1. Introduction

In the Age of Globalization, the technological development has made the distance and time not an obstacle for interaction among the people belonging to different countries. However, the existence of different languages – multilingual factor - is still an obstacle.

Especially multinational organizations and industries are under pressure to take or convey their messages such as brochures, guides to different countries in the local languages spoken there. For them, multilingual situation should not be an obstacle. Once the message is prepared in a particular language like English or French, in no time it should be translated into different languages. Thus languages also have become one of the important resources for the development of global industrialization and marketing.

Not only for industries, but for the political or administrative organizations like common European Parliament, the instant translation of their messages written in one language should be translated into other multiple languages.

Here, the development of electronic and information technology, especially the Language Technology, has come to help.

In the modern electronic age, many cognitive tasks of the human brain are being carried out by the computer. In addition to the ability of processing the numerical data, computer has been enabled to process the alphanumerical data. It undertakes many of the human
cognition tasks. One among them is the processing of natural language text.

From inputting the language alphabets to the translation task of texts in one natural language into another language, computer is now enabled to undertake most of language tasks. This branch of developing speech and language processing tools is called Language Technology.

Though the computer - the electronic brain - is enabled to undertake the above mentioned cognitive task of speech and language processing, the question is whether it could compete with the human brain - the bio-chemical organ. Many formalisms – both computational and linguistic ones – are being proposed to enable the computer to think and work like human brain. This branch of speech and Language processing is called Computational Linguistics.

The ultimate aim of Computational Linguistics is to achieve success in developing a computational model of speech and natural language processing. It attempts to evolve many formalisms and algorithms to be provided to the computer for the above process. This has led to the development of many language tools from spell checking to human – machine interface through natural language.

Machine Translation is part of the above endeavor. Automatic processing of speech and language texts in one natural language and translating those texts in to another language will be a challenging task. For this, all the levels of linguistic analysis should be automatically done as human brain does. Computational Phonology, Computational Morphology, Computational Syntax and Semantics have been developed for the above language analysis.
1.1. Automatic Machine Translation

As a part of Language Technology, Machine Translation has now achieved a significant success. For this venture, the knowledge of human language should be given to computer. That is, the language processing capacity of the human brain should be provided to the computer.

In developing the MT systems, the comparative and contrastive study of languages from the perspective of language technology is very much needed. Hence the development of Computational Linguistics all over the world.

1.2. Aim

The aim of the present thesis is to contribute to the development of the necessary language modules to be incorporated in the Machine Translation System to translate Tamil texts into English texts.

1.3. Objectives

The Objectives of the proposed thesis are:

1. To analyze the inflectional properties/morphosyntactic properties of the agglutinative language Tamil
2. To develop a Morphological Parser to parse the Tamil inflected words into stem and various affixes
3. To develop a Word-Class Tagger to categorize the Tamil Word-forms (inflected words) into different types
4. To find the English Grammatical Equivalences to the Tamil Inflectional properties
5. To develop the Transfer module to convert/translate the Tamil Inflectional properties into English Grammatical Equivalences.
1.4. Limitations of the thesis

1. This thesis restricts itself with the Modern Written Tamil.
2. Only Tamil inflected wordforms are the object of analysis.

1.5. Sources

The following Modern Tamil grammars form the sources for the interpretation of inflectional properties of Tamil:

1. Structure of Tamil Language (in Tamil)
   (Agesthialingom, S. 2002)
2. Contemporary Tamil Grammar (in Tamil)
   (Paramasivam, K. 2011)
3. A Grammar of Contemporary Literary Tamil
   (Kothandaraman (Porko), P. 1997)
4. Modern Tamil Grammar
   (Kothandaramam (Porko), P. 2006)
5. Fundamental Tamil Grammar (Nuhman, M.A. 2007)

For English, the following grammars form the sources:

1. A Comprehensive Grammar of the English Language
   (Quirk, R. et al 2010)
2. The Cambridge Grammar of the English Language
   (Huddleston & Pullum 2002)
3. A Student’s Introduction to English Grammar
   (Huddleston, R. & Pullum, G.K. 2005)
4. Cambridge Grammar of English
   (Carter, R. & Michael, M 2006)
1.6. Methodology

Normalization, Tokenization, Morphological Parsing, Word-class Tagging are the various tools developed based on Computational Linguistics.

Necessary Flow-charts are designed to understand the morphotactics of Tamil and to parse the wordforms. For Parsing, Left to Right direction is adopted.

The necessary computer program is developed in .NET platform using C# language.

1.7. The present thesis consists of seven chapters

1. Introduction
2. Review of Literature
3. Inflectional Morphology
4. Tamil Inflection
5. English Equivalence for Tamil Inflection
6. A Transfer Module in Tamil – English MT
7. Conclusion

Chapter 1: Introduction

Translation and Machine Translation system are explained. Various approaches and methods adopted in various MT systems are explained. Based on that, aim and objectives of the present thesis are being explained.
The sources for Tamil and English grammars are mentioned. The limitation of the present work is discussed. The structure of the present thesis and the contents of each chapter are explained briefly.

