

## English to Tulu Translator

Amrutha Shenoy M<sup>1</sup>, M Praveen Rao<sup>2</sup>, Varsha K Shenoy<sup>3</sup>, Vignesh P Kudva<sup>4</sup>, Vasanth Nayak<sup>5</sup>

<sup>1</sup>Information Science and Engineering, Canara Engineering College (India)

<sup>2</sup>Information Science and Engineering, Canara Engineering College (India)

<sup>3</sup>Information Science and Engineering, Canara Engineering College (India)

<sup>4</sup>Information Science and Engineering, Canara Engineering College (India)

<sup>5</sup>Assistant prof., Department of Information Science and Engineering, Canara Engineering College. (India)

\*\*\*

**Abstract**—The aim of this project is to devise an effective application that translates English language to a meaningful sentence in Tulu language. It can help the rural people of West Karnataka who only have knowledge of Tulu language. Also, the number of people in contact with Tulu or English will expand because of the knowledge gained with the help of this system. It will help the people who face difficulties in speaking and understanding Tulu language. This system explores the translation process of a machine translation system from English to Tulu based on rule-based machine translation. For ruled based approach, considering the structural difference between English and Tulu languages, syntax transfer based on methodology is adopted for translation. This translation engine is a parser, which analyzes the English text, and the corresponding target language structure is generated through transfer lexicon. To generate the proper Tulu sentence Morphological generator for Tulu is required. English is a universal language and Tulu, the language used by the majority of people in West Karnataka, we propose an English to Tulu machine translation system.

**Keywords**—Language Barrier, Lexicon, Morphological generator, Rule Based Machine Translation, Statistical Machine Translation.

### 1. INTRODUCTION

India is a multilingual country; different states have different languages but not all Indians know all languages. Majority of Indians, especially the remote villagers, do not understand, read or write English, therefore implementing an efficient language translation system is needed. Machine translation system, that translate text form English to Tulu, will enhance the knowledge of Indians without any language barrier. It is a translation system that translates one language to another language. Whether it is human translation or machine translation, the substance of the text is of the most importance, i.e. the meaning of the textual matter in English must be properly

translated into the Tulu. Human translation is more accurate, but at the same time it is more time consuming, expensive and lacks universality. Due to its applicability in overcoming the language barrier, Machine Translation systems have gained interest of many researchers. Advanced research in the area of Artificial Intelligence and Computational Linguistics made a promising evolution of translation engineering. This helped in the development of usable Machine Translation Systems in certain well-defined areas. It is a difficult task to get High quality, fully automatic machine translation system. Many organizations like Microsoft, Google and many others are engaged in the development of Machine Translation systems. The purpose of machine translation is to conform to ever-growing demand of translation of text. There are diverse kinds of users of machine translation who use it for business such as data mining, publishing content on a big scale. Government use machine translation for intelligence gathering and monitoring actions. There is an industry known as Language Service Providers which works on increasing productivity and ensuring quality standards. To gain more accuracy in translating, many new researches works and developments are being performed.

### 2. EXISTING SYSTEM

#### 2.1 Morphological Analysis for Rule based Machine Translation

[3] The Arabic language is a Semitic language. It exhibits systematic but complex morphological structure based on root-pattern design. English and Arabic languages are highly asymmetrical in terms of morphological structures. This system segments the word in the morphologically rich language into the sequence of prefixes stem suffixes. To make the desired morphological and syntactic symmetry, the system identifies morphemes to be merged or deleted in the morphologically rich language. The technique applied here aims to improve Arabic to English translation quality.

## 2.2 Rule Based Machine Translation System from English to Tamil

[4] The main goal of this system is to develop English to Tamil machine translation system using rule-based approaches. Considering the structural differences between English and Tamil languages, for rule-based approach, syntax transfer based methodology is adopted for translation. This translation system is a parser, which analyzes the English language (source text), and the corresponding target structure is generated through the transfer lexicon. To generate the proper Tamil sentences morphological generator for Tamil language is required.[5] The morphological transducer is responsible for creating correct Tamil word forms and is formed following the theory of lexical phonology which is responsible for the interrelationship between phonological rules and morphological rules in the context of lexical and post-lexical rules (Mohan, 1986) is.

