BILINGUAL CORPORA CREATION FOR SANSKRIT LANGUAGE USING NLP

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

ANUJ KUMAR

11506651

Supervisor

Mr. PRATEEK AGRAWAL

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

MAY 2017

PAC Approval Page

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE546 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0003

Supervisor Name : Prateek Agrawal **UID :** 13714 **Designation :** Assistant Professor

Qualification : M.Tech. **Research Experience :** 7.8 Years

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Anuj Kumar	11506651	2015	K1518	9023838213

SPECIALIZATION AREA : Intelligent Systems **Supervisor Signature:**

PROPOSED TOPIC : Bilingual Corpora creation for Sanskrit language using NLP

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10
1	Project Novelty: Potential of the project to create new knowledge	7.80
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.80
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.80
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	8.60
5	Social Applicability: Project work intends to solve a practical problem.	8.00
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	8.40

PAC Committee Members		
PAC Member 1 Name: Prateek Agrawal	UID: 13714	Recommended (Y/N): Yes
PAC Member 2 Name: Pushpendra Kumar Pateriya	UID: 14623	Recommended (Y/N): Yes
PAC Member 3 Name: Deepak Prashar	UID: 13897	Recommended (Y/N): Yes
PAC Member 4 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 5 Name: Anupinder Singh	UID: 19385	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): Yes

Final Topic Approved by PAC: Bilingual Corpora creation for Sanskrit language using NLP

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11011::Dr. Rajeev Sobti **Approval Date:** 26 Oct 2016

4/25/2017 12:41:00 PM

ABSTRACT

Language in any form is the fundamental requirement to communicate and interact within the human society. In this globalization era, we interact with people from different regions and linguistic backgrounds as per our interest in social, cultural, economic, educational and professional domain. It is quite tough, rather impossible to know all the languages. Thus we need a computerized approach to convert one natural language to another as per the necessity. The implementation of machine translation system for natural language such as Sanskrit is a very challenging task. Sanskrit language has richness of morphological analysis, so we use morphological analysis to identify noun, sarvnamm, verb, avyaya sentences. This research work creates Corpora for Sanskrit to Hindi approach for translating well-structured Sanskrit sentences into well-structured Hindi sentences. Machine translation in Sanskrit is an area of scope in Natural Language Processing. This is the application of NLP which require both the syntactic and semantic analysis at various levels. At the syntactic level, we develop algorithms or uses POS tagger to predict part-of-speech tags for each word in a given sentence as well as the various relationship between them. At the semantic level, we work on problems such as noun-phrase extraction, tagging the noun-phrases that refer to the same entity both within and across documents.

Keywords: Natural Language processing; Corpora, Tokenization, pattern matching, suffix stripping, prefix stripping, stemming.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "BILINGUAL CORPORA CREATION FOR SANSKRIT LANGUAGE USING NLP" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Prateek Agrawal. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

ANUJ KUMAR

11506651

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled **"BILINGUAL CORPORA CREATION FOR SANSKRIT LANGUAGE USING NLP"**, submitted by **Anuj Kumar** at **Lovely Professional University, Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Prateek Agrawal

Date:

Counter Signed by:

1) **HoD's Signature:** _____

   HoD Name: _____

   Date: _____

2) **Neutral Examiners:**

   (i) Examiner 1

   Signature: _____

   Name: _____

   Date: _____

   (ii) Examiner 2

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

I wish to acknowledge my sincere indebtedness and a deep sense of gratitude to Mr. Prateek Agrawal, HOD of Intelligent System, School of Computer Science and Engineering, LPU Jalandhar (Punjab) for his generous guidance, help, useful suggestions and making my work a pleasant experience. I shall always visualize his encouraging gesture, constant inspiration, and constructive and indefatigable zest to learn increasingly.

I express gratitude Head of School and other faculty members of Computer Science and Engineering Department, LPU for their intellectual support throughout the course of this work.

The words are not adequate to express my gratitude to my parents and entire family for their patience, support and active assistance in many ways. Finally, I indebted to all my colleagues, dears & nears and all whosoever has contributed in this dissertation work directly or indirectly.

Anuj Kumar

TABLE OF CONTENTS

CONTENTS **PAGE NO.**

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

Sanskrit since many thousands of years has been the original language of India. The SANSKRIT linguistic is engraved in diverse parts of India in the characters which are in use for the spoken dialects; but the alphabet which is regarded as most appropriate to it, and from where which the local alphabets are derived, is that which is termed **Nagari** or **Devanagari**, the alphabet of 'the city,' or of ' the city of the gods,' being a derivative from Nagara, ' a city,' compounded in the second form with Deva, 'deus,' ' a god.' It appears to have undergone various modifications from a period of remote antiquity down to the seventh or eighth century, when the letters assumed the form in which they now occur. Table 1.1 describe the basic information of Sanskrit language.

Table 1.1: Sanskrit Language Introduction

BASIS	**SANSKRIT**
Alphabets	42 characters
Number of Vowels	9
Number of Consonants	33
Number	Three: Singular, Dual and Plural
Sentence Order	Mostly Subject-Object-Verb
Tenses	6

Sanskrit is the base for most of the Indian Languages. Collection of texts of the written and spoken words is known as Language Corpora, which is collected in an organized way. Our corpora contains all the details of the Shabdhroop, Lakaar, Pronoun, and Interrogative Words completely. Corpora also have good collection of verb and common words. One set of shabdhroop words have been taken and manually evaluated,

and store into the database. The description of the Shabdhroop अकारान्त is shown in table 1.2. Similarly we collect the information for all the Shabdh types including vowels and consonants.

Table 1.2: अकारान्त शब्द रूप [26]

अकारान्त पुंलिङ्ग			
विभक्ति	**एकवचन**	**द्विवचन**	**बहुवचन**
प्रथमा	बालः	बालौ	बालाः
द्वितीया	बालम्	बालौ	बालान्
तृतीया	बालेन	बालाभ्याम्	बालैः
चर्तुथी	बालाय	बालाभ्याम्	बालेभ्यः
पन्चमी	बालात्	बालाभ्याम्	बालेभ्यः
षष्ठी	बालस्य	बालयोः	बालानाम्
सप्तमी	बाले	बालयोः	बालेषु
सम्बोधन	बाल	बालौ	बालाः
समरुप शब्द: बालक, देव, नर, सुर, राम			

Similarly, we collect the information for all the categories of the shabdhroop, lakaar, pronoun, interrogative words, verb and common words. For detailed description of each of the category you can refer **Appendix A.**

Communication is one of the vital parts of mortal behaviour and is a decisive element of our lives. In printed form it assists as an enduring record of familiarity from one group to succeeding. In verbal procedure it assists as our key means of directing our habitual behaviours with others.

NLP is the scientific study of natural languages and a field of computer science which makes computer interact-able with the human beings. Natural language processing was developed by a twenty years old psychologist student and an associate professor of linguistics in 1970 at California University. Both are working together in the field of Neuro linguistic programming. They found that there is a strong relationship between neurological processes, languages and behaviour pattern. Natural language processing is the sub area of an Artificial intelligence. Artificial intelligence is very broad area. There are two main parts of NLP:

1. **Natural Language Understanding:** Mapping the given input in natural language into useful representations.
2. **Natural Language Generation:** It is the procedure to convert the machine language into human readable language.

1.2 NATURAL LANGUAGE PROCESSING

When any natural language is handled by computer is known as NLP. It is an ability of a computer program to understand human speech as it is spoken. NLP have various subfields for research like Question Answering, Text categorization, Text Mining, Spell checking, Information retrieval, Information Extraction, Plagiarism detection, Sentiment analysis, Web mining etc. [24].

1.3 STAGES OF NLP

There are various steps of natural language processing (NLP). NLP steps are phonology, morphology, lexical analysis, syntax analysis, semantic analysis, discourse and pragmatic analysis. All the stages are related to one another. The output of one step is work as an input to the next step. Phonology deals with the concept of sound. When we design synthesizer then phonology is the main concern. The stages of NLP are shown in figure 1.1.

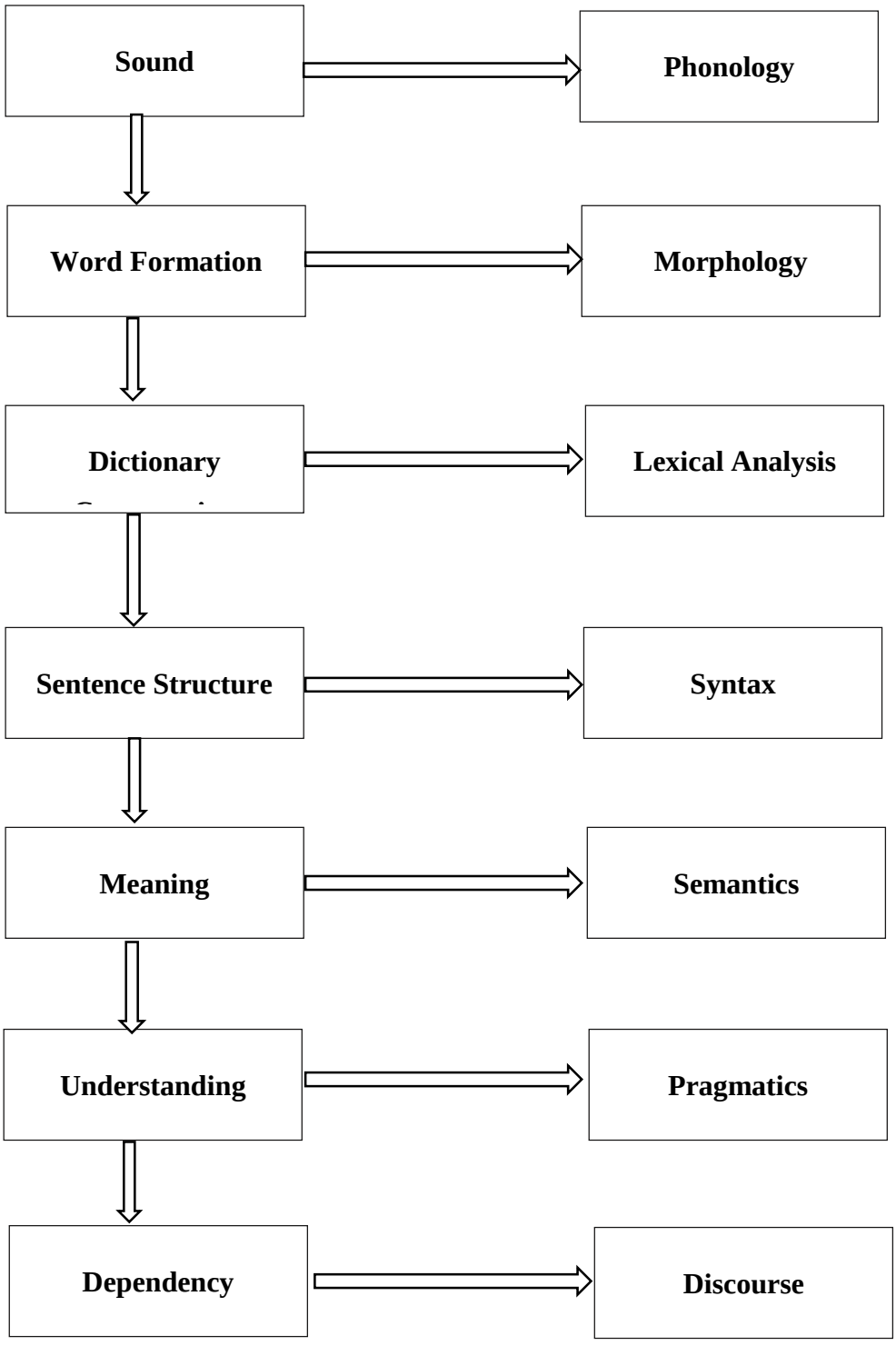

Figure 1.1: Stages of NLP [27]

i) **Phonology:** It is related to the sound. In this phase, the main task is to store the sound by using different levels of pitch and sound is processed in such a way that it is recognized by machine and store into the computer.

ii) **Morphology:** It is related to the words. In this steps we study about the different kinds of the words of natural language.

iii) **Lexical Analysis:** It is related to tokenization. In this step the words are divided into prefix, root and suffix part. Dividing a word into two or more parts is called sandhi witched.

iv) **Syntax:** Syntax analysis are used to check the grammatical correctness of the sentences. There are various techniques to check the correctness. The most popular is parsing tree. Parsing tree parse the sentence and tag each word like Noun, pronoun, verb, adjective, common noun etc.

v) **Semantic:** If any sentence is grammatical correct, it does not mean that it is semantically right. The main goal of semantic analysis is to resolve the problem of word ambiguity. So, semantic analysis identify the sense of the sentence and then use the correct meaning of the word. This is called word sense disambiguation.

vi) **Discourse:** At discourse level we find the dependency among sentences of a story. In this step, the main task is to replace the pronoun by the noun of the previous sentences. So, for that we store the relations between sentences in the form of relation table. For example:

राम पुस्तक पडता है | वह खाना खाता है | After discourse analysis, **वह** is replaced by the noun **राम.**

vii) **Pragmatic:** At the pragmatic level, grammatical rules are applied to the sentence to generate a meaningful sentence.

For example: **राम: पुस्तकम् पठति |** the output of the above sentence after syntax and semantic analysis is **राम ने पुस्तक को पडता है |** after pragmatic analysis output is **राम पुस्तक पडता है |**

1.4 AREAS IN NLP

i) **Question Answering:** In question answering application, the system is implemented in such a way that it is able to answer human language questions. The questions may be of any type. Sometime question answer could be one word like what is your age. But sometime question may be of descriptive type like describe yourself.

For example: IBM has developed the question answer application known as IBM Watson.

ii) **Text Categorization:** It is the task of assigning predefined categories to free-text documents. It can provide conceptual views of document collections and has important applications in the real world.

iii) **Sentiment Analysis:** It is process to identify the sentiment from the text and symbols. The text can be either structured or unstructured. The sentiments may be like happy, very happy, disagree, strongly agree, uncertain, like, dislike etc.

For example: Suppose online review are given to movie then we can perform sentiment analysis to determine the movie rating.

iv) **Information Retrieval:** It is the concept of searching that is finding the most related data for the users query. **For example:** Google, Bing, Yahoo Search engines. When we search something on Google then he find the most relevant information to our query.

v) **Information extraction:** Concept of meaningful information finding. For this first we perform information retrieval. It includes the lexemes, grammar, semantics and fact of languages.

vi) **Automatic summarization:** Finding the crisp part of the story or any document is called summarization. For example. We have a story of twenty lines then automatic summarization application generates a headline of one line which explain the story effectively.

vii) **Machine translation:** Communication is the vital part of human lives and language is the way of communication. By using language two or more persons are interact with each other. So, Machine translation is the procedure to convert one human readable language to another language by using computer. **For ex:** Converting Hindi language story to English by using software is called Machine Translation.

Application of NLP with which we are going to deal in our project is discussed below:-

1.5 MACHINE TRANSLATION

Machine translation is the field of linguistic. Machine translation is abbreviated as MT. It is a process to convert the speech or text from one language to another by using computer. When we convert speech then MT systems are known as synthesiser. Basically, Machine Translation performs the replacement of arguments from one natural language to the arguments in another language, but replacement alone usually cannot produce a proper translation of the text because the recognition of whole phrases and their closet counterparts in the target or specified language is needed or necessary. Solving this problem with wordnet and statistical techniques is the rapidly growing field which leads to better translation, handling differences in the linguistic typology, translation of the idioms, and the isolation of the anomalies [25].

The first set of proposals for computer-aided machine translation was presented in 1949 by warren weaver, a researcher at the Rockefeller Foundation, “Translation memorandum”.

