Chapter 1

INTRODUCTION

The various programmes developed for promoting use of IT in Indian languages have helped to generate expert manpower in diverse areas to offer solutions in Indian languages. As part of these programmes, work in the area of some of the foreign languages has also been carried out to provide solutions supporting those languages on personal computers. Research in language computing is progressing in many areas such as:

- machine translation
- speech processing
- optical character recognition (OCR)
- standards (character representation, fonts display, etc)
- localization
- development of applications like word processors, e-mail clients, etc.
- information extraction and retrieval
- search engines

In a large multilingual society like India, there is great demand for translation of documents from one language to another language. There are 22 constitutionally approved languages, which are officially used in different states. There are about 1650 dialects of these languages that are spoken by
different communities. There are 10 Indic scripts. All of these languages are well developed and are rich in content. They have similar scripts and grammars. The alphabetic order is also similar. Some languages use common script, especially Devanagari. Hindi written in the Devanagri script is the official language of the Union Government. English is also used for government notifications and communications. Even though India's average literacy level is 65.4 percent (Census 2001) less than 5 percent of people can either read or write English. As most of the state government works are in provincial languages whereas the central government’s official documents and reports are in English or Hindi, these documents are to be translated from and to the respective provincial languages to have an appropriate communication. Moreover, over 95 percent of the population is normally deprived of the benefits of information technology due to language barrier [1]. All these make language translation a necessary one.

Work in the area of Machine Translation in India has been going on for several decades. During the early nineties, advanced research in the field of artificial intelligence and computational linguistics made a promising development of translation technology. This helped in the development of usable machine translation systems in certain well-defined domains. It is extremely difficult to build a fully automatic high quality machine translation system (FGH_MT). In fact there is no system in the world which qualifies to be called FGH_MT. Many institutions like IIT(Kanpur), CDAC(Mumbai), CDAC(Pune), IIIT(Hyderabad), etc. are engaged in the development of MT systems under projects sponsored by Department of Electronics, state governments etc. since 1990 [2]. Research on MT systems between various Indian and foreign languages and also between Indian languages are going on in these institutions. Translation between structurally similar languages like Hindi and Punjabi is
easier than that between language pairs that have wide structural difference like Hindi and English. Having many parts of their grammars and vocabularies in common, it is easy to develop translation systems between closely related languages [3].

Development of Machine translation (MT) system requires very close collaboration among linguists, professional translators and computer engineers. In the development process, there are two major goals; (a) accuracy of translation and (b) 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. Innovative use of corpus analysis, efficient parsing algorithm, design of efficient data structure and run-time frequency-based rearrangement of the grammar are required to reduce the parsing and generation time [4].

1.1 Earlier works in Malayalam Language processing

C-DAC, Thiruvananthapuram (formerly ER&DC) is a pioneer institution involved in Malayalam language tool development. It is one of the thirteen resource centers for Indian language technology solutions set up across the country by the Ministry of Communications and Information Technology, Govt. of India under the TDIL (Technology Development for Indian Languages) programme.

Some of the Tools for Malayalam which C-DAC could develop so far include:

1. Tools such as portal, fonts, morphological analyzer, spell checker, text editor, search engine and code converters.
2. Knowledge resources like Malayalam corpora, trilingual (English-Hindi-Malayalam) online dictionary and knowledge bases for literature, art and culture of Kerala.

3. Human machine interface systems comprising of optical character recognition and text to speech systems.

4. Services like E-commerce application and E-mail server in Malayalam and

5. Language tutors for Malayalam and English.

1.2 Motivation for the work

Language translation between Malayalam and other languages can be said to be in very primitive stage now. There has been hardly any work done in the area of machine translation from Malayalam to English. It is also found that MT systems developed for other language pairs need improvements in:

- rule set
- dictionary design
- translation methodology

1.3 Research Objective

The main objective of this research is to study the issues in Malayalam to English translation and the development of a prototype system for Malayalam to English translation. The major goals in the design process are a) high accuracy of translation b) fast processing and c) low space requirement.
1.4 Research tasks

Major tasks include:

1. Study of the morphology of various lexical categories of Malayalam and English.

2. Identification of the part of speech tag set and chunk tag set for word sense disambiguation and structure transfer.

3. Derivation of syntactic structure of Malayalam and English sentences using morphology and word level dependencies.

4. Determination of the syntactic structure differences between Malayalam and English sentences.

5. Design of efficient lexicon in terms of speed of retrieval, space requirement and speeding up of the translation process.