## 2.3 Efficiency of a Machine Translation System

[2] In the field of machine translation large amount of work has been done to improve the essence of translation systems. Warren Weaver proposed the first idea of using computer for translation in 1949. Since then several approaches have been followed out to better the quality of translation systems. But even after continuing development in this field, we haven't reached the level where we could get the exact translation of the source text. The aim of this paper is to brief about different approaches of machine translation and introduce a mechanism for the evaluation of the efficiency of translation systems based on the lexical, syntactical and semantics differences between source text and translated text. This mechanism provides a statistical analysis of translation systems, the results of this analysis can be used as a rating mechanism on the translation systems used worldwide. In this paper, we have emphasized on translation between English and Hindi.

## 2.4 An Adaptive Machine Translator for Multilingual Communication

[1] Machine translation between natural languages is an extremely difficult problem in Natural Language Processing that is still being researched today. This research study explores the effectiveness of developing a translator using Lexical functional grammars. The main research objective is building a machine translation system for multi-lingual communication, which is developing a system where inputs are relating descriptions of a desired source and target language and whose output is a program that translates between the

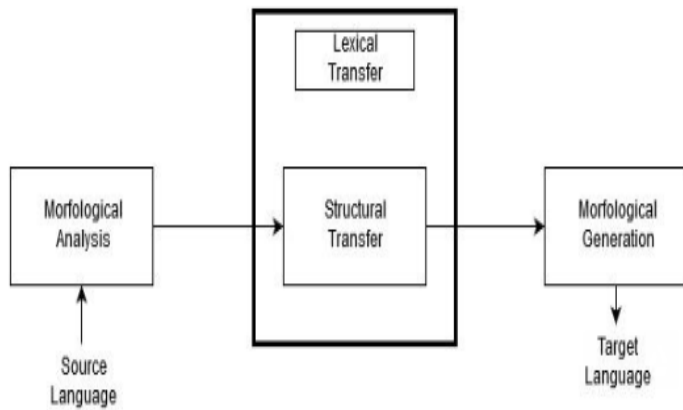
two natural languages. The benefits and drawbacks of this approach as generalized to Machine Translation systems are also discussed, along with possible areas of future work.[6] Data Analysis, Results and Discussion this section of the study focuses on the analysis of collected data. First, an overview of the translated strategies established in the data will be given. Researchers will then go on to describe the results and apply translation strategies in more detail. In many cases different examples of counterparts from the data will be given and analyzed. After that some statistics and figures of the strategies implemented will be given. The potential areas of future work as well as the benefits and drawbacks of this approach generalized to MT systems are also discussed.

## 2.5 Word Based Translation

[7] As the name suggests, the words in the input sentence are individually translated by the word, and these words are arranged in a specific way to ultimately achieve the target sentence. Alignment between words in input and output sentences normally follows some patterns in word-based translation. This approach is the first attempt in a statistical-based MT system that is comparatively simple and efficient. The main disadvantage of this system is the words oversimplified by word translation of sentences, which can reduce the performance of the translation system.

## 3. PROPOSED SYSTEM

Rule based machine translation (RBMT) which is a Classical Approach of Machine translation is a machine translation system based on linguistic information about source language and target language which is retrieved from dictionaries and grammars covering the main semantic, syntactic, and morphological regularities. The main approach of Rule based machine translation systems is based on referring the structure of the given input sentence with the structure of the demanded output sentence, necessarily preserving their unique meaning. Rule based translation systems use a combination of language and grammar rules with dictionaries for common words. To focus on certain industries and disciplines, special dictionaries are created based on dictionary entries, which means that words will be translated like how dictionaries does word by word, usually without much correlation of meaning between the words, followed by syntactic arrangement. In this system, dictionary looks up with the morphological analysis.