1.5.1 Machine Translation Process

Machine translation is the dominant area of research in natural language processing. Machine translation is the procedure to convert one human readable language to another language by using computer. The Figure 1.2 shows all the phases involved in the process of machine translation:-

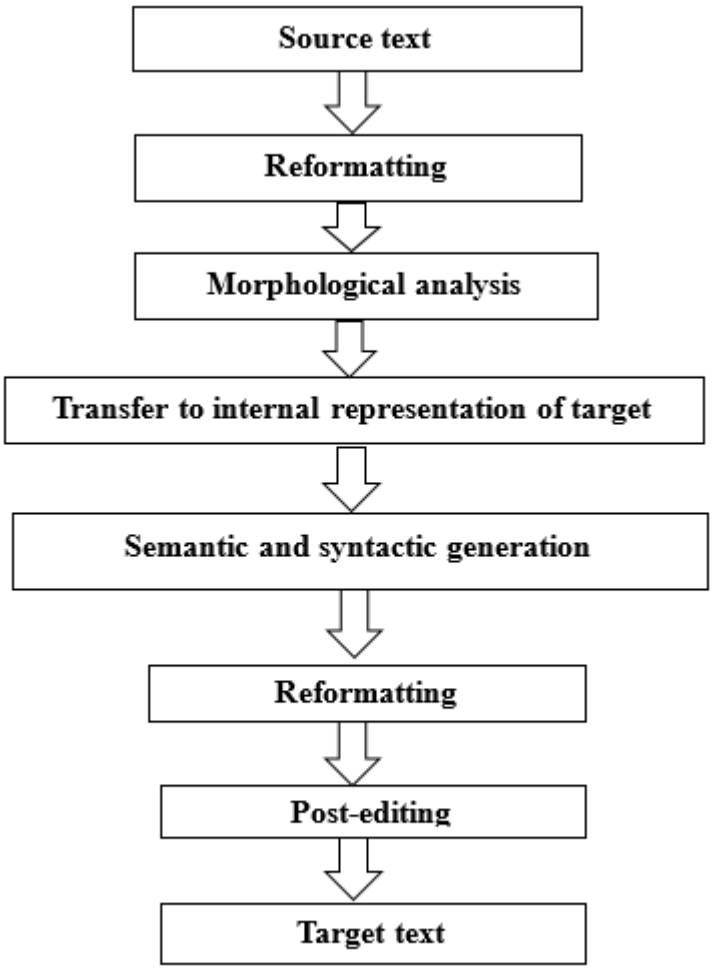

Figure 1.2: Steps of Machine Translation [27]

- **Source Text**

This is the first step of machine translation system. The input text is entered by the user of the system. Before start processing to the source text, initial testing is performed to check the correctness of the source language. The sentences are classified according to the difficulty level. At the initial stage we try to find the relationship between all the sentences. Because at discourse analysis, we need the relationship information to replace the pronoun by the noun of the previous sentence. This is

possible only when we store the sentence relationships. Interpretation of sentences could be done easily if we perform part of speech tagging. The source text is tokenized into different parts: prefix, root and suffix part. Suffix part is used to identify the type of word.

- **Deformating and Reformatting**

For making the machine translation process easier and qualitative, this step is involved. The source language text may contain figures, flowcharts, and tables etc. that do not require any type translation. So only that portion is identified which needs the translation. Once the source text is translated into the target text is to be reformatted after the post-editing.

- **Pre-editing and Post editing**

We can find the performance and an efficiency of any machine translation system by the level of pre editing and post editing. For some systems segmentation is required for dividing the long sentences into short sentences. Pre-editing involves the fixing up punctuation marks and blocking material that does not require translation. For making the quality of the translation up to the mark then post-editing is done. Post-editing is unavoidable or necessary especially for translation of the crucial information such as one for the health. Post-editing should continue till the MT systems reach the human-like systems.

- **Analysis, Transfer and Generation**

Morphological analysis is used to study the internal structure of the word. In morphological analysis we read the different components of a word. Syntax analysis are used to check the grammatical correctness of the sentences. There are various techniques to check the correctness. The most popular is parsing tree. Parsing tree parse the sentence and tag each word like Noun, pronoun, verb, adjective, common noun etc. Tagging can also be used to solve the problem of word disambiguation. Because some of the word have many meaning. Semantic analysis is the study of meaning. It is mainly used for word sense disambiguation. That is to resolve the problem of ambiguity. A word may have many meanings, so which one is used. For that semantic analysis is used.

- **Morphological analysis and generation**

Computational morphology deals with recognition and analysis of the words. Some types of morphological process are: - inflections, derivations, affixes and combining forms. Inflection is the most common and productive morphological process across all the languages. Inflection do the alteration of the form of the word in number, gender, mood, tense, aspect, person, etc. Morphological analyser gives information regarding the morphological properties of the words that it analysis.

- **Syntax Analysis and generation**

Syntax analysis are used to check the grammatical correctness of the sentences. There are various techniques to check the correctness. The most popular is parsing tree. Parsing tree parse the sentence and tag each word like Noun, pronoun, verb, adjective, common noun etc. Tagging can also be used to solve the problem of word disambiguation. Because some of the word have many meaning.

- **Grammar Formalism**

It is the framework that explain the basic structure of all languages. Researchers propose the following grammar formalisms:

  - Phrase structure grammar
  - Dependency grammar
  - Case grammar
  - Systematic grammar
  - Montague grammar

- **Parsing and Tagging**

Tagging can be defined as the identification of the linguistics properties of the individual words. Parsing can be defined as the assessment of functions of the words in relation with each other.

- **Semantic, Contextual analysis and generation**

Semantic analysis includes the meaning representation and assigning them the linguistic inputs. The semantic analyser basically uses lexicon and grammar for creating context independent meanings. The source of knowledge consists of meanings

associated with grammatical structures, meaning of words, and knowledge about the discourse context.

1.5.2 Applications of Machine translation

- **Publishing quality translation,** is developed by translations agencies and must be used for legal purposes and marketing and specific translation such as software localization. Publishing good quality translation is also used to provide the same information in various languages. By using quality translation system the websites can be work in a multi-linguistic way.
- **Next generation translation,** as available with translation enterprise, translation websites and translation application programming interfaces has many applications which has the ability to perform automatic email translation, open a website into different languages and also perform translation of various documents.
- **Enhanced machine translation,** developed by translation pro. It can be used for translation foreign documents. In can also be used for research purpose. Basically pro tool can be used to enhance communications such a traveling reservation, to manage your property abroad or simply communicate with any other domains.

1.6 BILINGUAL CORPORA CREATION

In this thesis work, we have described the methodologies that we have used in creating a Bilingual corpora creation for Sanskrit language using NLP. The basic step to create corpora is to collect the text data from the native user of the language. For the collection of vocabulary we used primary school text book as well as daily use words by the help of native speaker. For creating Sanskrit to Hindi corpora we used NLP technique. The first task of NLP is to perform sandhi witched. The proposed technique is implemented using Java.

The corpora creation for Indian languages was started in 1919 by the consortium of various departments. By the effort of department of electronic and Govt. of India the text data is available online for various languages first time. Every department has the responsibility to create corpora for different languages. Sanskrit, Kannada, Malayalam,

Telugu and Tamil database is developed by the central institute of Indian languages, Mysore.

Corpora contains a huge collection of knowledge rather than text data which makes the natural language processing task so easy. This information is divided into two parts:

i. Exact form of the collected data called representative knowledge.

ii. To add a descriptive information to the collected text.

To add a linguistic information to the text is also known as annotated corpus. The main advantage of annotated corpus is that the structural information can be retrieved at different level of language processing, which are the frequent requirement of the NLP researchers.

1.7 INTRODUCTION TO JAVA

Java is an object oriented programming language developed by sun microsystem in 1991. Initially it is known as oak and used to developed electronic devices. Later, in 1995 it is renamed as Java. Java is simple, portable and multithreaded language. Java is a platform independent language that is java application run on any hardware and operating system where java is installed.

Java Program Execution Model:

The java is platform independent language because java used a special environment for execution which is responsible to generate machine independent code known as Java Run Time environment (JRE). Figure 1.3 show an execution model.

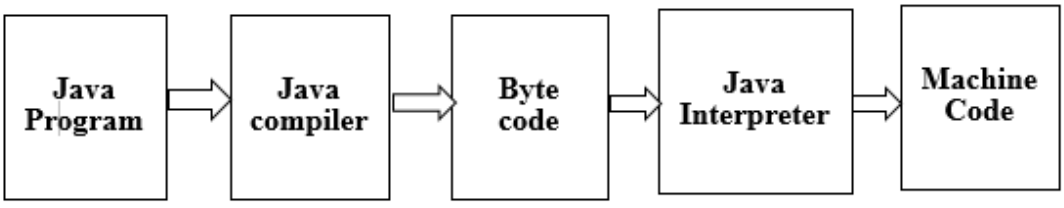

Figure 1.3: Conversion of Java Program to Machine Code

The java program is given to the compiler, which generates a .class file also known as Byte-Code [28]. The byte code is also known as machine independent code, which is

run on any platform where java is installed. The .class file is read by java interpreter, which is responsible to generate a machine code.

1.8 APPLICATION OF THE SYSTEM

The scope of the study is purely Natural language processing in Sanskrit language. There is not much work done in Sanskrit language. So, we are working in the field of Sanskrit language. We design corpora for Sanskrit to Hindi language using NLP in such a way that it is responsible for performing text-independent machine translation. Our corpora will be considered as machine translation from Sanskrit to Hindi language. NLP has various subfields like machine translation, sentiment analysis, automatic text summarization etc. Here, we are dealing with the machine translation area of NLP in which we are doing our work of bilingual corpus creation.

Various applications of our system are as follows:

i. **In various institutions for Sanskrit teachers:** Our system can be used by teachers to teach Sanskrit language in efficient and effective way. This system helps the students to learn Sanskrit grammatical rule quickly as well as teachers to explain efficiently. It make teaching more interesting.

ii. **For various writers:** The story writers who write the story in Sanskrit language and it is very difficult for them to publish the same story into Hindi. Our system will also help the writers to perform translation from Sanskrit to Hindi.

iii. **For awareness of Sanskrit language:** In today's era, we use English language mostly, Sanskrit language losing its importance and scope. New generation lost the basic knowledge of their mother language. So, this will increase the popularity of Sanskrit and Hindi language among the people of India.

CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

This chapter will survey the different types of research papers that are published in the field of natural language processing and machine learning. So we will try to explain various algorithms, approaches and technology that help us to finds the problem in which field we are working.

2.2 LITERATURE SURVEY

Md. Khalilur Rhaman et al. [1] build a model Bangla to English (B2E) bilingual translation using natural language processing. He used rule based approach to implement B2E. For Bangla to English translation first he performed morphological analysis for Bangla then he used rule based approach where he perform case analysis. To construct an English sentence from Bangla he used SVO grammatical rules. System perform B2E translation for Assertive-Affirmative, Negative and interrogative sentences. It used the following architecture to perform translation.

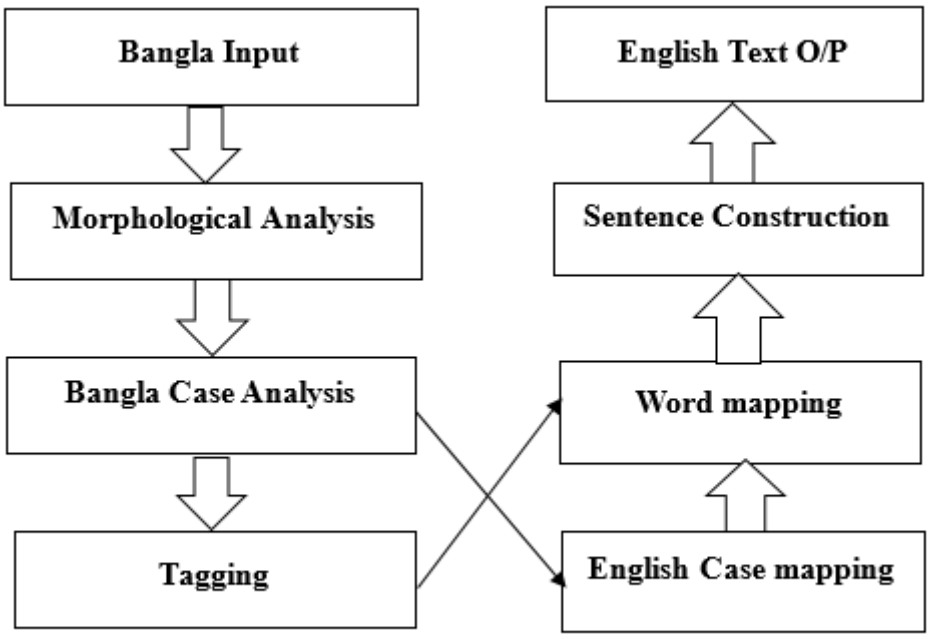

Figure 2.1: BTE process model [1]

In this paper system only considered 60 verbs, lexicon contains 1460 nouns, almost all pronouns, 650 adjectives and 19 prepositions. If any word is not found in database then system would considered it as a proper noun. During morphological analysis, it perform Sandi vichched on input text. It separate word in to two parts root word and suffix. On the basis of that information, the system will search into the corresponding entry in the database. He implemented Bangla to English translation by using JAVA. He used DFA to generate English sentence.

Sangavi G et al. [2] presents a machine translation system for English to Tamil and Tamil to English in 2016. Language is the way of communication and by using Machine translation communication is performed between human and computer. There are various available techniques for Machine translation. Some of them are Direct Machine Translation, Corpora based Machine translation, rule-based machine translation and statistical machine translation.

Sangavi G used statistical machine translation and the system is divided into two parts: Training and testing. The translation system designed by Sangavi G is show in Figure 2.2

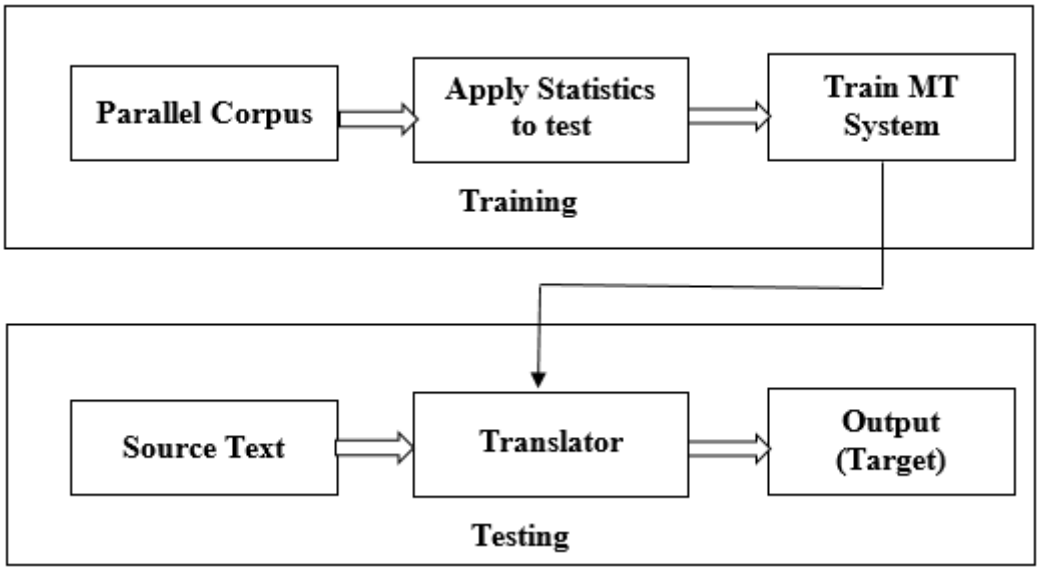

Figure 2.2: Machine Translation System [2]

The parallel corpora is designed to a particular domain and the training is started by making the vocabulary list for both the languages English and Tamil. The corpora is trained by using statistical machine translation technique.

Namrata Tapaswi et al. [3] describe a procedure for the construction of grammar for Sanskrit language using Parts-of-speech tagging and morphemes. The main problem with the construction of corpora is to understand the exact sense of the word. Because some words have more than one meaning. This problem is called ambiguity. For solving the problem of ambiguity there are two methods. The knowledge based and corpora based. The knowledge based used predefined collection of lexicons while corpora extract the information from the corpus to solve problem. Author also define the advantages of annotated corpora. Annotated corpora can be used at discourse and pragmatic analysis of natural language processing. By the help of annotated corpora we can easily identify the category of different words like Noun, pronoun, verb and adjective used in a sentence. There are mainly two types of taggers are used to tag database. First one is rule based and second is statistical tagger. In rule based tagger, first rules are designed and by using these rules the sentence is tag. In statistical tagger, Hidden Markov model is used and the accuracy of HMM is 100%.

