] B.K.Murthy, W.R Deshpande ., "Language technology in India: past, present and future", 1998,

[3] http://www.cicc.or.jp/english/hyoujyunka/mlit3/7-12.html [Dec 11,2011]

[4] S.Lalithadevi, P.Pralayankar , V.Kavitha. "Translation of Hindi se to Tamil in a MT System". Information systems for Indian languages, Berlin Heidelberg: Springer-Verlag, 2011, pp. 246–249.

[5] V Goyal, G S Lehal. "Advances in Machine Translation Systems". *Language In India,* Vol. 9, No. 11, 2009, pp. 138-150 .

[6] G.S.Josan , G.S. Lehal. "Evaluation of Hindi to Punjabi Machine Translation System". *International Journal of Computer Science Issues*, vol4 no1, 2009, pp 243-257.

[7] V.Goyal, G.S. Lehal . "Web Based Hindi to Punjabi Machine Translation System". *Journal of Emerging Technologies in Web Intelligence.* Vol.2., 2010, pp.148-151.

[8] S.K.Goutam. "The EB-Anubad translator: A hybrid scheme". *Journal of Zhejiang University Science*, Vol.6, 2005, pp.1047-1050.

[9] D.S Parikh P.Bhattacharyya "Interlingua Based English Hindi Machine Translation and Language Divergence", *Machine Translation* , Vol.16, 2001, pp.251-304.

[10] R.M.K. Sinha. "A hybridized EBMT system for Hindi to English Translation". *CSI Journal*, volume 37 no. 4, 2007, pp.3-9.

[11] 10. R.M.K. Sinha. "Designing Multi-lingual Machine- Translation System: Some Perspectives". International Conference on Machine Learning: Models, Technologies & Applications (MLMTA 2007), 2007 , pp. 244-249.

[12] 11. S..M Idicula, D. S. Peter. "A morphological processor for Malayalam language". *South Asia Research*. vol. 27 (2): 2007, pp.173-186.

[13] 12. L. Pandian, T.V.Geetha "Morpheme based Language Model for Tamil Part of Speech Tagging" *Polibits* (38) , 2008, pp.19-26 .

[14] 13. L.R.Nair, D.S. Peter. "Development of a rule based learning system for splitting compound words in Malayalam language". IEEE Recent advances in intelligent and computational systems(RAICS), 2011, pp.751-755.

[15] 14. L.R.Nair, D.S. Peter, “Shallow parser for Malayalam Language using finite state cascades”, 4th international congress on image and signal processing, China, 2011, pp.2464-2467.

[16] 15. S. Abney. “Partial parsing via finite state cascades”. *Journal of Natural Language Engineering*, **2**(4), 1996, pp. 337-344.

[17] 16. D.Jurafsky ,J.H Martin. *Speech and natural language processing.* India: Pearson Education, 2000,pp 657-671.

[18] 17. E. Rich, K. Knight, S. B Nair. *Artificial Intelligence*. New Delhi,India: The Tata McGraw Hill, 2009 pp 295- 300.

[19] 18. S.L.Devi, P.Pralayankar. “Verb Transfer in a Tamil to Hindi Machine Translation System”. International Conference on Asian Language Processing. Harbin, China, 2010, pp.261-264.