6. Improving speed and accuracy of the translation process.

1.5 Conclusion

This chapter gave an introduction about the research work. The motivation behind the research work, the major objectives of the research and the tasks involved in the work were clearly described. The following chapter lists the various MT approaches and major MT systems developed in India using these approaches. Out of the various approaches listed one which is best suited for the language pair is chosen.
Chapter 2

APPROACHES FOR MACHINE TRANSLATION

2.1 Introduction

This chapter describes the various approaches used for machine translation. Researchers used different formalisms that are best suited to their applications. They are mainly i) Direct translation approach ii) Rule based approach and iii) Corpus based approach. Direct translation is appropriate for structurally similar languages. Among the rule based approaches transfer based systems are more flexible and it can be easily extended to language pairs in a multilingual environment. The interlingua based systems can be used for multilingual translation. The Universal Networking Language has been proposed as the interlingua by the United Nations University for overcoming the language barrier. Over the past decade data-driven approaches to machine translation have come to the fore of language processing research. Hybrid systems are found to have better performance compared to the ones with the component technology.

2.2 Direct machine translation systems.

As the name suggests, these systems provide direct translation, without using any intermediate representation. This is done on a word by word translation using a bilingual dictionary usually followed by some syntactic arrangement. The steps involved are:

a. Identification of root words by removing suffixes from source language words.
b. Dictionary look up to get the target language words/morphemes.

c. Rearrange the word order to match the target language. For English to Malayalam, this may be reordering of prepositions to postpositions and changing subject-verb-object to subject–object–verb structure.

A sample output from such a system for English to Hindi is given below:

Input(English): Rama played in the garden.

Output (Hindi) after word translation: Rama khela maim baag.

Output after syntactic rearrangement: Rama baag maim khela.

A direct translation system is appropriate for similar languages like Hindi and Punjabi [5, 6]. Vishal Goyal and Gurpreet Singh Lehel of Punjab University have developed a web based Hindi to Punjabi MT system with 95% accuracy. Their system has additional modules for training the system for generating the lexicon using already existing corpus, input text font conversion into Unicode format to make the system free from specific font dependency, Hindi text normalization to handle spelling variations for the same word due to variation in dialects, replacement of collocations by keeping a lexicon for collocations, named entity recognition and replacement, word by word translation using bilingual dictionary and transliteration of unknown words. They also perform word sense disambiguation using a dictionary of ambiguous words. It uses a trigram approach with a sample corpus for word sense disambiguation.

A similar direct translation approach has been applied to Anusaraka systems which translate between two closely related Indian languages using the principles of paninian grammar [7]. The Anusaraka project started at IIT
Kanpur, by Prof. Rajeev Sangal and the research is continued at IIIT Hyderabad. Anusaarakas have been constructed from Telugu, Kannada, Bengali, Punjabi and Marathi to Hindi. Anusaraka systems or the language accessors are based on the principle of 'information preservation'. As a consequence, the Anusaaraka output follows the grammar of source language. Hence before using Anusaaraka systems to access the information, the reader has to undergo a short training to read and understand the output. Anusaaraka provides 'glosses' in target language for each meaningful lexical unit. There are cases where the meaning is too general or too specific. Such cases are handled by introducing some special notation to either narrow down or widen the meaning. An attempt is made to find the underlying thread (called 'shabda sutra' or 'wordthread’) that connects different senses of the polysemous word. A kind of formula ('sutra’ also means a formula in Sanskrit) is then evolved that faithfully and unambiguously represents the connection between these different senses. The core Anusaaraka engine has four major modules viz.