Shahnawaz et al. [4] performed conversion between Hindi and Urdu Language. He explained various machine translation approaches namely, example based, statistical based, corpus based, direct machine translation and hybrid approach. But he used direct machine translation approach to perform conversion between Hindi and Urdu. A direct machine translation start with giving the input text to the system. Then system perform morphological analysis to the text. In morphological analysis, the internal structure of the word is examined. For example, in word Unbelievable the morphological analysis determines three things. The first is prefix (UN), root is belie and suffix is able. After performing morphological, the tokenization process is applied and the stemmed words are stored into the dictionary for machine translation.

Ved Kumar Gupta et al. [5] defined Machine Translation as a step by step process which converts one natural language to other by using computer. Authors implemented Sanskrit to English machine translation system by using rule based technique. They also define various machine translation techniques like Corpora based machine translation, example based, direct machine translation, statistical based and rule based

technique. They perform machine translation in three steps. In first step, they perform tokenization, that is, the words are divided into root and suffix part. In second step they compare suffix word to identify the type of the word. The word could be either noun, pronoun, verb, adjective etc. and store the information into an index number. In final step, the information is displayed according to the value of index number.

Farshad Kiyoumarsi et al. [6] compare summary generated manually to the summaries generated by various automated tools. The manual summaries are finding by the expertise of English languages. Automatic summaries are generated by tools and tools implement two technology. First one is fuzzy methods and second is vector approach. The best summaries are generated by fuzzy tool instead of vector approach. Automatic text summarization is the sub field of natural language processing. There are two methods for automatic text summarization. Summary based on abstraction and summary based on extraction.

Sarita G. Rathod et al. [7] perform translation of sentences from English to Sanskrit. The translation done by using various methods but they used only dictionary based approach to translate simple English texts involving the need for tokenize, applying grammar rule create parse tree into corresponding appropriate sentences in Sanskrit. It has two modes Text to Text translator module and text to speech synthesizer module.

N. Murali et al. [8] perform analysis on kridantas for Sanskrit language. It is a novel approach. They describe that Sanskrit words are divided into two categories: Declinable and Indeclinable. A word that can be changed according to the cases are known as declinable. Declinable includes noun, pronoun, verb and adverb. Indeclinable words are called avyayas. In the word kridantas, kri is used for normalization. They use tokenization process to identify the declinable words and design an avaya analyser tool to identify indeclinable words.For example: **raama:** is divided as **raam+a:** and **lataabhyaam** is tokenized as **lat+aabhyaam.**

Namrata Tapaswi et al. [9] define morphological and lexical analysis of the Sanskrit Sentences. Sanskrit is a free ordering language (or syntax free language) and there is no ambiguity in the form of the words order change. Both the phases are based on Sandhi-Wichched (SW), so they design an algorithm for SW:

Algorithm:

Step 1: Begin

Step 2: Receive a word for doing Sandhi-Wichched.

Step 3: Try iteratively breaking the word in two parts: root word and suffix.

Step 4: End

Sarita G. Rathod et al. [10] implement machine translation system for English to Sanskrit language by using the combination of various machine translation techniques. Author used the combination of rule based and example based approach to implement English to Sanskrit translator. The main module of the developed system are text input, tokenization, spell checker, Parser, and RBMT and EBMT translator. The task of tokenization is to generate lexemes, spell checker is the advantage of the system. Parser is used to check the grammatical correctness of the system. RBMT and EBMT both use previous modules to generate outputs. The main goal of machine translation software is to convert text of one natural language to another human language by using computer. The basic requirement to design a machine translation system is to design translation algorithm, corpora and grammatical rules. The translation algorithm is implemented in any programming language.

Nandini Sethi et al. [11] describes an approach for finding the central theme of the story using NLP. The document is in English. The system takes English story as input and produces the title after applying various approaches like frequency prioritization, noun and adjective combination or idiom based title. System used Hidden Markov Model (HMM) part of speech tagger to tag the complement English story. HMM tagger is used at pragmatic analysis step.

Wei Yen Chong et al. [12] performed experiments on tweets sentimental analysis. This examine is used to determine the sentiment based on subject that exists in tweets. For performing sentimental analysis he used natural language processing approach. The experiment consist mainly three steps, subjectivity, classification, semantic association,

and polarity classification. It used part of speech tagger to identify noun, verb, adverb, adjectives. At the initial step system perform pre-processing to convert unstructured tweets into the structured format.

Damodar Magdum et al. [13] presents methodology for designing and creating Hindi speech corpus. Speech corpus plays a very important role in text-to-speech applications. The overall quality of speech to text conversion is totally depend on the size of the corpus and quality of corpus. So, they use various steps to design Hindi corpus as shown in Figure 2.3

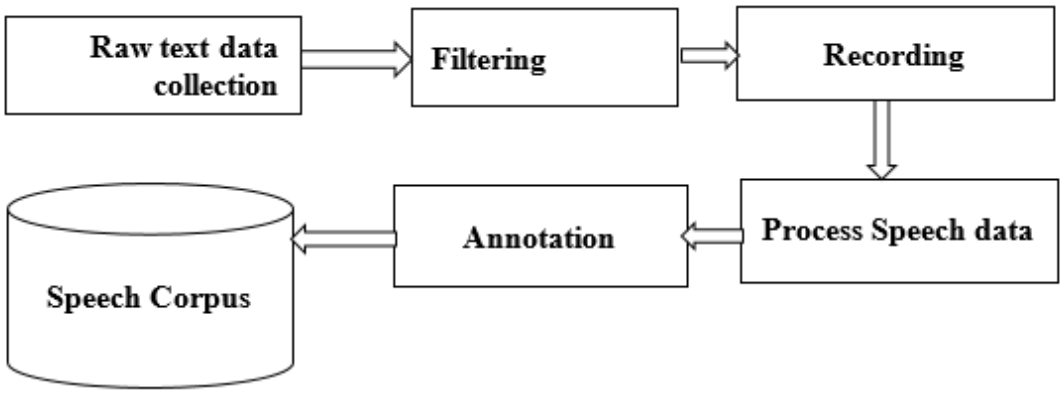

Figure 2.3: Methodology for creating Hindi speech corpus [13]

First, they collect text data from various domain including financial, news, banking sectors, government etc. along-with predefined dictionaries. The crawled text is filtered for getting accuracy. Filtering includes various tasks like spelling mistakes, word length, and validity of Hindi etc. The filtered words then analysed carefully and ensures that the collected text is correct. Finally, the data is recorded by the native professional and then recorded data is processed and annotated to generate the final speech corpus.

Pramod Salunkhe et al. [14] describes a hybrid machine translation approach for English to Marathi known as hybrid translator. They perform statistical and rule base approach for the input. The corpus contains information of various sources like agriculture domain, tourism and medical. They also implemented Marathi wordnet to increase dictionary and to incorporate better translation result. This system is currently working for text documents but it can also be extended to speech and voice. They also

compare their hybrid translator with Google translator that shows the better result from Google translator.

Neeha Ashraf et al. [15] defines various machine translation techniques and their comparative study. Communication is one of the vital parts of mortal behaviour and is a decisive element of our lives. In printed form it assists as an enduring record of familiarity from one group to succeeding. In verbal procedure it assists as our key means of directing our habitual behaviours with others. Information presents in different language and structure creates a barrier in information retrieval. Many times government documents are presented in English language or any other document published in some other foreign language where an untrained guy from Hindi linguistic contextual discoveries trouble to recognize material. So, we need an automated translation system which works in cross field material recovery. A lot of research work has been done on paraphrase of English to Hindi, Tamil Bangla and various distant languages. MT is challenging research area due to ambiguity, structural and lexical mismatches. There are various machine translation techniques are used to present automated translation. This research work presents a comparative study of various machine translation approaches used for multilingual translation and creation of execution framework for comparing machine translation techniques using open source translation tools.

Sunita Chand [16] gives survey of Machine Translation Tools in 2016. She perform comparative analysis on various machine translation tools available online and give a result that there is no such machine translation tool available till now that performs better result or same as human translation. The tools tested for this research as show in table 2.1:

Table 2.1: Comparison of Various Online MT Tools

Machine Translation Tool	**Technique Used**
ANGLA-BHARTI-II	Hybrid MT
ANUBAAD	Rule based approach
GOOGLE	Statistical machine translation
BING	Statistical machine translation
IM Translator	Statistical machine translation

Kanika, Ankur et al. [17] presents a review of English to Indian languages translator Anusaaraka. It is a kind of language translator that translate English to any indian language. They defined various machine translation tools Angla Bharti, Mantra, Shiva, Shakti, Sampark, ANUSAARAKA. Mantra is a machine assited machine translation tool developed by C-DAC and it is used by government of india to translate documents. This tool is specific to some domains like agriculture, information technology and health care. At the starting point, this tool convert only English to Hindi but now it is capable to perform translation to various other indian languages. Angla Bharti uses pattern based matching. This tool is based on context free grammar technology. The source language English is converted into an intermediate code called Pseudo Linga for indian languages (PLIL). This code is then converted into the destination language. Shiva is a machine translation tool based on corpus based technique. Shakti used two techniques rule based and corpus based to perform machine translation. Sampark is developed by consortium of institutions IIIT Hyderabad, IIT Kharagpur, CDAC and many more. This tool perform conversion from Punjabi, Tamil, Telgu, Urdu to Hindi.

Shaharban et al. [18] describe approach for pragmatic analysis of Malayalam sentences. They mainly represents how Malayalam sentences are used in different context or different situation. They perform pragmatic analysis on different steps. Pragmatic analysis are also known as speech act analysis. This system contains different modules:

1. **Tokenization:** In tokenization the given Malayalam sentences is divided into the words.
2. **Parts of speech tagging:** In POS tagging, different words are labelled as Noun, verb, pronoun, and adverb.
3. **Classifier:** It is responsible for performing sentence categorization. Sentences started with This, that is classified as fact. Sentences classified as SVO.
4. **Dynamic corpus:** Initialize the Dynamic corpus with some valid examples. When sentences are input to the system, if they are match with the existing example then ok otherwise check the validity of the new sentence. If new sentence is valid then stored it to the dynamic corpus.

B.N.V Narasimha Raju et al. [19] define a statistical machine translation system for indian languages. This system consists of Language model, decoder and translation model. Language model is used to calculate the probability of the target language sentence and the probability of the source language sentence is computed by translation model. The probability of the sentences can be computed by using n-gram. By using n-gram probable model, it determines how words are continuous. The probability is decomposed into different parts by using Markov chain rule. The methodology used by this system is shown in figure 2.4.

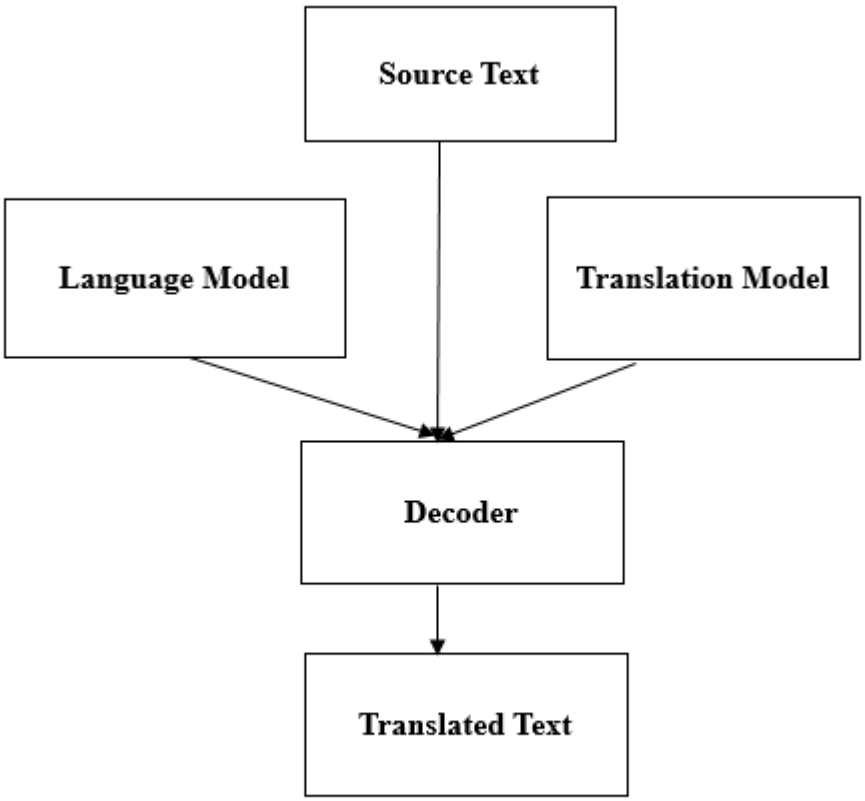

Figure 2.4: Statistical Translation Model [19]

Pooja Malik et al. [20] gives an improvement in BLEU metric for English-Hindi machine translation system. The BLEU (Bi lingual evaluation understudy) metric is developed by IBM used to evaluate the performance of the machine translation system. The BLEU is based on the concept of n-gram. It compares machine translation result with the human result. The main problem with the BLEU metric is the synonym problem. It consider the synonyms as a different words that decrease the final score of the above equation. For solving this problem they integrate synonym replacement module in the improved BLEU metric as different module.

Pooja Alva et al. [21] described a hidden Markov model for Part of speech tagging in word sense disambiguation. The process of identifying the different meaning of a word in sentence is called as sense disambiguation. POS tagging is used to identify the type a word in a given sentence. Type refers to either noun, pronoun, verb, adverb or adjective. Hidden Markov model helps us to find the probability of the particular words and is also used to predict the sequence of the words. For example, if we find an article "An" then there are 45% chances that next word is a noun, 40% chances that next word

is an adjective and 20% chances that next word is a number. Hidden Markov model has two main components:

1. **Transition probabilities:** We know that the main task of Markov model is to assign grammatical categories to the word. Then there may be a chance that we assign wrong category to a word and now the requirement is to change the category. So, the transition of a category from one to another is called as transition probabilities.
2. **Emission probabilities:** This matrix show the probability that a given word can have a particular part of speech.

M. Rajani Shree et al. [32] defines a novel approach for sandhi witched. They perform sandhi witched for kannada language at character level. They split each kannada word into morphemes. After performing splitting they manually tag each word into root and suffix. Implemented system perform tagging on to the split words. The tagged data generated by system is compared by manually generated data. They provide sandhi bound kannada word to the system as an input. Then consonants and vowels are identified in the word and according to vowels and consonants a pattern is generated that is splitting is performed.

Leena Jain, et al. [23] defines a Text independent root word identification in Hindi language using natural language processing. They stem Hindi words and identify the prefix, root and suffix words. They design a stemmer that reduced the word to its root form that helps to perform various natural language processing tasks. They performed test on inflected word list of 800 words, prefix list of 160 words, root list of 910 words and suffix word list of 340 words. In this technique, pre-processing of text is required at the time of stemming a word which improves its time complexity of giving responses

CHAPTER 3
PRESENT WORK

3.1 PROBLEM FORMULATION

The scope of the study is purely Natural language processing and Machine Translation. NLP field has various areas like: Machine translation, speech recognition, automatic text summarization, Sentimental analysis, Text categorization. Machine translation deals with converting the sentence from one language to another language. Speech recognition is the process of converting the speech into the text. Automatic text summarization generate the summary with important lines of a document. Sentimental analysis identifies the sentiments from the given text like: sad, happy, satisfy, disagree. Information extraction is the concept of meaning full information finding.

Here, we are dealing with the area of the natural language processing that is machine translation. Translation in which we are doing our work of Bilingual Translation and Corpora Creation.

The scope of Machine translation is also in the field in which we are working for our research. This deals with both syntactic and semantic analysis at various levels. At the syntactic level, Syntax analysis are used to check the grammatical correctness of the sentences. There are various techniques to check the correctness. The most popular is parsing tree. Parsing tree parse the sentence and tag each word like Noun, pronoun, verb, adjective, common noun etc. Tagging can also be used to solve the problem of word disambiguation. Because some of the word have many meaning. On the semantic phase, we work on problems such as noun-phrase extraction, tagging noun-phrases as person, organization location or common noun. Clustering noun-phrases that refer to the same entity both within and across documents.

The story or novel writers who write story easily in Sanskrit language but it becomes very difficult for them to write the story into Hindi language. So, it becomes a major problem for writers to write the story into different languages. So, our system will helps writers to convert the Sanskrit story into its equivalent Hindi.

New generation does not have much more knowledge about Indian languages like Hindi and Sanskrit. Indian histories, ved puranas are written in Hindi and Sanskrit. So, that new generation does not read about Indian histories and ved puranas. So, this system will help to increase the awareness about Indian languages among new generation. System can also be used by college and school teachers to teach Sanskrit language more efficient and effectively. Students can also use the software to learn Sanskrit grammar easily and quickly.