**Chapter 2: Review of Literature**

The works already done in MT all over the world are analyzed under five heads as follows:

1. First Period (1948-1960)
2. Second Period (1960 - 1966)
5. Fifth Period (1990 - Present)

Also some important MT systems already developed around the world are discussed here.

- **SYSTRAN**  
- **SUSY**
- **LOGOS**  
- **METEO**
- **Ariane (GETA)**  
- **Eurotra**
- **METAL**  
- **Rosetta**

The MT attempts for Indian languages are studied. Some of the following notable MT systems are explained here:

- **ANGLABHARTI**  
- **ANUBHARTI**
- **ANUSAARAKA**  
- **MANTRA**
- **MaTra**  
- **SHIVA**
- **SHAKTI**  
- **ANUVAADAK**
The articles and books on Machine Translation systems have been reviewed in this chapter.

Chapter 3: Inflectional Morphology

This chapter provides the foundation for the later chapters. The Inflectional process is explained. Many fundamental concepts – Part-of-speech (POS), Sub-categories, Word (Lexeme), Word-Form (Inflected lexemes), Form-Class, Inflectional categories, Inflectional/Morphosyntactic properties – are discussed here.

The contribution of Inflectional properties to the meaning of the words in a sentence is discussed. Various Inflectional Categories with their properties/values found in human natural languages are explained with suitable examples in this chapter.

Chapter 4: Tamil Inflectional Morphology

In this chapter, Tamil Part-of-Speech classification is studied in detail here. The various criteria used for the classification POS in Tamil is discussed.

Various Inflectional Categories found in modern Written Tamil are studied here. And the respective properties of each Inflectional category are explained with suitable examples.

The various Word-Classes or forms found in modern Written Tamil are studied. The classification of word-forms into different types in Tamil is explained.

Tamil suffixes, morphotactics and morphophonemics used in the development of Tamil morphological parser are studied in detail.
Issues found in the Tamil morphological parsing are discussed. Then, how the different inflected words could be grouped into different word-classes/form-classes is explained.

Based on the above knowledge, three computational linguistic tools – Tokenizer, Morphological Parser and Word-class Tagger – are developed for the analysis of Tamil Inflection.

Then, how the above three linguistic tools could be accommodated in the Tamil – English MT system is discussed in this chapter.

**Chapter 5: English Equivalence for Tamil Inflection**

The fifth chapter analyses the various inflectional categories of English as well as some syntactic categories which are relevant to Tamil Inflectional categories.

Two important linguistic factors are to be considered in this context. One is, Tamil and English belong to two different language families. Tamil belongs to Dravidian whereas English belongs to Indo-European. The second one is, Tamil is a Head-last language whereas English is a Head-first one.

These two linguistic factors make the translation of Tamil inflectional properties into English a difficult and complex task.

The lexical equivalences could be found by the translator with the help of a good bilingual dictionary. But the knowledge of the above mentioned equivalences for grammatical properties could be acquired by the translator only with the comparative study of Tamil and English grammars.
In this chapter, the English equivalences to Tamil inflectional properties are discussed with suitable examples. For some Tamil inflectional properties such as Singular/Plural, Genitive cases, Past Tense/Present tense, equivalences are available in English morphological structure. But, for many other Tamil inflectional properties such as Masculine/Feminine (Gender), various Aspects and Modals, Voice, there is no inflectional/morphological level equivalence in English. They are being expressed by individual function words and syntactic structure in English.

The above issue is an important and difficult one in the development of Tamil – English MT. This is studied in detail in this chapter. This helps us in developing a Transfer module to translate Tamil inflectional properties into English.

**Chapter 6: A Transfer Module in Tamil – English MT**

This chapter deals with the issues found and solutions provided in the development of a Transfer Module to be accommodated in the Tamil – English MT system.

This module helps to translate Tamil inflectional properties into their respective English Equivalences. The English Equivalences may be either morphological inflection or some function words or syntactic structure.

This module consists of the following:

1. Tamil Morphological Parser
2. Tamil Word Class Tagger
3. Rules to transfer Tamil inflectional properties into English
This Transfer Module will help the MT system in transferring/translating Tamil inflectional properties to English equivalences. Thus it will help to translate the Tamil Wordforms – Lexemes plus inflectional suffixes into English equivalences

Chapter 7: Conclusion

From the analysis made in the previous chapters, this chapter summarizes its contribution in the following areas from the perspective of the development of Tamil – English Machine Translation system:

1. The study of inflectional properties/morphosyntactic properties of the Tamil and English
2. The development of a Morphological Parser to parse the Tamil inflected words into stem and various affixes would be developed.
3. The development of a Word-Class Tagger to categorize the Tamil Word-forms would be provided.
4. Transfer Rules to translate Tamil Inflectional properties into English Grammatical Equivalences.

Along with the above contribution, this final chapter explains how this contribution could be used for further Research and Development of Machine Translation of Natural languages in general. To demonstrate the application program, necessary screen shots have been added in this chapter.

Screen shots, Charts and Tables are added wherever necessary. Bibliography is added at the end.