(i) Word level substitution
(ii) Word sense disambiguation
(iii) Preposition placement, and
(iv) Target language generation

The output generated by the system may not be grammatically perfect and can be understood by a reader after some training. A sample output of an Anusaaraka system from Telugu to Hindi is shown below:

Input(Telugu): mIru pustakam caduvutunnArA
Output(Hindi): aap pusthak paTh raha [hai|thha] kya ?
The translation is done in a morpheme by morpheme basis. The tense, aspect and modality information of the verb from the source sentence is extracted and is translated as raha [hai|thha] kya?. The possible suffixes for tense information [hai/thha] are given and the user has to take the correct one based on the context.

Anusaraka based on a new architecture together with a user-friendly interface is convenient for a user-cum-developer. This new architecture makes a clear-cut distinction between the resources that are in principle reliable and those that are in principle probabilistic.

2.3 Rule based translation

Rule based machine translation (RBMT) systems parse the source text and produce an intermediate representation, which may be a parse tree or some abstract representation. The target language text is generated from the intermediate representation. These systems rely on specification of rules for morphology, syntax, lexical selection and transfer, semantic analysis and generation and hence are called rule based systems. Depending on the intermediate representation used, these systems are further categorized as Transfer based machine translation and Interlingua based machine translation.

2.3.1 Transfer based machine translation

This method needs parsing of input text to get the structure of the input sentence. It has three modules: analysis module, transfer module and generation module [8,9]. The analysis module produces source language structure. The language grammar rules can be used to generate the hierarchical syntax tree for
Approaches for Machine Translation

the source language sentence. A set of hierarchical rules for forming the syntax tree for English sentences are given below:

\[ S \rightarrow \text{NP} \text{VP} \mid \text{VP} \]
\[ \text{NP} \rightarrow \text{N} \mid \text{NP PP} \mid \text{Det N} \]
\[ \text{VP} \rightarrow \text{V} \mid \text{VP PP} \mid \text{VP NP} \]

The transfer module transfers the source language structure representation to a target language structure representation. This module needs the subtree rearrangement rules by which the source language sentence syntax tree can be transformed into target language sentence syntax tree.

A set of transfer rules for a English to HindiMT system are:

<table>
<thead>
<tr>
<th>English structure</th>
<th>Hindi structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP-&gt;V NP</td>
<td>VP-&gt;NP V</td>
</tr>
<tr>
<td>PP-&gt;P NP</td>
<td>PP-&gt; NP P</td>
</tr>
<tr>
<td>VG-&gt;ADV V</td>
<td>VG-&gt;V ADV</td>
</tr>
</tbody>
</table>

The generation module generates target language text using target language structure. Syntactic structure transfer for verbs from Tamil to Hindi is discussed in [10,11]. It involves lexical transfer of verbs, transfer of auxiliary verb for tense, aspect and mood and transfer of gender, number and person information. Syntactic and lexical ambiguities are better resolved in this approach than in
direct translation approach. An example for structure transfer from English to Hindi is shown below:

Input(English): Ram sat on a chair.

Output(Hindi): Ram ne ek kursi mein baita.

We have used two transfer rules. One is to switch the preposition and noun in the PP chunk and the other is to place the verb at the end of the sentence. The structural transfer and Hindi generation are shown in Figure 2.1. Transfer systems are more realistic, flexible and adaptable in meeting the needs of different levels and depths of syntactic and semantic analysis.

![Figure 2.1 Structural transfer in a English-Hindi MT system](image)

The English to Hindi MT system Mantra, developed by Applied Artificial Intelligence(AAI) group of CDAC, Bangalore, in 1999 uses transfer based
Approaches for Machine Translation

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 [12,13,14,15,16]. This system also works well on other language pairs such as English-Bengali, English_Telugu, English_Gujarati and Hindi_English 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. The salient features of Mantra are : i) format retention ii) input can be in .rtf or .htm file or output of a speech recognition program or an optical character recognition package. 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. Mantra is one of CDAC’s major achievements.

An English to Kannada MT system is developed at Resource centre for Indian Language Technology Solutions (RC_ILTS), University of Hyderabad by Dr. K. Narayan Murthy [12]. 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 [17]. The technique is applied to English_ Telugu translation as well.