3.2 OBJECTIVES OF THE STUDY

Objective is the collection of tasks that are incorporated into research work or any other activity. The main objective of our research is the Bilingual Corpora creation of Sanskrit language using NLP. Our research objective includes the following task:

I. Transliteration (From English to Sanskrit).
II. Tokenization.
III. Corpora creation for Sanskrit to Hindi Language.
IV. Testing of the system.

Objectives are the requirements which the user can expect from the software, project or from the research. Objectives are of two types: social objective which deal with the objectives of the society that is in which way our system is beneficial for the society. Technical objective is that which deal with the technical requirements of the user from the project or the software.

3.2.1 Social Objectives

It is to design and build a software which will help the society in the following ways:-

I. It will very helpful for Sanskrit teachers to teach students the Sanskrit language in efficient and effective way.
II. It will help the users of Sanskrit Language for translating the Sanskrit story into Hindi.
III. It will help the human being to learn the concepts of Sanskrit language that is for those who will don’t know about this language.

IV. Not much work has been done for Sanskrit to Hindi translation, so this will increase the popularity of Sanskrit to Hindi Translation among the people of India.
V. The system user interface is very user friendly so any one can operate it and will understand it.
VI. The system can be used by the people for the purpose of getting knowledge about the Sanskrit and Hindi language.

3.2.2 Technical Objectives

The technical objective of our research is to design and build software that will analyse, understand, and generate languages that humans are naturally, so that eventually we will be able to address our computer as though we were addressing another person. The natural language is easiest for humans to understand, learn and use, but is hardest for a computer to master. So, our technical objective is to provide training to machine in such a way that it can work efficiently. By doing this, our machine can solve the problem very efficiently with high speed and with high accuracy than human being. So, technical objectives of our research are as follows:

I. Building and Designing User-friendly software.
II. It should provide results efficiently with high speed.
III. The result produce by the software must be accurate, this can be done only if we train our machine properly with the language on which we are working.
IV. It must consume less time in comparison with the human being.
V. The system will be highly reliable and efficient that it will be based on the specified, accurate rules of grammar.
VI. The system will also help them economically as this is one time investment only.

3.3 RESEARCH METHODOLOGY

3.3.1 Overview

The whole idea of this thesis work is based on NLP. The main concept is that, we collect the raw text data of Sanskrit language and create corpora using NLP techniques. Corpora is created in such a way that it is able to perform text-independent machine translation of Sanskrit language to Hindi language. First, we perform sandhi-witched. After performing sandhi-witched the word is divided into two parts root word and suffix word. Suffix word is used to tag the words as Noun, Pronoun, Verb, Interrogative words, Common Words. There are various approaches to perform machine translation like Corpora based approach, Rule based approach, Statistical approach. But we used text-independent hybrid machine translation approach. The system has mainly two parts as shown in Figure 3.1.

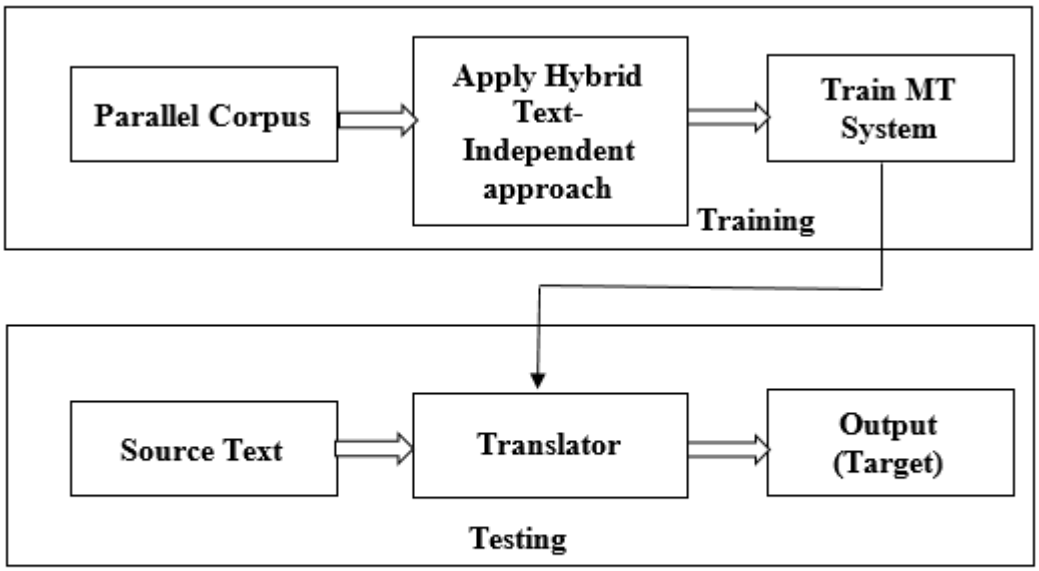

Figure 3.1: Bilingual Machine Translation System

1) **Corpora creation:** For the collection of vocabulary word we used the primary school text book as source as well as daily life word which is understandable to all native speakers. The word from primary text book found to be grammatically rich.

2) **Translator:** The translator system is built on top of the model to translate Sanskrit sentences into the equivalent Hindi Sentences. The translator work used text-independent machine translation hybrid approach.

Here, we are dealing with the most important application area of NLP is machine translation. The first challenge in implementing machine translation techniques is the platform or the framework through which human can interact with the computer. NLP techniques are used to generate Hindi sentences corresponding to Sanskrit sentences. There are many challenges to provide communication between humans and computer. Computers can understand only binary digits i.e., 0 and 1 but human can't understand binary language. So we require a database which store the huge data for processing of human understandable words by computers.

The efficiency and performance of any intelligent application is totally based on the size of the corpora. The machine translation of complex stories will purely depend on the size of the database being used in the research work. Our corpora contains approximately all the pronouns, interrogative words, Shabdhroop, lakaar and for including the remaining words, our database is designed in such a way that it is able to perform text-independent translation from Sanskrit to Hindi sentences.

Here, the main aim is to design "Sanskrit to Hindi" corpora and build software that will analyse, understand and generate natural language that human will understand easily, so that eventually we will be able to address our computer as though we were addressing another person. This task is not so easy. "Understanding" Sanskrit language means, knowing what concepts a Sanskrit word or the phrase stands for and then knowing how we link those concepts together in a meaningful way.

Figure 3.2, describes the stepwise phases of methodology which is followed for our research:

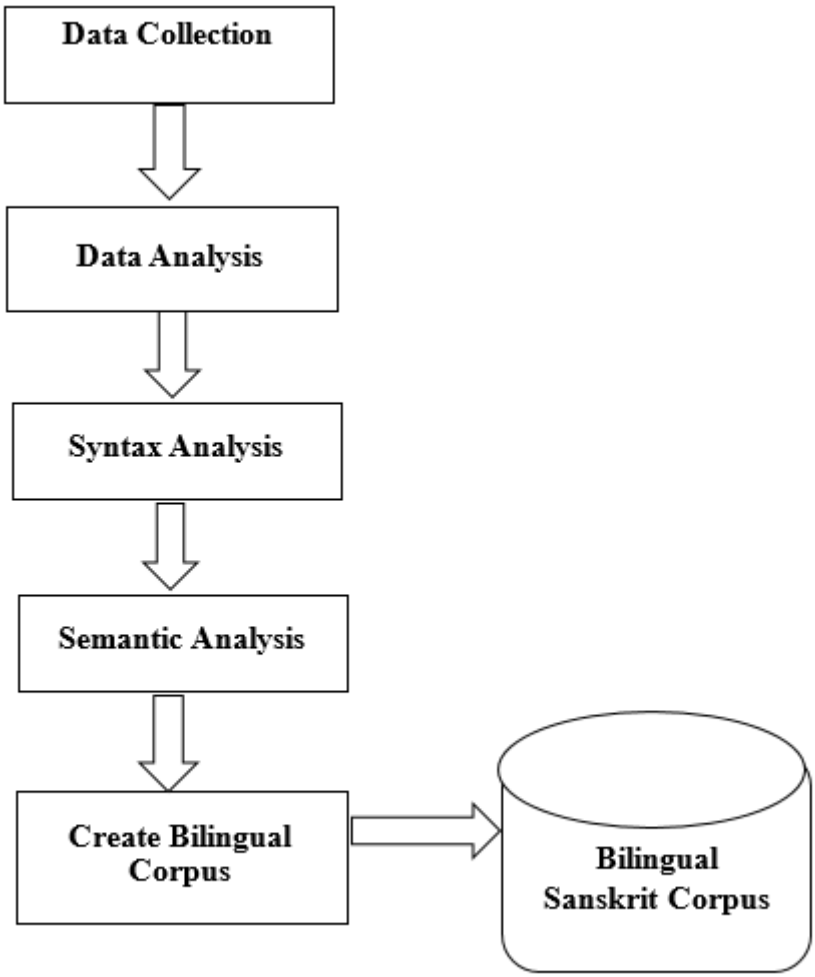

Figure 3.2: Flowchart of Methodology Part-I

3.3.2 Data Collection

In this step, we collect the data in Sanskrit language i.e. Sanskrit word net which includes Shabdh type, lakaar, noun, pronoun, verb, interrogative words, pronoun, and some common words. The step-by-step process of collecting Sanskrit words is as follows:

3.3.2.1 From Sanskrit Grammar Books: For collection of Shabdh type, lakaar, noun, pronoun, verb, helping verb, pronoun, and interrogative words we studied grammar rules related to Shabdh and lakaar type. After that we create a file of all these words in Sanskrit which will helpful in further processing of sentence. We studied books of various classes for finding the different kind of words.

3.3.2.2 From Sanskrit Professor Tutor: We met Sanskrit professor for taking help in understanding the meaning of different words. The professor help me to resolve the problem of ambiguity that arrive at stemming steps. When we perform stemming (for ex. **paThati =paTh+ati** related to lakaar) then the same suffix word is related to both lakaar and Shabdh type, and also help me to understanding the meaning of different words.

3.3.2.3 From Internet: we surf various sites to collect the Sanskrit words and understand the meaning of them. Internet also help me to find the huge collection of Shabdh and lakaar type. Our corpora contain almost all the Shabdh type and lakaar.
Now, we explain various types of collection that our corpora contains:

i. **Collection of Shabdh type:** In this step, we collect all the details related to different kind of shabdhroop. There are two types of shabdhroop either vowels or consonants. Every Shabdh type in Sanskrit have eight vibhakti and three vachan type. On the basis of vibhakti all the shabdhroop are distinguishing identifiable.

ii. **Collection of lakaar:** In this step, we collect different types of lakaar details from Sanskrit grammar books. There are mainly five types of lakaar; लट् (वर्तमान), लोट् (आज्ञा), भूत, भविष्य, विधि. Every lakaar has three purush types' pratham purush, madhyam and uttam purush. For detailed description of the lakaar refer **Appendix A.**

iii. **Collection of common words, pronoun and Interrogative words:** In this step we collect pronoun, interrogative words and different common words useful for performing bilingual translation. For detailed description of the words refer **Appendix A.**

iv. **Collection of Verb and Helping Verb:** In this step, we collect the verb from Sanskrit to Hindi language that are helpful for performing bilingual translation. Our corpora will contain good collection of verb. For detailed description refer **Appendix A.**

3.3.3 Data Analysis

The main problem occurred during Natural language processing is ambiguity. So, in this step we are dealing with ambiguity problem. First, we divide the story into sentences and then finding the ambiguity in the words. Some words are treated as noun and verb as well. So, these kinds of words are placed in both the files. Data analysis can be done with the help to following steps:

3.3.3.1 Syntax Analysis

Syntax analysis are used to check the grammatical correctness of the sentences. There are various techniques to check the correctness. The most popular is parsing tree. Parsing tree parse the sentence and tag each word like Noun, pronoun, verb, adjective, common noun etc. Tagging can also be used to solve the problem of word disambiguation. Because some of the word have many meaning. In figure 3.3, it shows the parsing of a simple Sanskrit sentence.

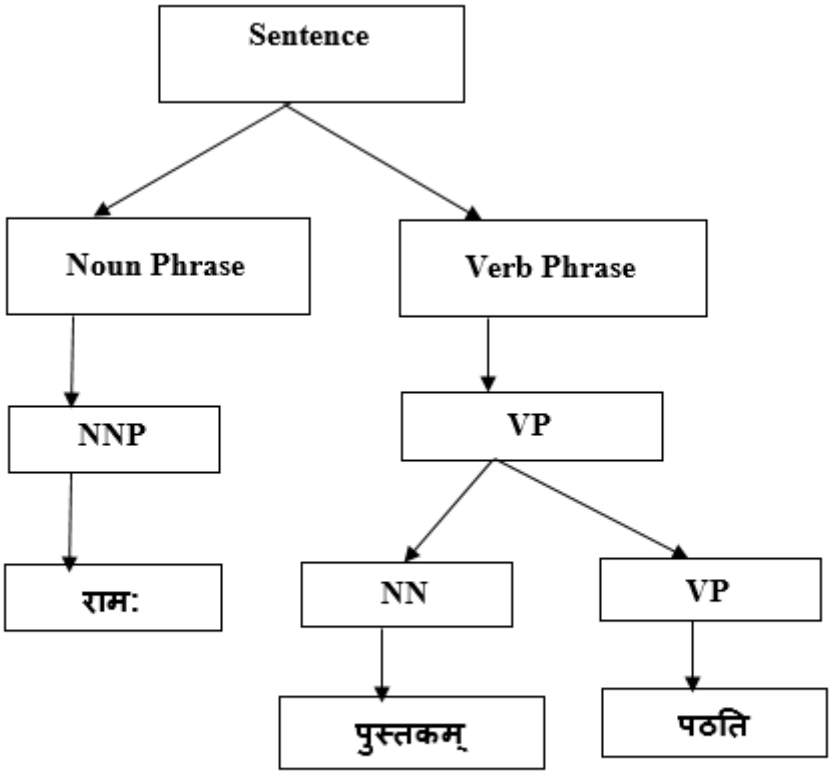

Figure 3.3: The result of syntactic analysis of "राम: पुस्तकम् पठति"

3.3.3.2 Semantic Analysis

If any sentence is grammatical correct, it does not mean that it is semantically right. The main goal of semantic analysis is to resolve the problem of word ambiguity. So, semantic analysis identify the sense of the sentence and then use the correct meaning of the word. This is called word sense disambiguation. So, semantic analysis is the concept of meaning.

3.3.4 Create Bilingual Corpus:

After collecting the data and analysis, now we create bilingual corpus. Basically, Corpora plays a key role to perform translation. Language translation is totally depends on corpus. Figure 3.4 show the steps for creating Bilingual corpus.

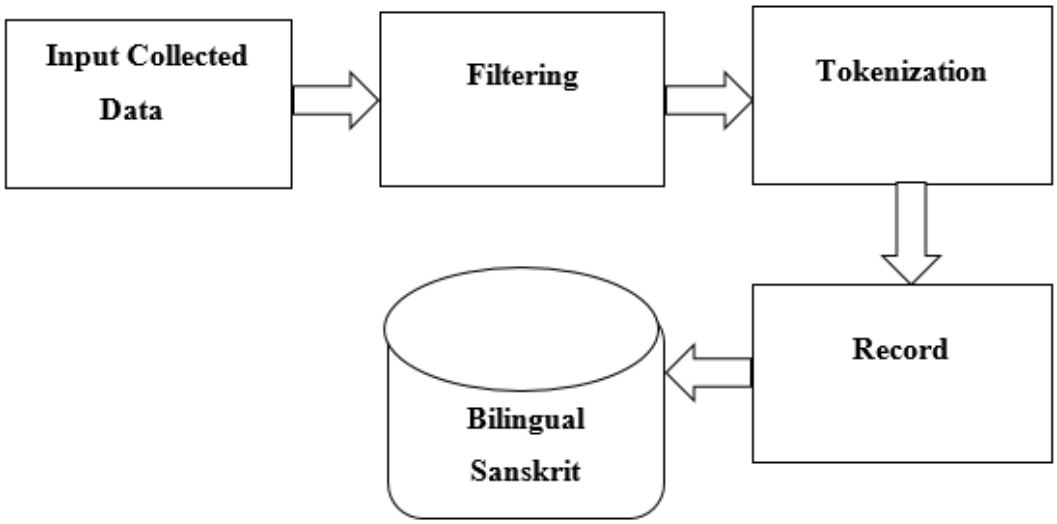

Figure 3.4: Bilingual Sanskrit Corpus Creation Flow

3.3.4.1 Input Collected Data: In this step, we input the collected and analysis data to the system. So, that the process of corpora creation could be started. The quality of corpus is directly-depends on the quality of collected text. For quality text collection we have to cover all the frequent domain that are essential for day to day life communication.