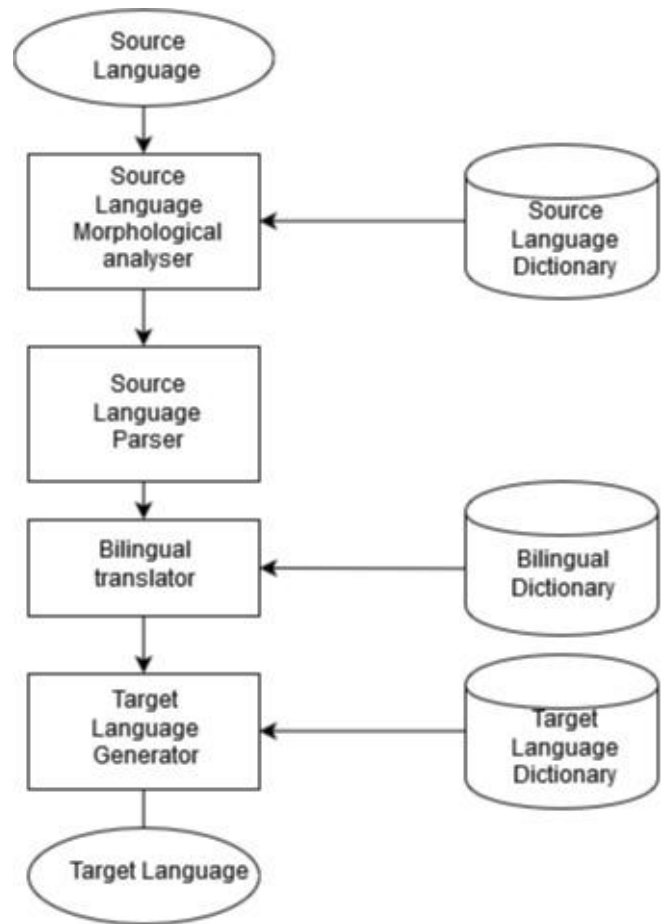
**Figure 1 Architectural diagram for Machine translation**

The architecture for this process is shown in Figure 1. Different phases represent different steps of the translation process. To fix up the punctuation marks, symbols etc., Pre-editing is required which do not require any translation, but post editing is to ensure that the translation is done efficiently. Morphological, syntactic and semantic analysis will be done on the source text. It then checks for any prepositions in the text. If any prepositions exist, then it extracts the word that precedes, and succeeds the preposition, and consider that as a triplet. Triplet will be translated based on the preposition in that and then the whole sentence will be translated and reordered based on the semantic structure of Tulu language. If preposition is not present in the input text, this step of extracting the triplet and then translating won't be necessary. The rule-based approach on the sentences in English (source language) is identified, corresponding rule for the Tulu (target language) will be identified. In the next stage, internal representation of the sentences in Tulu will be generated. The next phase, contextual semantic and syntactic generation will help in finding the exact words in Tulu and frames the words or sentences.

**4. IMPLEMENTATION**

An architecture description is a formal description and representation of a system, organized in a way that supports reasoning and behavior of the system.

**4.1 Architectural design**



**Figure 2 Represents the Architectural Design**

Figure 2 represents the architectural design of the system where source language is fed to Morphological analyzer with the help of language dictionary. The output generated is then passed into Parser. Bilingual translator then looks up in its dictionary. Target language generator translates to target language and displays it.

**4.2 Modular Design**

Modular design is an approach (design theory and practice) that subdivides a system into smaller parts called modules or skids, that can be independently created and then used in different systems. A modular design can be characterized by functional partitioning into discrete scalable, reusable modules, rigorous use of well-defined modular interfaces and making use of industry standards for interfaces.