Other systems 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. Some of the MT systems for non Indian languages using this approach are Apertium, interNOSTRUM etc. [18,19,20,21]

2.3.2 Interlingua based machine translation

The interlingua approach was inspired by Chomsky’s claim that regardless of varying surface syntactic structures, languages share a common deep structure [22]. In this approach, translation is a two step process: analysis and synthesis. During analysis, the source language text is converted into a language independent meaning representation called interlingua. In synthesis phase the interlingual representation is translated into any target language. Thus it can be used for multilingual translation. The amount of analysis needed in interlingual approach is more than that in a transfer based approach. This requires semantic analysis and the representation can be used for information retrieval.

For the sample sentence “Raman played in the garden”, the interlingua generated is,

(play (tense: past)
   (Mood declarative)
   (subject (Raman(number singular)))
   (location(garden(reference definite)
       (number singular))))
The major difficulty in using this approach is in defining a universal interlingua which preserves the meaning of a sentence.

Anglabharti is an MT system for translation from English to Indian languages which uses pseudo interlingua approach. It was developed in 1991 by Prof. R.M.KSinha and his team at IIT Kanpur. The system analyses English sentences and creates an intermediate structure called PLIL(Pseudo Lingua for Indian Languages). It performs most of the disambiguation. The effort required for analysis phase is 70% and the generation phase takes 30%. So with an additional effort of 30% a new translator for another Indian language could be built. A context free grammar like structure is used to create the PLIL structure. It also uses statistical analysis of a corpus to identify the movement rules for the PLIL structure. Its beta_version is Angla Hindi for English to Hindi translation and is available at http://anglahindi.iitk.ac.in. [23,24].

The World Wide Web contents are mostly in English and cannot be accessed without proficiency in this language. The universal networking language has been proposed by the United Nations University for overcoming the language barrier. An English to Hindi MT system which uses Universal Natural Language (UNL) as the interlingua has been developed by Pushpak Bhattacharya at IIT Bombay. This system has an English analyzer which converts the sentence into UNL form which is then given to a Hindi generator which generates the target sentence in Hindi. 95% of the UNL expressions were correctly converted to Hindi. It also does part of speech disambiguation and sense disambiguation for postposition markers and wh_pronouns. The system handles language divergence in a better way [22]. Currently on an MT system for English to Marati and Bengali is in progress.
Chapter 2

2.4 Corpus based Machine Translation

Corpus based MT systems have gained much interest in recent years. The advantage of these systems are that they are fully automatic and require 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. Corpus based systems are classified into statistical machine translation(SMT) and Example based Machine Translation(EBMT) [23,24,25].

2.4.1 Statistical Machine Translation

In this the input is considered as a distorted version of the target language sentence and the task is to find the most likely source language sentence given the translation.

The task involves three steps:

i) Estimating the language probability $P(t)$

ii) Estimating the translational model probability $p(s/t)$

iii) Devising an efficient search for the target text that maximizes their product.

We have to find the sentence $T$ for which $p(s,t)$ is maximum.

$$P(s,t)= \arg \max_t p(s,t) = \arg \max_t p(t)p(s/t)$$

In the above model $s$ is the source language sentence and $t$ is the target language sentence. The probabilities are to be calculated from the parallel corpus.
Smoothing techniques are required for handling data sparsity problem that occurs in any noisy channel model.

A phrase-based Hindi-English translation system was tried by Kamal Kuzhinjedathu and ShravyaShetty at Department of Computer Science, State University of New York. The translation model was generated using a Hindi-English parallel corpus. Since the parallel corpus is only sentence aligned the freely available online tool called GIZA++ was used to perform word alignment. The alignment produced were then processed to create the phrase based translation model. Two sets of parallel corpuses were used: EMILE (Enabling minority language engineering) corpus distributed by the European Language Resources Association and the Hindi and English bibles from www.Hindibible.org. Bilingual dictionaries available on the internet were used to augment the statistical model. The SRI language model toolkit was used to obtain a statistical language model. After getting translation model and language model a phrase based decoder, Moses was used to translate the test sentences.