3.3.4.2 Filtering: Text is divided into sentences and each sentence is divided into words and given as input to the Filter. Because raw text may contain any invalid words like unrelated word, lengthy words and other language words. So it is necessary to filter all

the unrelated words. If all the unrelated words are stored directly then there may be a chance that system behave unrecognizable and the output must not be correct.

3.3.4.3 Tokenization: In this step, the word is divided into two parts, root and suffix part, i.e. we perform sandhi-witched. The information of each word is stored on the basis of suffix. Suffix plays a key role in our corpus and it is responsible for performing text-independent machine translation and also it works as a tagger at semantic and syntactic analysis. At, this stage we create a suffix table that store all the suffix part of our data.

For example:

All the words are tokenized into root and suffix word. The suffix table work as a parent table in our corpus.

i. **रामः** is divided into two parts **राम** and **अः**. **राम** is root word and **अः** is suffix.

ii. **लता** is divide into two parts **लत** and **आ** . **लत** is root word and **आ** is suffix

iii. **रामाभ्याम** is divided into two parts etc.

Table 3.1: Sandhi Witched

Actual Word	Root Word	Suffix
रामः	राम	अः
लता	लत	आ
रामाभ्याम	रामा	आभ्याम
नदी	नद	इ
नदीषु	नद	ईषु
रामेभ्यः	राम	एभ्यः
पठत्	पठ	एत्

3.3.4.4 Record:

After performing tokenization, the next step is to save the data into the corpus. We maintain separate files for pronoun, interrogative words, Shabdh, lakaar, verb and common words. Recording of word and sentences is different. We analyzed that words selected from sentences give more accurate result and make corpora more strong.

Figure 3.5, represents the various next phases of methodology. These steps includes designing an interface, designing a transliteration algorithm and validate corpus.

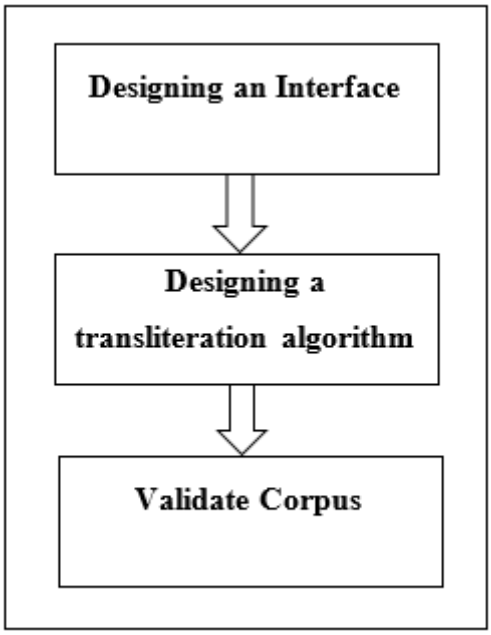

Figure 3.5: Phases of Methodology-II

3.3.5 Interface Design:

In this step, we designed an interface through which a user can interact with our system. The user interface is built by using some user interface control like buttons, Text area, panel, label and Frame. The interface is user friendly, i.e. everyone can operate software very easily. The implementation is done in java language. The below figure shows the interface of our system.

Whatever we write in the text area 1, the corresponding Sanskrit will be transliterated to the below to the text area 1.

3.3.5.1 Input method:

In this system user can input the text into two ways:

i. User can type the text in English in text box 1 and the corresponding Sanskrit will be transliterated in to the Sanskrit text area. For this process we use English and Sanskrit Unicode, which make one to one mapping between English alphabets to Sanskrit alphabets Unicode.

ii. User can directly paste the Sanskrit story into the Sanskrit related text area.

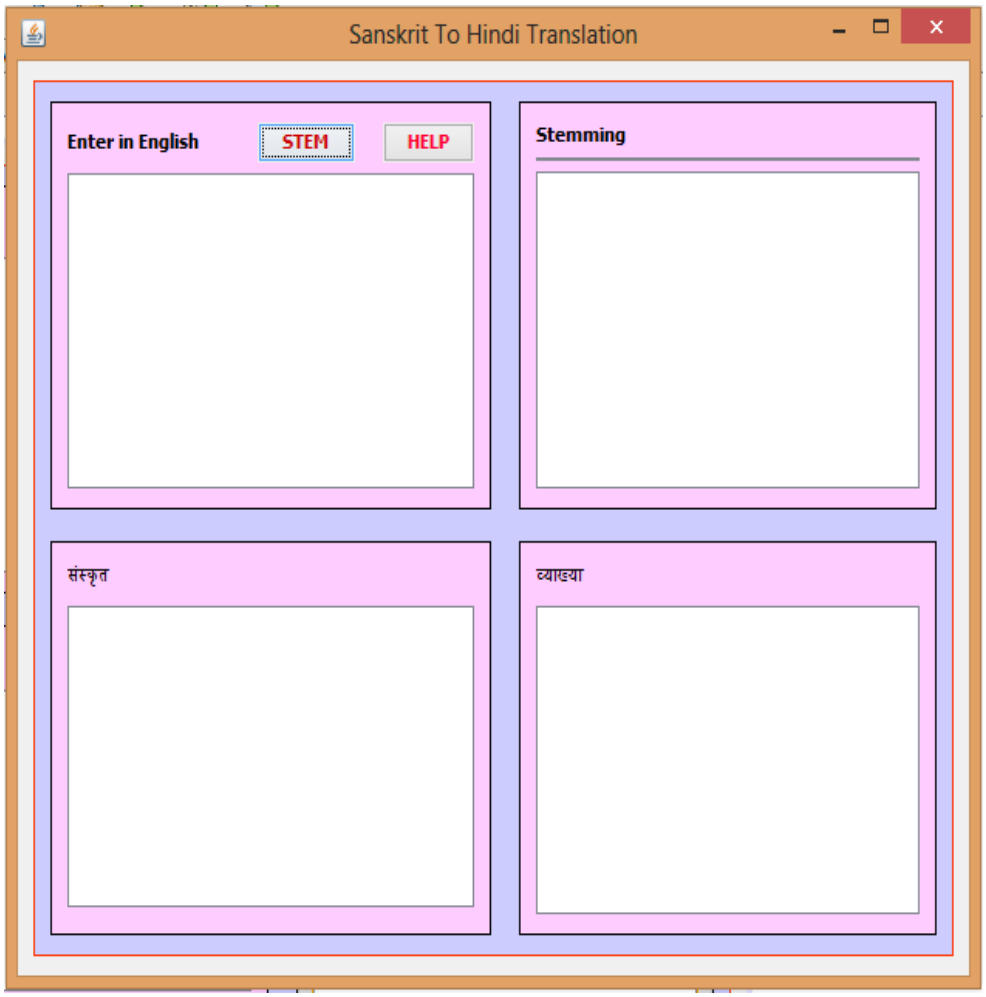

Figure 3.6: Interface of the System

3.3.5.2 Help Module:

This help window helps the user to write Sanskrit word correctly [23]. This window will be more helpful for those who are not native Sanskrit speaker or the beginner of Sanskrit. In this help window we display those words that are difficult for user to type. It contains both vowels and consonants as show in below figure 3.7.

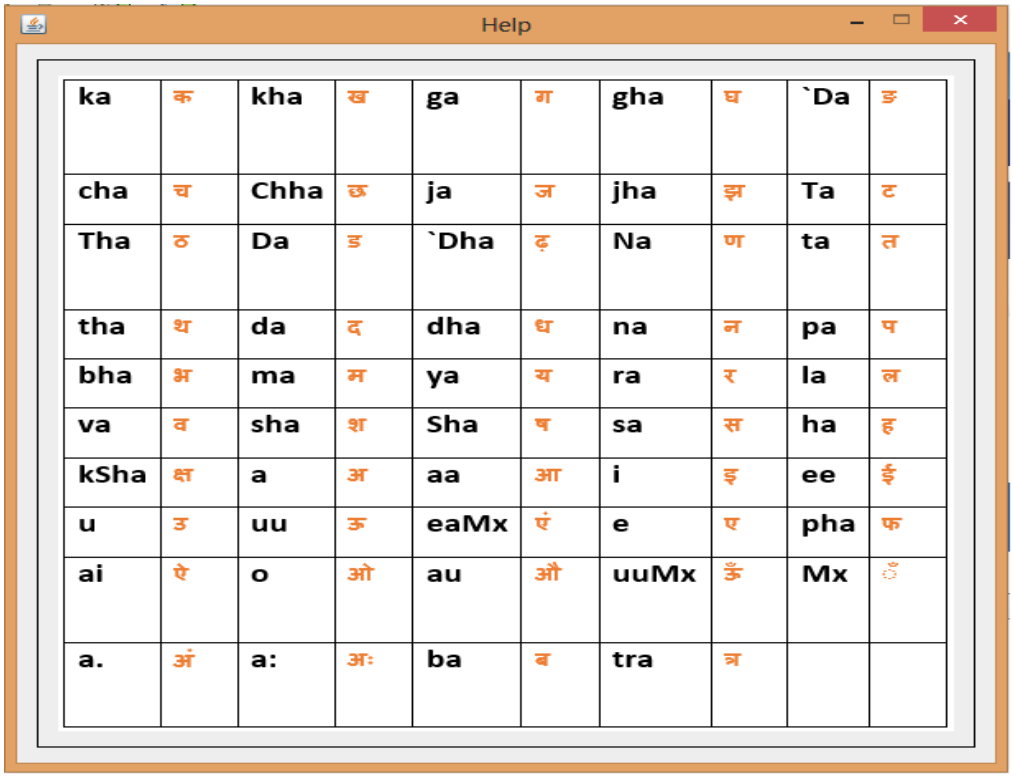

Figure 3.7: Help Interface for Input

3.3.6 Transliteration Algorithm:

Whatever you are writing in English in above interface in text area 1, the transliteration algorithm will convert it into the corresponding Sanskrit language. This process is known as language transliteration [29]. For this, we use Unicode of English and Sanskrit language, which will make the one to one mapping between the English to Sanskrit alphabets [31].

Figure 3.8: Unicode

Algorithm: This algorithm is used for language transliteration. The text is converted from English text to Sanskrit character by character. This is the most general algorithm for language transliteration.

Transliteration (English [], Sanskrit [], textarea, textarea1)

English [] is used to store English alphabets

Sanskrit [] is used to store Unicode numbers

Textarea contain English text written by user

Textarea1 contain Sanskrit text

t:= index number

For each (char c in textarea)

```
{       t:= find position of c in English []
        If(c is in English [])
                Textarea1=Sanskrit[t];
}
```

3.3.7 Validate Corpora:

The final step is to validate the corpus. For validating our created corpus we use Natural language processing steps like lexical analysis, syntax analysis and semantic analysis. Discourse and pragmatic analysis are the future of research area.

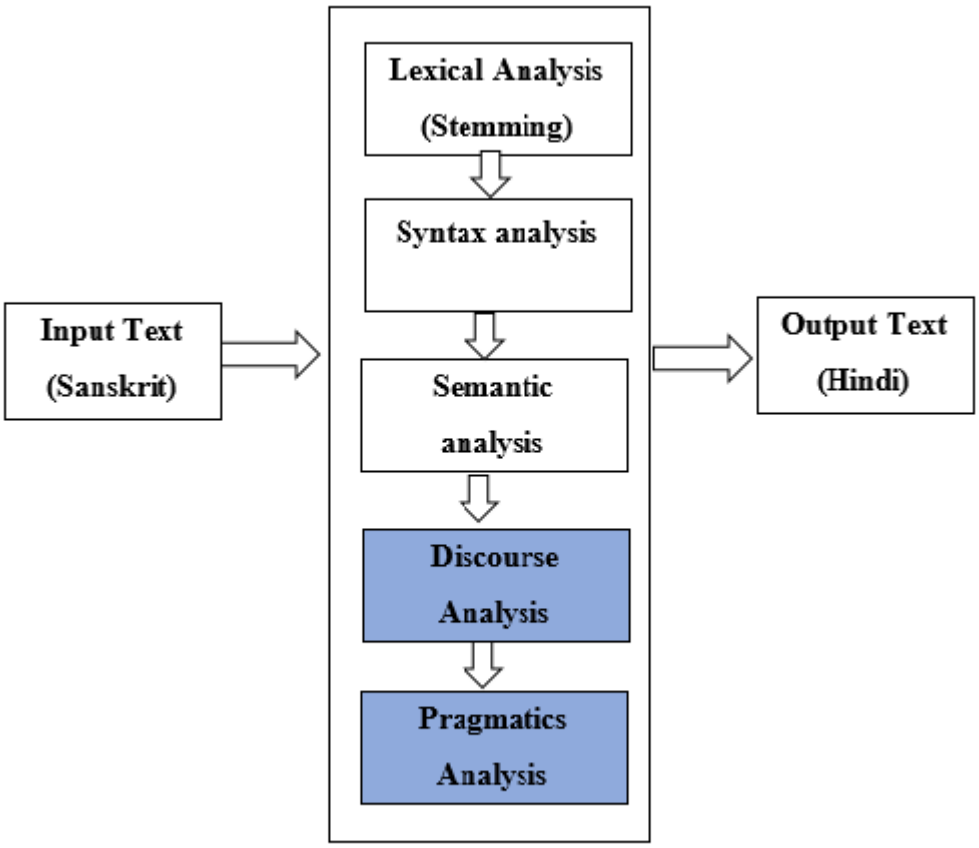

Figure 3.9: Processing steps to validate corpus

3.3.7.1 Lexical Analysis:

Lexical analysis is the concept of characters. The text is divided into tokens, tokens are nothing but a sequence of word.

1. The text is entered into the textarea.
2. Space between two words are considered as completion of single word.
3. The symbols like "|", "!", "." are considered as delimiter, denote the end of the sentence.

Algorithm:

The text entered in textarea 1 in user interface are stored in INTEXT. This algorithm performs tokenization. That is the word is divided into two parts: root word and suffix. The root word is stored in BASEWORD and suffix is stored into the SUFFIX.

Rules:

1. Vowel + Vowel=invalid
2. Consonant + Consonant=invalid
3. Consonant + Vowel =valid
4. Vowel + Consonant= invalid

LexemeGeneration (INTEXT, BASEWORD, SUFFIX, WORD [])

INTEXT is used to store input story.

BASEWORD is used to store root word.

SUFFIX is used to store suffix part.

{

Split story into sentences by the help of delimiters.

The sentences is divided into words and all the words are stored in an array WORD [].

For each (word w in WORD [])

{

w is tokenized into root and suffix part by using above rules 1 to 4.

BASEWORD=root

SUFFIX=suffix

}

}

3.3.7.2 Syntax and Semantic Analysis:

Syntax analysis are used to check the grammatical correctness of the sentences. There are various techniques to check the correctness. The most popular is parsing tree. Parsing tree parse the sentence and tag each word like Noun, pronoun, verb, adjective, common noun etc. Tagging can also be used to solve the problem of word disambiguation. Because some of the word have many meaning.

If any sentence is grammatical correct, it does not mean that it is semantically right. The main goal of semantic analysis is to resolve the problem of word ambiguity. So, semantic analysis identify the sense of the sentence and then use the correct meaning of the word. This is called word sense disambiguation.

Algorithm

Translation (Basew, NNP, PRN, VB, SUFFIX, NOUNSUF, CMNWORD, VERBSUFFIX)

Translation algorithm will use stemming algorithm to find the root and suffix word. NNP represents the noun, PRN denotes the pronoun and VB denotes the verb. Basew is used as temporary variable to store each word for processing.

{

Split story into sentences.

Call Stemming algorithm.

If Basew: =Empty && SUFF: = Empty then exit.

If Basew: = NNP && SUFF: = NOUNSUF then search the corpora to find the valid details and store these values (Basew and SUFF) for future use.

If Basew: = VB && SUFF: = VERBSUFFIX search the corpora to find the valid details and store these values (Basew and SUFF) for future use.

