Chapter 1 : INTRODUCTION

In automatic Natural Language Processing (NLP) applications, language data are processed by a computer system that does not really understand human language. The various abilities of these systems among others are, to correct spelling mistakes, part of speech tagging, retrieve relevant documents from large databases from the internet, speech to text conversion and vice-verse, detection of bad grammar in written texts, etc. One of the fundamental and basic components in language processing computer applications is morphological analyzer; and it is an established fact that, in computational linguistics, a morphological analyzer is a starting point for many natural language processing applications (Pretorius & Bosch, 2003; Yona & Wintner, 2005). Morphological analysis deals with the process of separating and identifying existing words into their constituents, called morphemes, and describing how these morphemes are combined to form words. According to its grammatical function within a word structure, a morpheme can be described in many forms such as root, stem, prefix, suffix, particle, infix, circumfix, etc. Thus a morpheme may consist of a word, such as cow, or a meaningful piece of a word, such as the –ed in look**ed**, that can no more be divided into smaller meaningful parts. One of the major ways morphologists investigate the internal structure of words and their formation is through the identification and study of morphemes, the smallest linguistic pieces with a grammatical function (Mark Aronoff, Kirsten Fudeman, 2005). A traditional way of doing morphological analysis has been in manual fashion. Linguistic experts took several months to years in order to describe a single language. Recently, computer algorithms have been applied which can automate and therefore speed-up the process, deploying various techniques from machine learning to improve their performance with increasing numbers of words or analyzed word examples they have access to.

1.1 MORPHOLOGICAL ANALYSIS

The two general problems faced in the analysis of word structures of a language is said to be A) morphotactics/ morphosyntax, the identification of the morpheme order and B) the morphophonemic alternations occurred at the morpheme boundary within the word structure.

A) **Morphotactics/ Morphosyntax:**

The morphological complexity of words of less inflected languages like English are simpler compared to that of agglutinative languages like Manipuri. English words forms are

(1.1) cats→cat + s →cat+PL

(1.2) happiness →happy + ness →happy + ADJ + NZR

i.e.

Lexical Form Surface Form	Cats cat +PL
Lexical Form Surface Form	happy + ADJ + NZR happiness

In contrast to the simple word structured languages like English, the word forms in Manipuri are characterized by the presence of a number of morphemes concatenated one after another like beads on a string. The following Manipuri word demonstrates the agglutinative nature of Manipuri,

(1.3) pusinbihənləmgədəbəni

The meaning of which stands in Manipuri is

(I) could have let the (the other person/some carrier) do the work of bring in (something).

The above word can be broken in to its lexical/grammatical components as

(1.4) pu+sin+bi+hən+ ləm+gə+də+bə+ni

→ _pu+DIR+HON+CAUS+DCT+IRSP+DUBT+NZR+COP_

Lexical Form	pu+DIR+HON+CAUS+DCT+IRSP+DUBT+NZR+COP
Surface Form	pu +sin +bi +hən +ləm +gə +də +bə +ni

These morphemes do not combine in a free order but follows some strict rules. These rules specify the order of arrangement of the morphemes in a word structure; for instance the number suffix when occurs with case markers in a word structure, the former strictly precedes the later. Examples for ordering constraints in the Manipuri nominal paradigm[1] are as shown:

(1.5) keIthel –siη –gI _*keIthel –gI –siη_

market –PL –CASE _*market –CASE –PL_

(1.6) khudəm –siη –bu _*khudəm –bu –siη_

example –PL –CASE _*example –CASE –PL_

B) Morphophonemic Alternation:

Morphemes can be realized in many different forms, e.g. /-gI/-kI/, /-bu/-pu/, /-i/-mi/-pi/-li/-ηi, /-bə/-pə/, /-thok/-dok/-tok/, etc. These are conditioned by the phonological properties of the surrounding sound segment of the affixes.