An English to Hindi MT system which combines RBMT and phrase-based SMT approach was developed at IIIT Hyderabad in 2010. Though SMT systems are able to handle local re-orderings by themselves, in case of long-distance transformations they benefit considerably from external guidance (the RBMT system in this case). The system works in two stages. In the first stage, the source analyzer performs extensive linguistic analysis by running Brill’s POS tagger and the Stanford dependency parser on the input sentence. It then converts the source into a chunk-based unordered dependency tree. In the next stage, the Transfer Grammar performs local and long-distance re-orderings. By chunking the source sentences and converting them into a dependency structure,
the RBMT system separates local (intra-chunk) reordering decisions from global (inter-chunk) re-orderings. This allows for separate specifications of local and long-distance rules; thus, greatly reducing the number of rules that must be written into the grammar [26].

Translation of English into the Dravidian language, Malayalam SMT system has been tried in Cochin University of Science and Technology, Cochin, in 2010. By using a monolingual Malayalam corpus and a bilingual English/Malayalam corpus in the training phase, the machine automatically generates Malayalam translations of English sentences. The alignment model is improved by incorporating the parts of speech information into the bilingual corpus. Removing the insignificant alignments from the sentence pairs by this approach has ensured better training results. Pre-processing techniques like suffix separation from the Malayalam corpus and stop word elimination from the bilingual corpus are carried out for effective training. The structural difference between the English - Malayalam pair is resolved by a decoder using the order conversion rules [27].

2.4.2 Example Based Machine Translation (EBMT)

An Example based Machine translation (EBMT) system maintains a corpus consisting of translation examples between source and target languages. An EBMT system has two modules: Retrieval module and an adaptation module. The retrieval module retrieves a similar sentence and its translation from the corpus for the given source sentence. The adaptation module then adapts the retrieved translation to get the final corrected translation.

Consider the English to Hindi translation for the following sentence:
Rama sings a song.

The retrieval module retrieves the following sentence and its translation from a list of approximately matching sentences in the corpus. It uses some similarity measures based on word similarity or syntactic and semantic similarity to identify this set of approximately matching sentences. From these the system selects the sentence with closest match with the input sentence.

If the system selects “Rohit sings a song” and its translation “Rohit geet gaathai hai” as the closest one, it replaces Rohit with Rama and gaata with gaathi and finally forms the translation

Rama geet gaati hai

Here the adaptation is required to replace the word and suffix replacements. This method may not work in case of translation divergence where structurally similar sentences of the source language get translated into a different structure [28].

Anubharti, is an EBMT approach based MT system developed at IIT Kanpur by Prof. R.M.K Sinha and his associates. Along with basic EBMT it uses some grammatical analysis to reduce the size of the parallel corpus. This is done primarily by generalizing the constituents and replacing them with abstracted form achieved by identification of syntactic groups from the raw examples [29,30].

Vaasaanubaada is another system for translating bilingual Bengali-Assamese news texts using EBMT technique. The work involves machine translation of bilingual texts at sentence level. In addition, it also includes preprocessing and post-processing tasks. The work is unique because of the language pair that is
chosen for experimentation. The bilingual corpus was constructed and aligned manually by feeding real examples using pseudocode. The longer input sentence is fragmented at punctuations, which resulted in high quality translation. Backtracking is used when an exact match is not found at the sentence/fragment level, leading to further fragmentation of the sentence. Since bilingual Bengali-Assamese languages belong to the Magadha Prakrit group, the grammatical form of sentences is very similar and has no lexical word groups. The system gives quality translation [31].