If Basew: = PRN then print detail of the word and copy the index position

}

CHAPTER 4

RESULTS AND DISCUSSIONS

This work will generate the corpora for Sanskrit to Hindi language translation. For validating the corpus, we use NLP techniques. At syntax analysis we perform grammatical validation and at semantic analysis we perform the word sense disambiguation.

4.1 USER INTERFACE FOR VALIDATING BILINGUAL CORPUS:

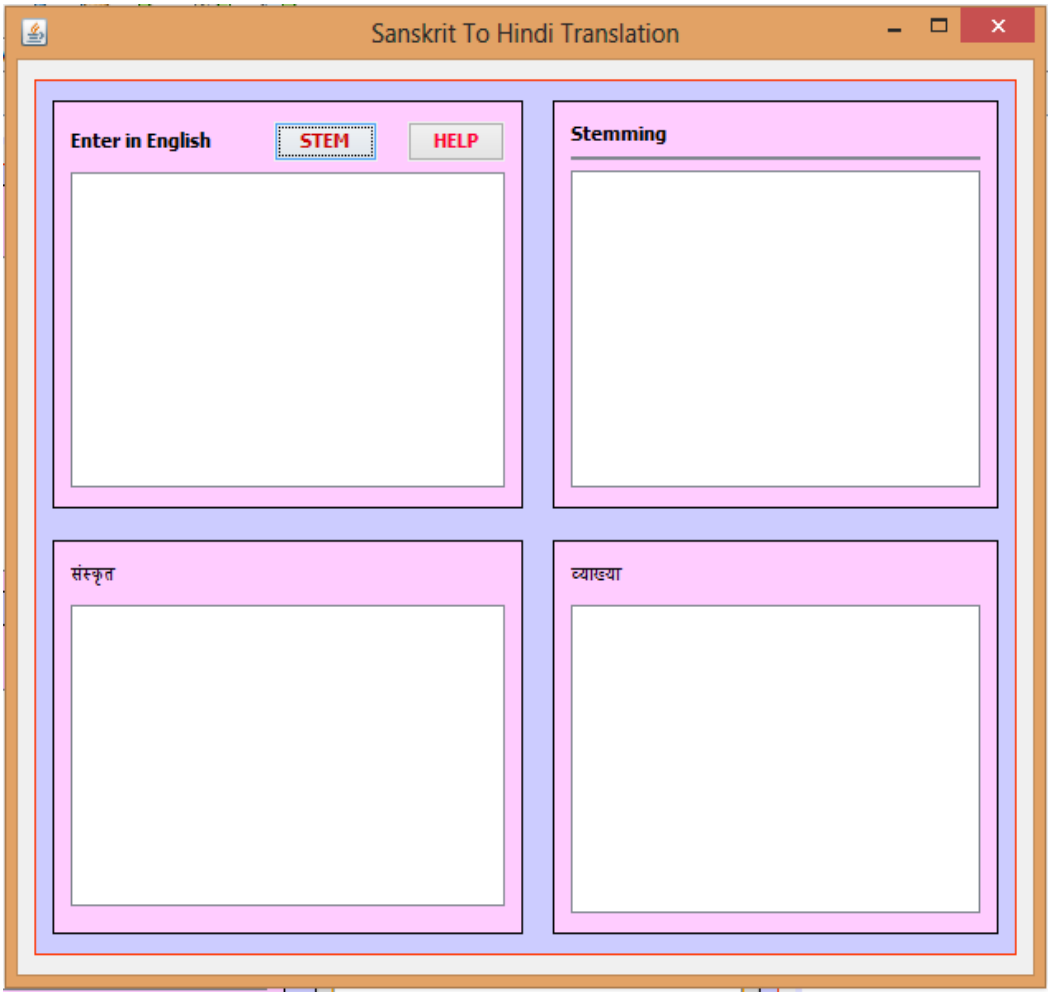

Figure 4.1: User Interface of System

The outcome of this research is creation of corpora for Sanskrit to Hindi language. This system can be used by novel writers, teachers and students to learn Sanskrit efficiently. This system can also be used by government to convert our ved puranas written in Sanskrit language to Hindi to increase the awareness of Indian history.

4.2 RESULTS:

There are various objective of our research. So, we discuss our all the objectives one by one and at the end we provide the accuracy of the system. The list of objectives is as follows:

1. **Transliteration:**

   Whatever we are typing in the English in the user interface of the proposed system the corresponding Sanskrit is transliterated in the text area. This is the first objective of our system as shown in the figure 4.2.

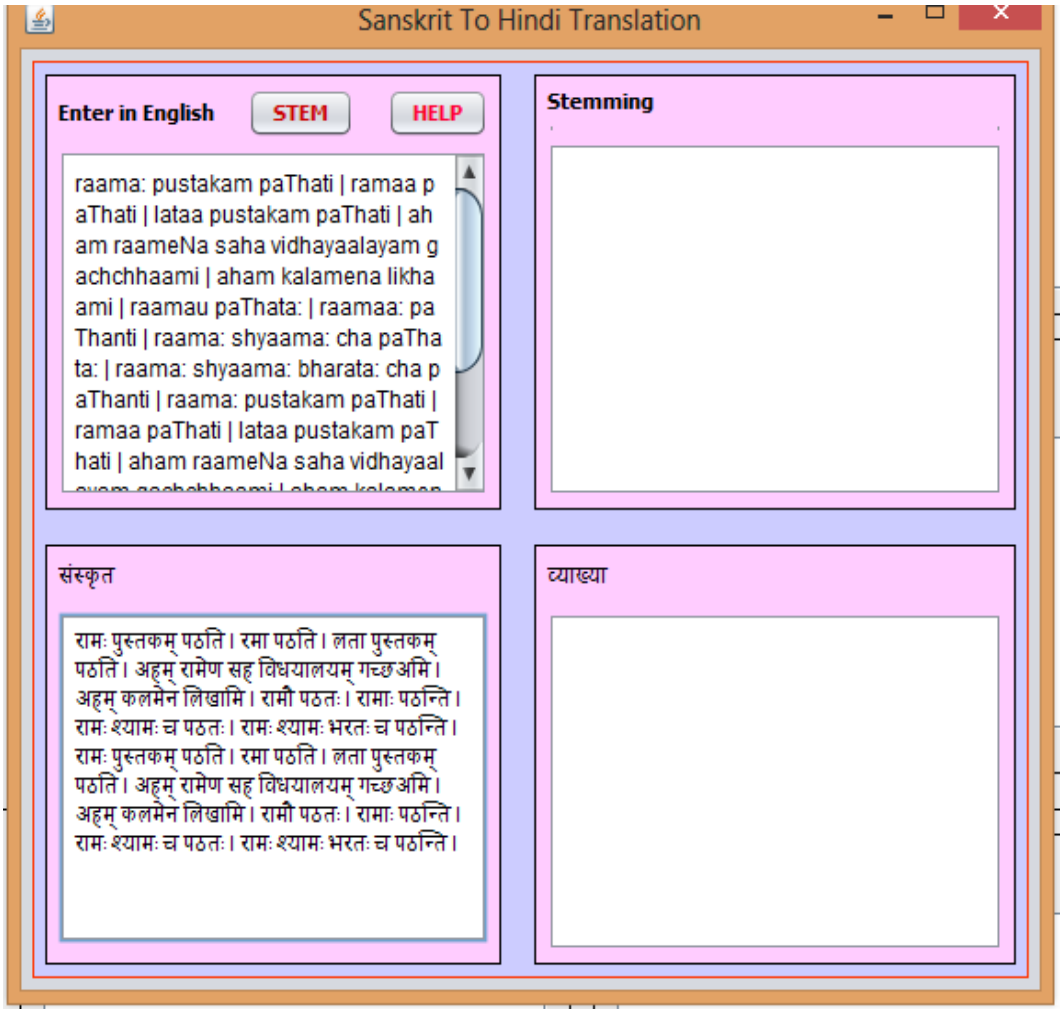

Figure 4.2: Transliteration from English to Sanskrit Language

2. **Stemming:** Stemming is used to perform sandhi witched. It divides the word into two parts: root and suffix parts. This system will perform sandhi witched for all the Sanskrit words. The sandhi witched of Sanskrit paragraph is performed by our system is show in figure 4.3.

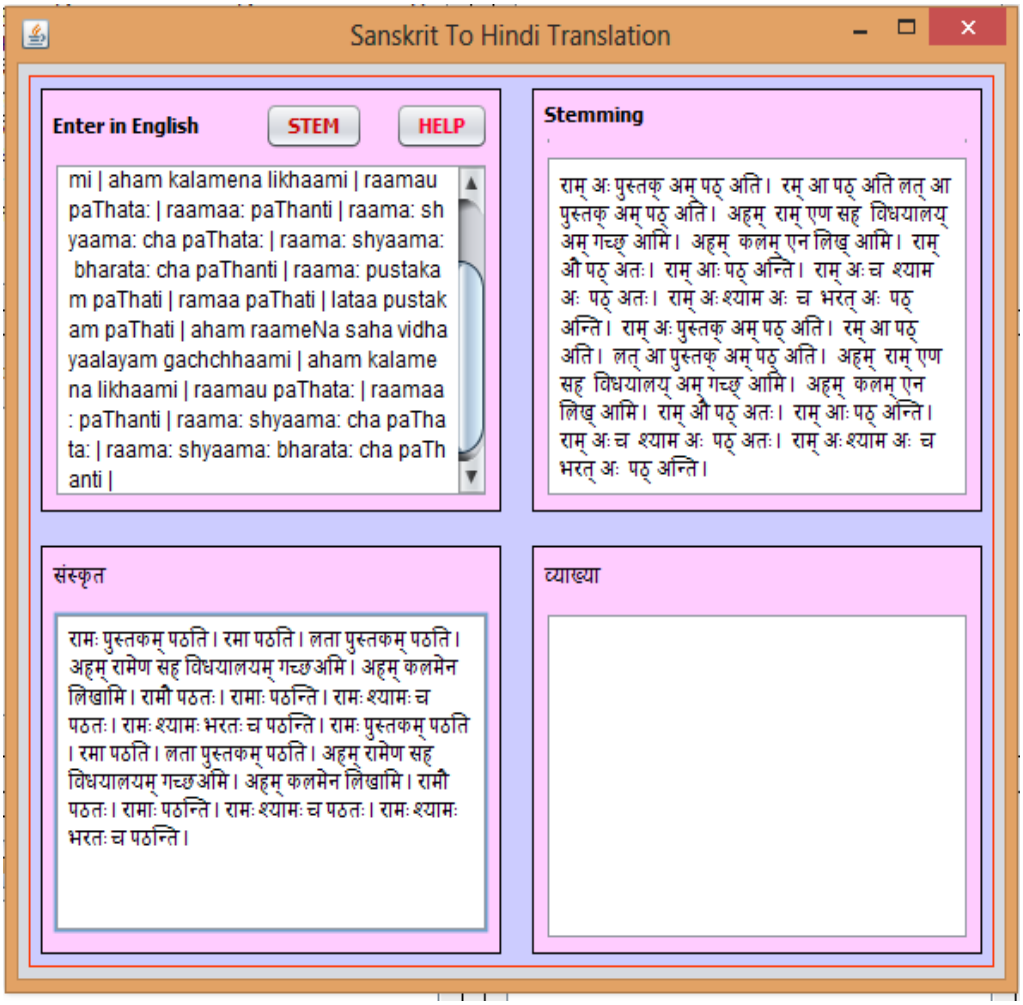

Figure 4.3: Sandhi Witched of Sanskrit

3. **Bilingual Corpora verification of different words:** Our corpora contains a large collection of vocabulary. Vocabulary contains all the details about Shabdhroop and lakaar data. Corpora also contains good collection of Verb, it cover approximately all the details about pronoun and interrogative words. Now, we show one by different words recognize by our system correctly.

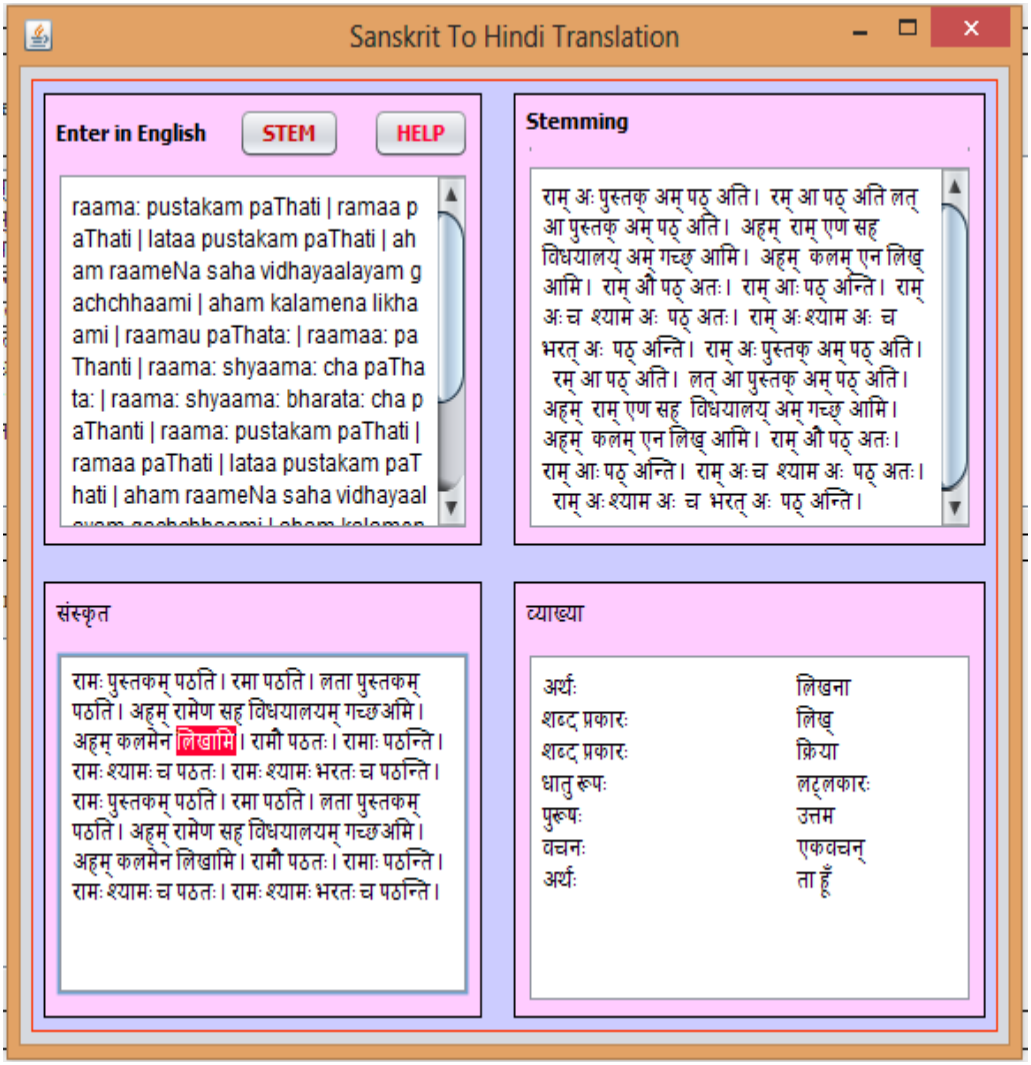

Figure 4.4: Description of the word लिखामि

It is clearly visible in figure 4.4 that there are four textarea. Whatever we are writing in the textarea 1, the corresponding Sanskrit is reflected to the below of the textarea 1. There are two buttons Help and Stem. When we clicked on Help, a help window is displayed that provides information related to transliteration. When we clicked on the Stem button then tokenization is performed. That is the words are divided into two parts: root and suffix part. Finally, when we clicked on the Sanskrit word, then their corresponding Hindi translation is displayed in the textarea 4 and the clicked word is visible in red colour.

For more results refer **Appendix B.**

4.3 TESTING OF THE SYSTEM

The implemented system was tested for different vocabulary of Sanskrit language and the corresponding Hindi details is displayed. The results acquired were good for Sanskrit corpora. The obtained results are cross verified by using Sanskrit teachers and text book of U.P. State Board. Now, we show the comparison between the actual result and the result obtained by our tool.