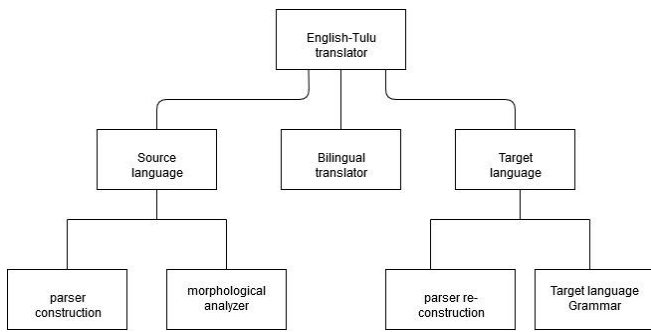


Figure 3 Diagram Representing Modular Design

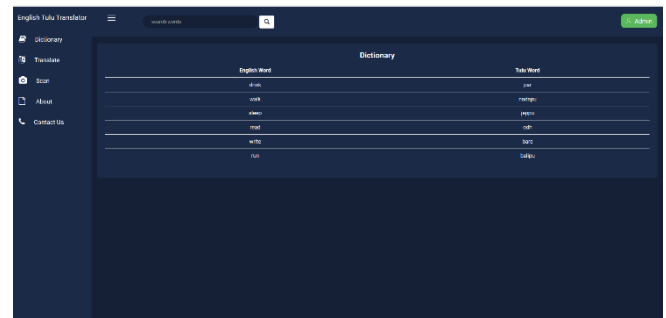


Figure 5. Dictionary Section

### 4.3 Natural Language Processing

Natural language processing (NLP) is a field of artificial intelligence in which computers analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

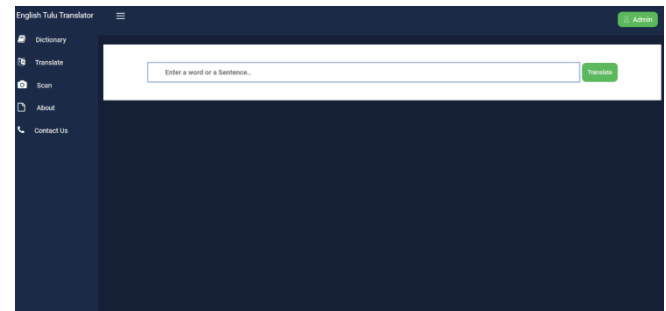


Figure 6. Translate Section

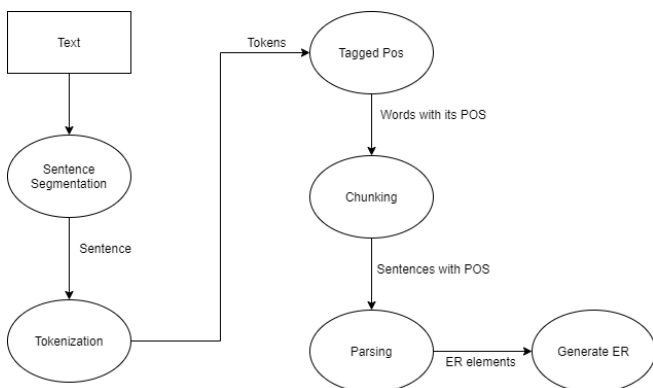


Figure 4 Natural Language Processing Diagram

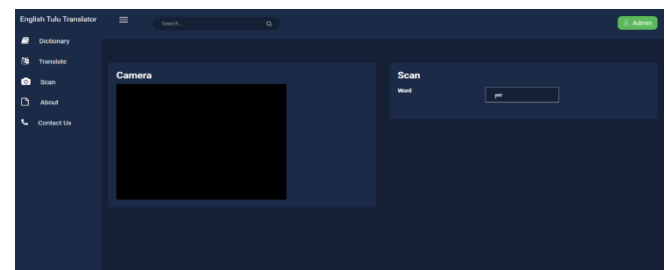


Figure 7. Scan Section

## 5. RESULTS

### 5.1 User page

User page allows the users to view dictionary, translate sentences via entering texts or sentences or by scanning images.

### 5.2 Admin page

Admin page allows admin to add, edit or delete words in dictionary.