Siva, an English to Hindi MT system developed jointly by Carneige Mellon University, USA, IISc Bangalore and IIIT Hyderabad is using EBMT approach. In addition to the hard coded linguistic rules it uses a statistical approach for learning new rules [15]. An English to Sanskrit EBMT has been tried at Banaras Hindu University. Many hybrid systems which combines rule based approach and corpus based approach has been tried and has produced encouraging results [32,33,34,35,36]. The systems developed for Indian languages are listed in Table 2.1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Name of the MT system and Language Pair</th>
<th>Developer</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>Anusaarak systems(Telugu,Kannada,Bengali,Marathi to Hindi)</td>
<td>Started in IIT Kanpur, continuing in IIIT, Hyderabad</td>
<td>1995</td>
</tr>
<tr>
<td></td>
<td>Punjabi to Hindi MTS</td>
<td>PunjabUniversity, Patiala</td>
<td>2007</td>
</tr>
<tr>
<td></td>
<td>Web based Hindi to Punjabi MTS</td>
<td>PunjabUniversity, Patiala</td>
<td>2010</td>
</tr>
<tr>
<td>Method</td>
<td>Name of the MT system and Language Pair</td>
<td>Developer</td>
<td>Year</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Rule Based</td>
<td>Mantra-English to Hindi, Telugu,Gujarathi,Hindi-English</td>
<td>CDAC,Pune</td>
<td>1995</td>
</tr>
<tr>
<td>Transfer based</td>
<td>Matra-English to Hindi MTS</td>
<td>CDAC,Pune</td>
<td>2004</td>
</tr>
<tr>
<td></td>
<td>Shakti-English to Hindi,Marathi,Telugu MT</td>
<td>IIISc Bangalore and IIIT, Hyderabad</td>
<td>2004</td>
</tr>
<tr>
<td></td>
<td>Anubaad-English-Bengali MTS(n-gram approach for pos tagging)</td>
<td>CDAC, Kolkata</td>
<td>2005</td>
</tr>
<tr>
<td></td>
<td>English to Kannada MTS using UCSG</td>
<td>University of Hyderabad</td>
<td>2006</td>
</tr>
<tr>
<td></td>
<td>English – Malayalam MTS</td>
<td>Amrita Institute of Technology</td>
<td>2009</td>
</tr>
<tr>
<td>Interlingua</td>
<td>Anglabharti, English-Hindi,Tamil MTS; uses pseudo interlinguaPLIL</td>
<td>IIT,Kanpur</td>
<td>1991</td>
</tr>
<tr>
<td></td>
<td>AnglaHindi(combines example based approach and AnglaBharti approach)</td>
<td>IIT,Kanpur</td>
<td>1991</td>
</tr>
<tr>
<td></td>
<td>English to Hindi MTS Using UNL</td>
<td>IIT, Mumbai</td>
<td>2001</td>
</tr>
<tr>
<td></td>
<td>Hindi-English MTS (Additional layer over AnglabhartiII)</td>
<td>IIT,Kanpur</td>
<td>2005</td>
</tr>
<tr>
<td>SMT</td>
<td>Tamil-Sinhala MTS</td>
<td>Carnegie-Mellon University, Pittsburgh, USA</td>
<td>2002</td>
</tr>
<tr>
<td></td>
<td>English-Hindi MTS</td>
<td>IIIT, Hyderabad</td>
<td>2010</td>
</tr>
<tr>
<td></td>
<td>English- Malayalam MTS</td>
<td>Cochin University ,Cochin</td>
<td>2010</td>
</tr>
<tr>
<td></td>
<td>Hindi-English MTS</td>
<td>StateUniversity of New York.</td>
<td>2010</td>
</tr>
<tr>
<td>EBMT</td>
<td>VAASAAANUBAADA - Bengali-Assamese</td>
<td>PondicheryUniversity</td>
<td>2002</td>
</tr>
<tr>
<td></td>
<td>Anubharti; Hindi-English MTS</td>
<td>IIT,Kanpur</td>
<td>2004</td>
</tr>
<tr>
<td></td>
<td>Siva- English-Hindi</td>
<td>IIIScBangalore,IIITHyderabad</td>
<td>2004</td>
</tr>
<tr>
<td></td>
<td>English-Sanskrit</td>
<td>I.T, Banaras Hindu University</td>
<td>2008</td>
</tr>
</tbody>
</table>
2.5 Conclusion

This chapter explains in detail the various approaches tried by different researchers in this field. A number of Machine Translation systems between Indian and non-Indian languages have already been developed. Most of the MT systems are for Hindi and there are only very few systems for south Indian languages. More research has to done in these areas to overcome the language barrier faced by India. The MT systems developed has many shortcomings in terms of rule set, dictionary, translation methodology and it is apparent from the survey that further work is needed in MT as a whole to produce intelligible translations. The amount of analysis needed in interlingual approach is more than that in a transfer based approach. From among the various approaches the transfer based approach has been chosen due to its flexibility and extensibility to other language pairs in a multilingual environment.