Accurate Considerable Not Relevant

Table 4.1: Result Analysis

S. No.	Sanskrit word	Actual Detail	Generated By Tool
1	राम:	वचनः एकवचन् विभक्तिः i अर्थः ने शब्द प्रकार् अ लिंग प्रकार् पुल्लिन्ग्	वचनः एकवचन् विभक्तिः i अर्थः ने शब्द प्रकार् अ लिंग प्रकार् पुल्लिन्ग्
2	भोजनम्	वचनः एकवचन् विभक्तिः ii अर्थः को शब्द प्रकार् अ लिंग प्रकार् नपुन्लिन्ग्	वचनः एकवचन् विभक्तिः ii अर्थः को शब्द प्रकार् अ लिंग प्रकार् नपुन्लिन्ग्
3	रामेण	वचनः एकवचन् विभक्तिः ii अर्थः से/के साथ/के द्वारा शब्द प्रकार् अ लिंग प्रकार् पुन्लिन्ग्	वचनः एकवचन् विभक्तिः ii अर्थः से/के साथ/के द्वारा शब्द प्रकार् अ लिंग प्रकार् पुन्लिन्ग्
4	नदिषु	वचनः बहुवचन् विभक्तिः vii अर्थः ओं में/पर शब्द प्रकार् इ लिंग प्रकार् स्त्रिलिन्ग्	वचनः बहुवचन् विभक्तिः vii अर्थः ओं में/पर शब्द प्रकार् इ लिंग प्रकार् स्त्रिलिन्ग्

5	सः	अर्थः वह शब्द प्रकार् सर्वनाम लिंग प्रकार् पुल्लिन्ग्	अर्थः वह शब्द प्रकार् सर्वनाम लिंग प्रकार् पुल्लिन्ग्
6	लता	वचनः एकवचन् विभक्तिः i अर्थः ने शब्द प्रकार् आ लिंग प्रकार् स्त्रिलिन्ग्	वचनः एकवचन् विभक्तिः i अर्थः ने शब्द प्रकार् आ लिंग प्रकार् स्त्रिलिन्ग्
7	तौ	अर्थः वह दोनो लिंग प्रकार्ः पुल्लिन्ग्	अर्थः वह दोनो लिंग प्रकार्ः पुल्लिन्ग्
8	किम्	अर्थः क्या लिंग प्रकार्ः नपुन्लिन्ग्	अर्थः क्या लिंग प्रकार्ः नपुन्लिन्ग्
9	कलमेन	वचनः एकवचन् विभक्तिः iii अर्थः से/के साथ/के द्वारा शब्द प्रकार् अ लिंग प्रकार् नपुन्लिन्ग्	वचनः एकवचन् विभक्तिः iii अर्थः से/के साथ/के द्वारा शब्द प्रकार् अ लिंग प्रकार् पुन्लिन्ग्
10	अयोध्यात्	वचनः एकवचन् विभक्तिः v अर्थः से प्रथक शब्द प्रकार् अ लिंग प्रकार् नपुन्लिन्ग्	वचनः एकवचन् विभक्तिः v अर्थः से प्रथक शब्द प्रकार् अ लिंग प्रकार् पुन्लिन्ग्
11	पठति	अर्थः पढना शब्द प्रकारः पठ् शब्द प्रकारः क्रिया धातु रूपः लट्लकारः पुरूषः प्रथम वचनः एकवचन् अर्थः ता है	अर्थः पढना शब्द प्रकारः पठ् शब्द प्रकारः क्रिया धातु रूपः लट्लकारः पुरूषः प्रथम वचनः एकवचन् अर्थः ता है

In the above table, we had shown the results of some sanskrit words with their actual and the tool generated description. In which the orange colored description are exactly matched with the orignal titles which are already present with the word. The green color indicates that the description are relevant according to the survey of various vocabulary. Green colored descripton are relevant and correct according to the validation of various sanskrit professional tutors and grammatical rules. The red colored description are somewhat not relevant, but only some of the descriptions are red colored which are generated by our system. Our system will generate description of the sanskrit.The results generated by the system are satisfactory according to various Sanskrit tutors and Sanskrit language students. Our system is approximately generating the correct description. According to testing of the system 90% of the descripton are accurate and relevant.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

The main task of this thesis is to provide the bilingual corpora information from Sanskrit to Hindi language. The collected data works well as we shows the results.

5.1 Conclusion

This research work presents different algorithms to create bilingual corpora. The algorithms are designed in such a way that they make text-independent translation. In this proposed system, Bilingual Corpora creation for Sanskrit language is created using NLP. This system can be used by scholars, technical writers, students and teachers. The quality of the translation is depend on the size of corpora. Corpora will include verbs, pronouns, interrogative words, lakaar etc. It will be an educational device for the students. They can learn the basics of the Sanskrit language quickly. By using this tool faculties can explain each word very easily and also helps to increasing learnability. It promotes the more research in the field of Sanskrit language.

5.2 Future Scope

This research work can further be extended to perform discourse and pragmatic analysis to the Sanskrit story. This work can also be extended to perform text independent machine translation from Sanskrit to Hindi. It can also be used in the field of automatic title generation and story summarization.

The techniques used here for designing corpora for Sanskrit to Hindi language can also be used for other language like Tamil, Telugu, and Kannada etc. It can be the basis for any machine translation application.

REFERENCES

[1] Md. Khalilur Rhaman (2012) "*A Rule Based Approach for Implementation of Bangla to English Translation*", International Conference on Advanced Computer Science Applications and Technologies, Kuala Lumpur, pp. 13-18.

[2] Sangavi G (2016) "*Analysis on Bilingual Machine Translation Systems for English and Tamil",* International Conference on Computation of Power, Energy Information and Communication, pp. 245-250.

[3] Namrata Tapaswi (2015*) "An Approach for Grammatical Constructs of Sanskrit Language using Morpheme and Parts- of-Speech Tagging by Sanskrit Corpus",* International Journal of Latest Trends in Engineering and Technology (IJLTET), Volume 5-No. 3, pp. 476-483.

[4] Shahnawaz (2015) *"Conversion between Hindi and Urdu",* International Conference on Computing, Communication and Automation, pp. 309-313.

[5] Ved Kumar Gupta, Prof. Namrata Tapaswi and Dr. Suresh Jain (2013) *"Knowledge Representation of Grammatical Constructs of Sanskrit Language Using Rule Based Sanskrit Language to English Language Machine Translation",* International Conference on Advances in Technology and Engineering (ICATE-2013), Mumbai, pp. 1-5.

[6] Farshad Kiyoumarsi (2014) *"Evaluation of Automatic Text Summarizations Based on Human Summaries",* 2nd global conference on LINGUISTICS and FOREIGN LANGUAGE TEACHING (LINELT-2014), Dubai, pp. 83-91.

[7] Sarita G. Rathod, Shanta Sondur (2012) *"English to Sanskrit Translator and Synthesizer (ETSTS)",* International Journal of Emerging Technology and Advanced Engineering, Vol. 2-No. 12, pp. 360-380.

[8] N. Murali, Dr. R.J. Ramasreee 2and Dr. K.V.R.K. Acharyulu (2014) *"KRIDANTA ANALYSIS FOR SANSKRIT"*, International Journal on Natural Language Computing, Vol. 3-No.3, pp. 33-49.

[9] Namrata Tapaswi and Dr. Suresh Jain (2011) *"Morphological and Lexical Analysis of the Sanskrit Sentences"*, MIT International Journal of Computer Science & Information Technology, Vol. 1-No. 1, pp.28-31.

[10] Sarita G. Rathod (2014) *"Machine Translation of Natural Language using different Approaches: ETSTS (English to Sanskrit Translator and Synthesizer)"*, International Journal of Computer Applications, Volume 102– No.15, pp. 26-31.

[11] Nandini Sethi and Prateek Agrawal (2016) *"Automated Title Generation in English Language Using NLP"*, International Journal of Control Theory and Applications, Volume 9-No.11, pp. 5159-5168.

[12] Wei Yen Chong (2014) *"Natural Language Processing for Sentiment Analysis"*, 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology (ICAIET-2014), pp. 212-217.

[13] Damodar Magdum and Manisha Shukla (2015) "*Methodology for designing and creating Hindi speech corpus*", International Conference on Signal Processing and Communication Engineering Systems (SPACES), pp. 336-339.

[14] Pramod Salunkhe and Shrikant Jadhav (2016) "*Hybrid Machine Translation For English to Marathi: A Research Evaluation In Machine Translation",* International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 924-931.

[15] N. Ashraf and M. Ahmad (2015) "*Machine Translation Techniques and their Comparative Study",* International Journal of Computer Applications Volume 125 – No.7, pp. 25-31.

[16] Sunita Chand (2016) *"Empirical Survey of Machine Translation Tools",* International conference on Research in Computational Intelligence and Communication Networks (ICRCICN), pp. 181-185.

[17] Kanika, Ankur and Divyanjali (2014) "*A Review of English to Indian Language Translator: Anusaaraka",* International Conference on Advances in Computer Engineering & Applications (ICACEA-2014), GZB, pp. 1-6.

[18] Shaharban T.A (2016) *"Pragmatic analysis of Malayalam sentences",* International Conference on Inventive Computation Techniques (ICICT-2016)

[19] B.N.V Narasimha Raju and M S V S Bhadri Raju (2016) *"Statistical Machine Translation System for Indian Languages",* 6th International Advanced Computing Conference (IACC-2016), pp. 174-177.

[20] Pooja malik and Anurag Singh Baghel (2016) *"An Improvement in BLEU Metric for English- Hindi Machine Translation Evaluation",* International Conference on Computing, Communication and Automation (ICCCA2016), pp. 331-336.

[21] Pooja Alva and Dr. Vinay Hegde (2016) *"Hidden Markov model for POS tagging in Word Sense Disambiguation",* International Conference on Computational Systems and Information Systems for Sustainable Solutions (ICCSISSS), pp. 279-284.

[22] Mrs. Namrata Tapaswi, Dr. Suresh Jain Mrs. Vaishali Chourey (2012) *"Parsing Sanskrit sentences using lexical functional grammar"*,International Conference on Systems and Informatics (lCSAI 2012), Yantai, pp. 2636-2640.

[23] Leena Jain and Prateek Agrawal (2015)," *Text independent root word identification in Hindi language using natural language processing*", Int. J. Advanced Intelligence Paradigms, Vol. 7-No. 3/4, pp. 240-249

[24] "*Natural Language Understanding*" by James and Allen. The Benjamin/Cummings Publishing Company, Inc... First Edition, ISBN: 0-8053-0334-0

[25] "*Machine Translation an Introductory Guide*" by Douglas Arnold. Publisher: NCC Blackwell Ltd., First edition, ISBN: 1855542-17x.

[26] *"A Higher Sanskrit Grammar"* by M. R. Kale, Delhi M. Banarassidas Publisher, Fourth Edition, ISBN: 9788120801783

[27*] "machine-translation-process":*language.worldofcomputing.net/machine-translation/machine-translation-process.html //accesed on 21st Feb 2017

[28] "Java Basics," 2003: www.tutorialpoint.com/java/ //accesed on 11th Jan 2017

[29] P. Lavanya, P. Kishore, and G. Madhavi, "A simple approach for building transliteration editors for Indian languages," *J. Zhejiang Univ. SCI*, vol. 6A, no. 11, pp. 1354–1361, 2005.

[30] https://www.javatpoint.com/java-tutorial //accesed on 21st Jan 2017

[31] Varsha Tomar and Manisha Bhatia *"Localization of Text Editor using Java Programming,"* International Journal of Computer Applications, Vol. 89-No.12, pp. 49-54.

[32] M. Rajani Shree and Sowmya Lakshmi (2016) *"A novel approach to Sandhi splitting at Character level for Kannada Language,"* International Conference on Computational Systems and Information Systems for Sustainable Solutions, pp. 17-20.

APPENDIX A

A.1 Corpora Description

During Sanskrit corpora creation we collect the information of all types of Sanskrit words including Shabdh type, lakaar, verb, common words, pronoun, interrogative words and connective words. On the basis of this collected information we able to maintain our corpora in such a way that it is able to perform text-independent machine translation. The sample of the collected Sanskrit information is shown in below tables.

Table A.1: वर्तमान काल

वर्तमान काल			
प्रथमपुरुष	भवति	भवत:	भवान्ति
मध्यमपुरुष	भवसि	भवथ:	भवथ
उत्तमपुरुष:	भवामि	भवाव	भवाम

Table A.2: आज्ञा

आज्ञा			
प्रथमपुरुष	भवतु	भवताम्	भवन्तु
मध्यमपुरुष	भव	भवतम्	भवत
उत्तमपुरुष:	भवानि	भवाव	भवाम

Table A.3: विधि

विधि			
प्रथमपुरुष	भवेत्	भवेताम्	भवेयुः
मध्यमपुरुष	भवे:	भवेतम्	भवेत्
उत्तमपुरुष:	भवेयम्	भवेव्	भवेम

Table A.4: भूतकाल

भूतकाल			
प्रथमपुरुष	अभवत्	अभवताम्	अभवन्
मध्यमपुरुष	अभव:	अभवतम्	अभवत
उत्तमपुरुष	अभवम्	अभवाव	अभवाम

Table A.5: भविष्यत्काल

भविष्यत्काल			
प्रथमपुरुष	भविष्यति	भविष्यत:	भविष्यन्ति
मध्यमपुरुष	भविष्यसि	भविष्यथ:	भविष्यथ
उत्तमपुरुष:	भविष्यामि	भविष्याव:	भविष्याम:

Table A.6: Common Words

शब्दः	अर्थ	लिङ्
भ्राता	भाई	पुल्लिङ्ग
भगिनी	बहन	स्त्रिलिन्ग
सुता	पुत्र	पुल्लिङ्ग
मातुला	मामा	पुल्लिङ्ग
अश्व	घोडा	पुल्लिङ्ग
कन्दुक्	गेन्द	नपुन्लिङ्ग
अजा	बकरी	स्त्रिलिङ्ग
भोजन	खाना	नपुन्लिङ्ग
काल	समय	नपुन्लिङ्ग

Table A.7: Pronoun

शब्दः	अर्थ	लिङ्
स:	वह	पुल्लिङ्ग
अयं	यह	पुल्लिङ्ग
इमे	ये सब	पुल्लिङ्ग
इमा:	ये सब	स्त्रीलिङ्ग
सा	वह	स्त्रीलिङ्ग
तौ	वह दोनो	पुल्लिङ्ग
ते	वे सब	पुल्लिङ्ग

Table A.8: Interrogative Words

शब्दः	**अर्थ**	**लिङ्**
कदा	कब	नपुन्लिङ्ग
कः	कौन	पुल्लिङ्ग
कौ	कौन दोनो	पुल्लिङ्ग
का	कौन	स्त्रिलिङ्ग
के	कौन दोनो	स्त्रिलिङ्ग
किम	क्या	नपुन्लिङ्ग
कानि	कौन सब	नपुन्लिङ्ग

Table A.9: Verbs

शब्दः	**अर्थ**
पठ	पडना
पत	गिरना
वद	बोलना
क्रीड	खेलना
लिख	लिखना

APPENDIX B

B.1 Results of Bilingual Corpora verification

We test our system approximately for all the words that corpora contain. The results generated by our tool is satisfactory which are validated by Sanskrit textbook and Sanskrit teachers. Below figures shows the detailed result of the words that our corpora contain.