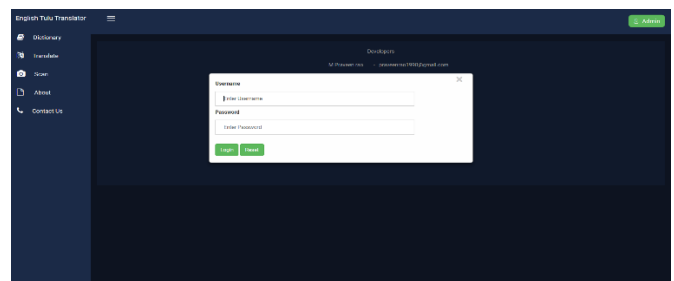
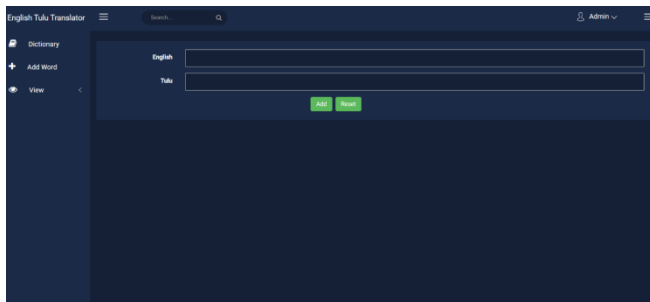
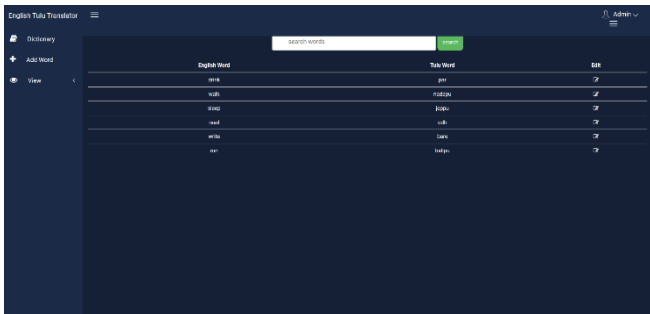
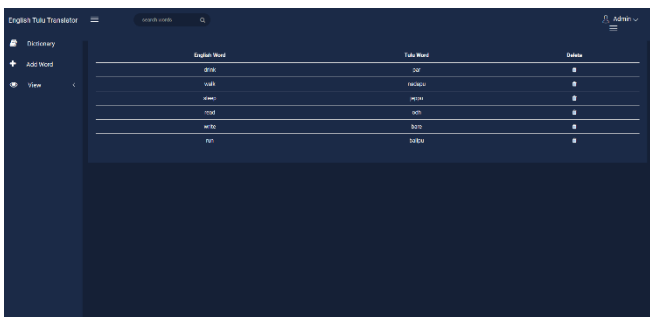


Figure 8. Admin Login Page

**Figure 9. Add Word Section****Figure 10. Edit Section****Figure 11. Delete Section**

## 6. CONCLUSION

In this paper, a new approach for machine translation system from English to Tulu is proposed and this is based on Rule based machine translation (RBMT). The work that is done in English to Tulu translation is complex and among the different approaches that are available, this is unique. We are working on writing effective set of rules and building a dictionary that is rich in words which consists of general words and other words, so that our

translation system can deal with any kind of sentences. Classification of words is another complicated task, which we are working on. We want to extend this work to identify the phrases and idioms, other functional words and translate them in a better manner instead of a direct translation.

## REFERENCES

- [1] (2007). 'Moses: Open source toolkit for statistical machine translation in ACL'.
- [2] Ryan Lane and Ajay Bansal, "An Adaptive Machine Translator for Multilingual Communication", 2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises.
- [3] Sandeep Kharb, Hemant Kumar, Manoj Kumar, Dr. Arun Kumar Chaturvedi, "Efficiency of a Machine Translation System", International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.
- [4] Arwa Hatem, Nazlia Omar, Khalid Shaker, "Morphological Analysis for Rule based Machine Translation", 2011 International Conference on Semantic Technology and Information Retrieval 28-29 June 2011, Putrajaya, Malaysia.
- [5] Renganathan, V. (2002). "An ineractive approach to development of English-tamil machine translation system on the web". The international Tamil Internet 2002 Conference and Exhibition (TI2002).
- [6] SaharAhmadi and Saeed Ketabi. 2011. "Translation Procedures and problems of Color Idiomatic Expressions in English and Persian". The Journal of International Social Research, Volume: 4 Issue: 17.
- [7] Antony P.J, "Machine translation approaches and survey for Indian languages," ACLCP, Vol. 18, No 1, March 2013.