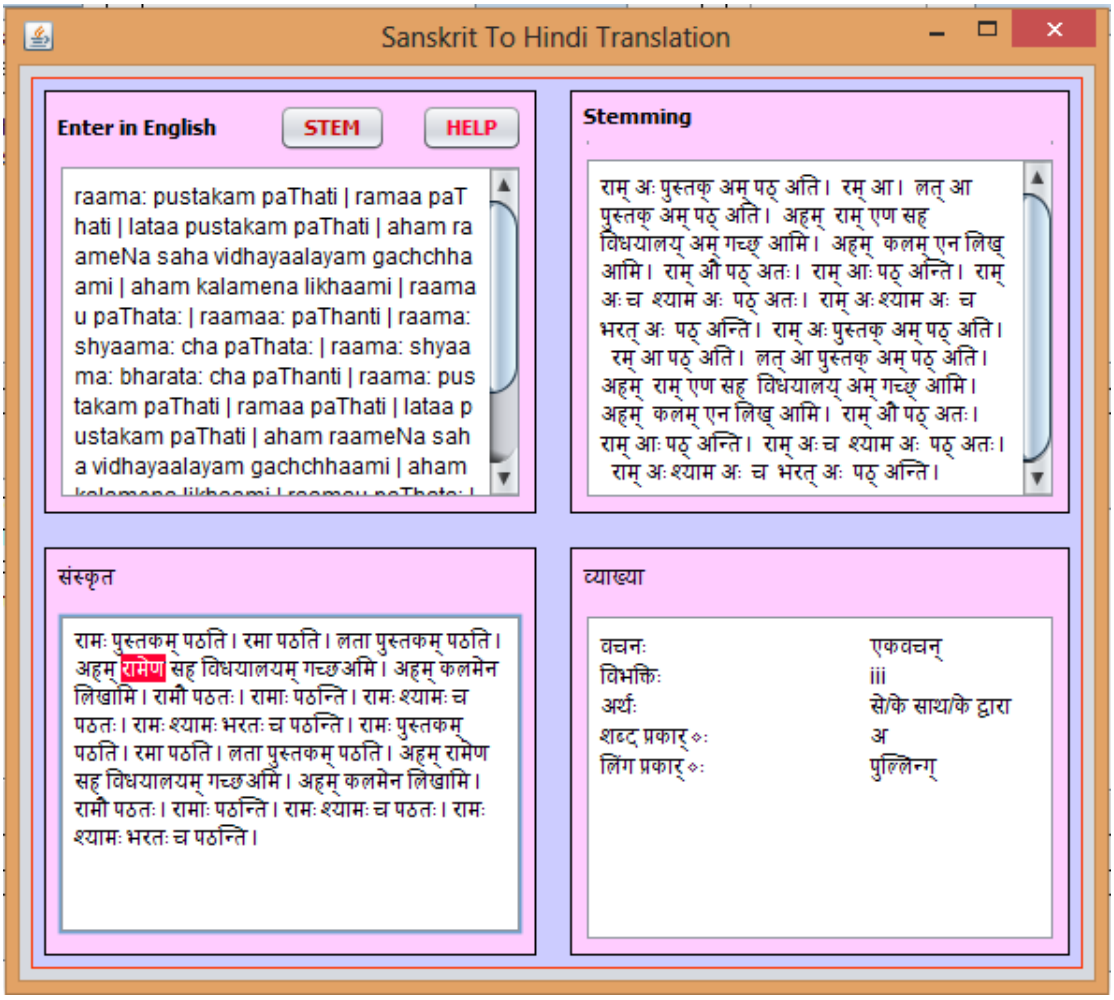

Figure B.1: Description of the Word रामेण

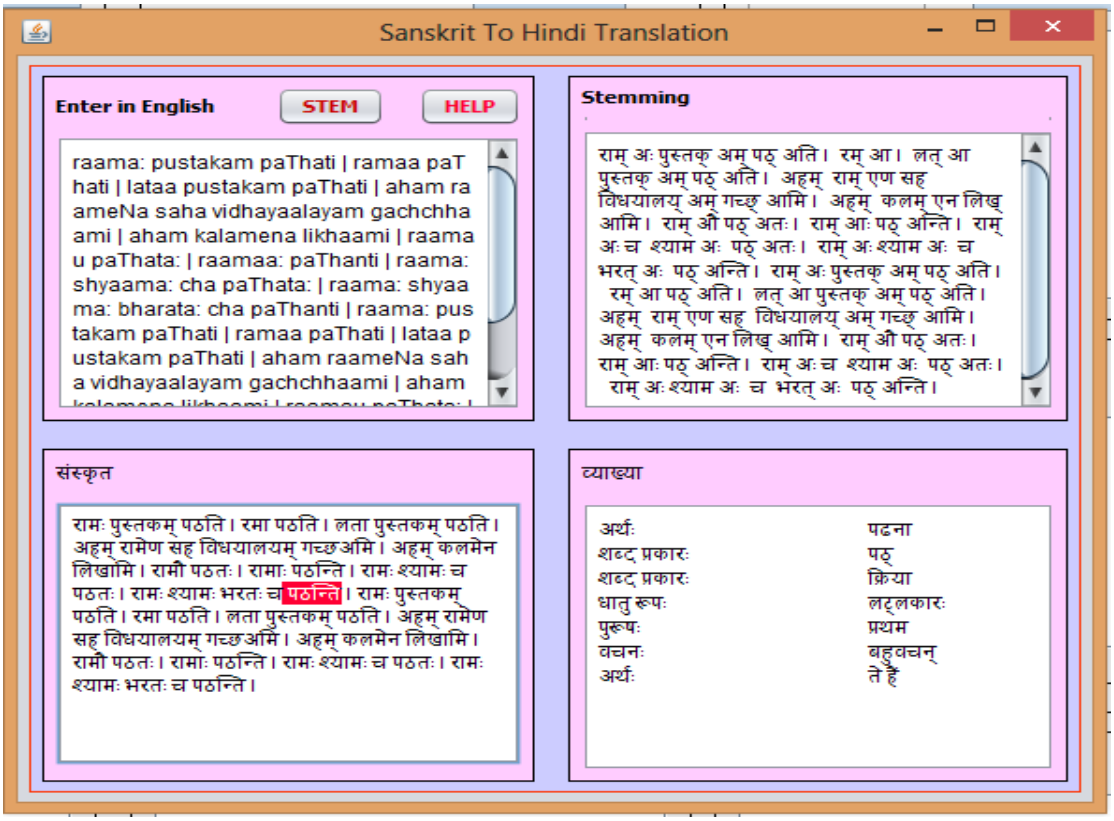

Figure B.2: Description of the Word पठन्ति

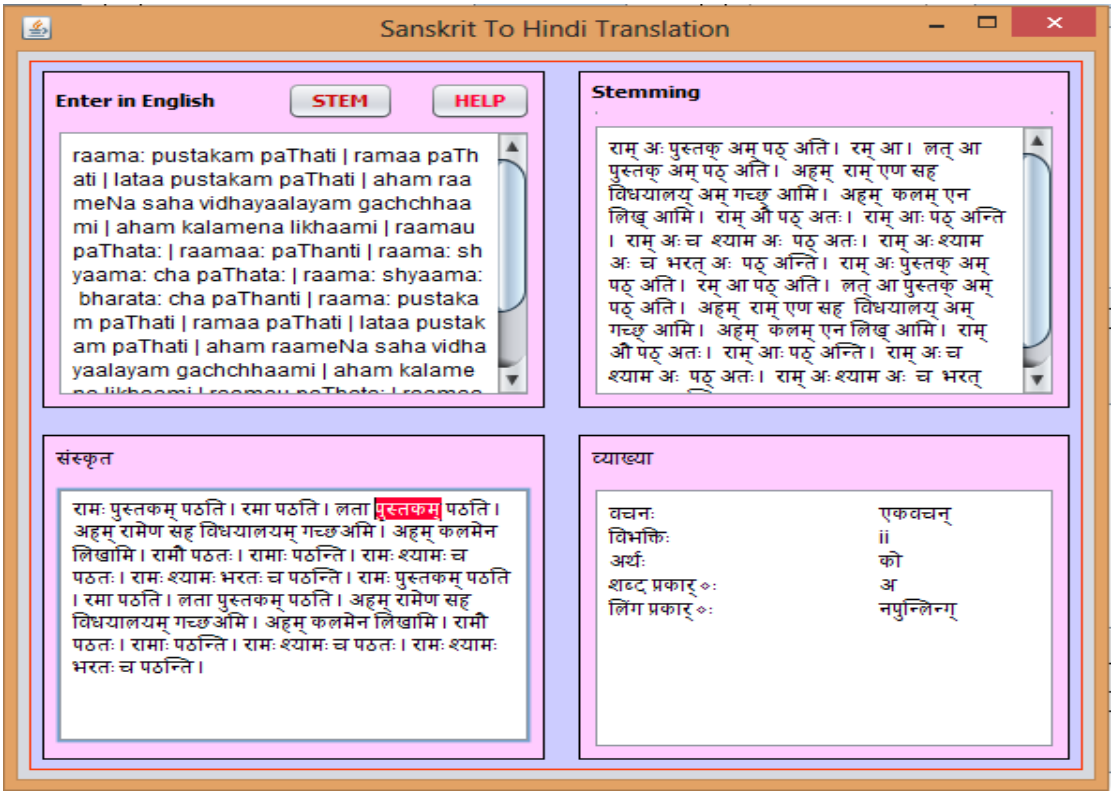

Figure B.3: Description of the Word पुस्तकम्

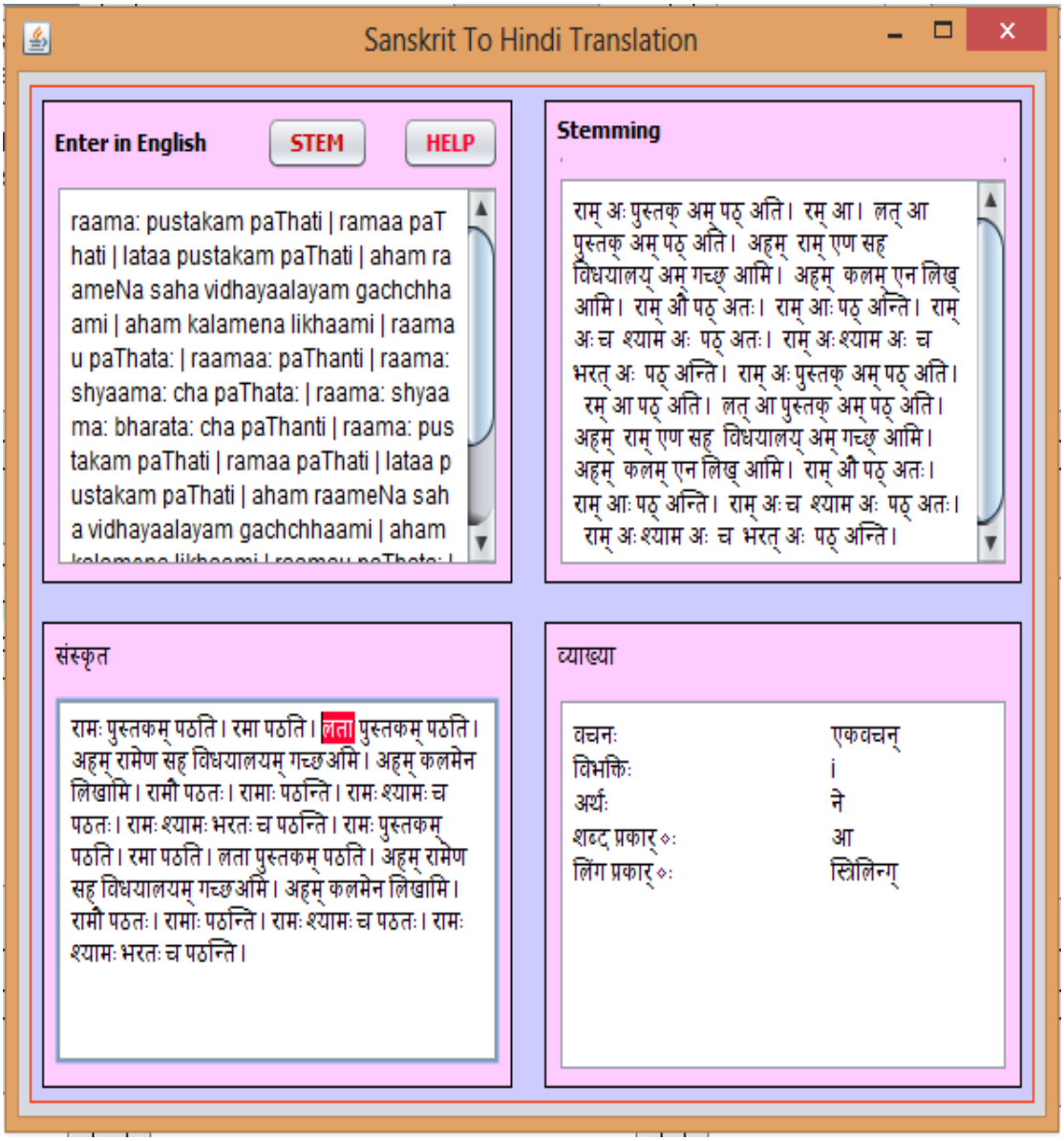

Figure B.4: Description of the Word लता

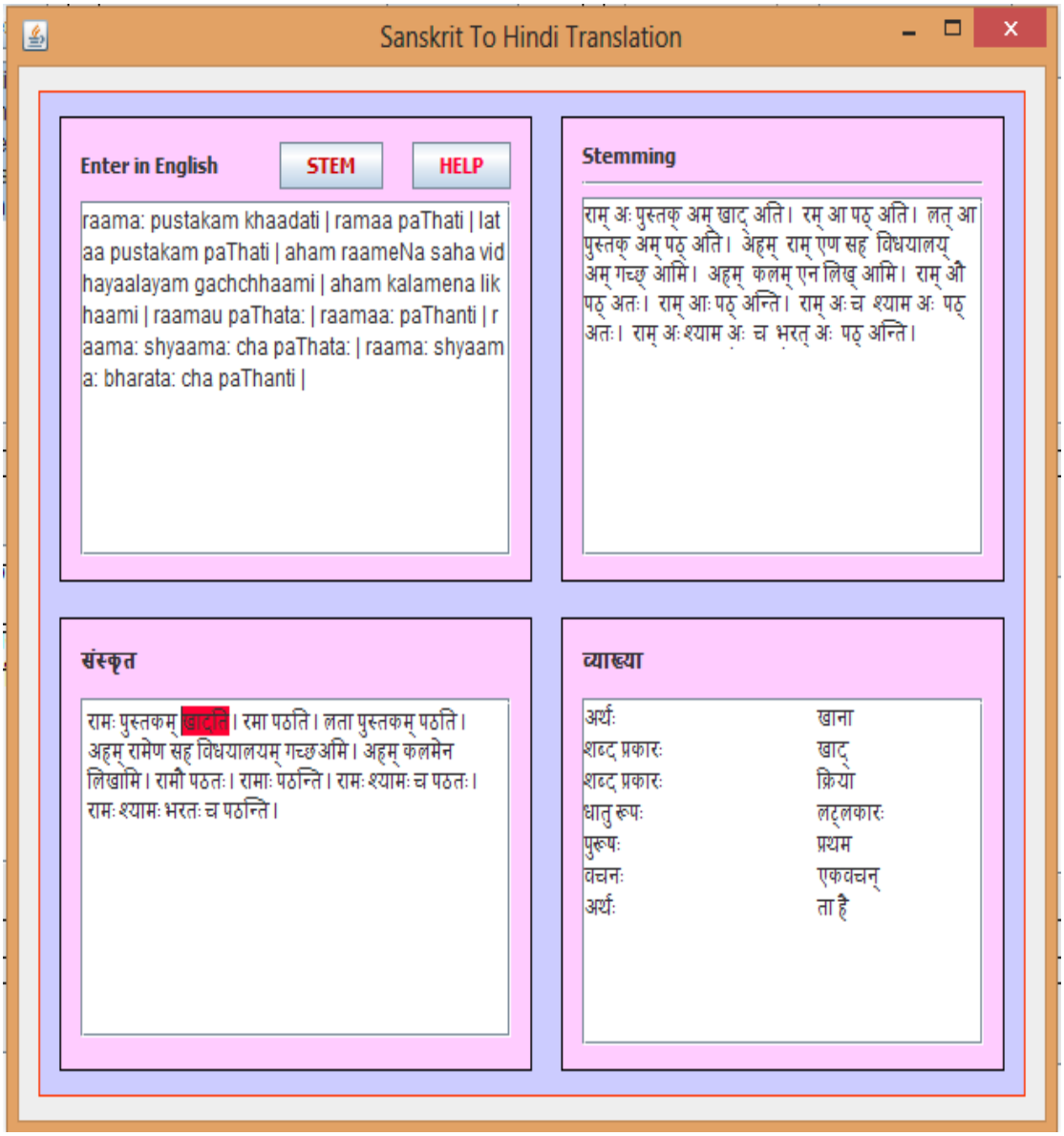

Figure B.5: Description of the Word खादति

PUBLICATIONS

LIST OF PUBLISHED PAPER

[1] Nandini Sethi, Prateek Agrawal and Anuj Kumar (2016) *"Automated Title Generation in English Language Using NLP"*, International Journal of Control Theory and Applications, Volume 9-No.11, pp. 5159-5168.