_(1.7) pu –bə_ _*pu –pə_

take –NZR _*take –NZR_

(1.8) set –pə _*set –bə_

wear –NZR _*wear –NZR_

_(1.9) ca –i_ _*ca –mi_

eat – ASP _*eat –ASP_

(1.10) ca –thok –pə _*ca –dok –pə_

eat –DIR –NZR _*eat –DIR –NZR_

[1] Examples marked as * are considered ungrammatical or non-felicitous

These changes at the morpheme boundaries pose a real challenge for the computational linguists.

Looking at the above phenomena in the word structures, it is crucial to note that in the context of NLP applications, doing morphological analysis of the word structures of a language, especially for agglutinative languages with rich and complex morphological word structures, requires substantial amount of grammatical, linguistic and vocabulary insights of the language in question. So in order to implement the morphological model of a language, understanding linguistic insights of the word structure of a language is an essential task from computational point of view.

Very often, developed and well-described languages with a large amount of speakers get more attention and focus in terms of its resource development. Resource development for less developed languages like Manipuri is never an easy task and may face various challenges. Successful statistical approach and the methods developed for other languages with good resources may not be suitable for this specific language.

1.2 MOTIVATION

Numerous indigenous languages of India especially the north-east Indian languages have received far less attention from computational linguists than the other standard Indian and European languages. Given the non- availability of standardized document of the language and a very complex linguistic features and phenomena observed in Manipuri language, from a morphological point of view it is safe to say that the level of complexity it offers, by large, exceeds what computational linguists deal with in European languages.

Manipuri is an agglutinative language. A word form in the language very often is equivalent to a long English sentence. The word formation processes observed in Manipuri language are also very complex. These two features of the language pose a very special challenging task for morphological analysis and generation using computational approach, but nonetheless may also lead to interesting insights on the possibilities and limitations of the standard methods applied for morphological analysis and generation.

It is to be noted that in India with more than thousand languages, survival of indigenous languages (with less number of speakers) is very tough as almost all of the languages with less number of speakers struggle with strong social pressure from dominant languages (mainly Hindi, English, Bengali, etc. for Manipuri language). A large number of languages have already become extinct, viz. Tripuri in Tripura, Cachari in Cachar, etc. in north-east India. If an attention from the computational field towards Manipuri language may help increase the social prestige of the language, then this should constitute another good reason for my investigation in this direction.

For the successful implementation of any morphological tool, an indispensable requirement is the standard and comprehensive linguistic description of the language in question. Though Manipuri lacks such a standardized grammatical documentation of the language in terms of spellings, etc. it has some extensive documentation on its linguistic features. Besides these, I am being a native speaker of the language the language is a natural choice for my research topic.

1.3 OBJECTIVES

The of main objectives of this study is to develop a language model for **Morphological Analysis of Manipuri** language with the help of **Finite State Techniques.** Implementation of the analyzer with the help of a finite-state tool naturally give rise to various interesting questions. The work is guided by the following four main principle objectives:

> **Linguistic insight**: The in-depth linguistic examination of Manipuri word forms from computational perspective is must for a probable natural language processing application.
>
> **Efficiency & Sufficiency**: Finite-state technique and tool used here for the purpose is efficient enough. Also the technique used is sufficient to handle the

idiosyncrasies and other irregular features occurred in the word structure of Manipuri language.

Coverage: A satisfactorily wide range of Manipuri word forms can be analyzed using this technique.

Effort: The comparison in the efforts of implementation of Manipuri language with other languages using finite state techniques.

1.4 OUTLINE OF THE THESIS

This thesis is organized in the following way:

Chapter 1: The first chapter gives an introduction to the topic of the study, the motivation behind solving the problem and the objectives set for the study.

Chapter 2: The chapter two discusses various approaches available in the literature for morphological analysis with the proposed methodology and design issues for the same.

Chapter 3: The third chapter gives a brief detail about Manipuri language, phonological sound system and its representation using Bengali script and some of the salient linguistic features of the language with a perspective from computational linguistics point of view.

Chapter 4: This chapter is about the theoretical framework of our work that introduces the finite-state theories of automata and transducers along with its mathematical properties and characteristics. Moreover, the current chapter discusses the use of automata and transducers in the context of developing a language model for natural language processing applications, especially for morphological analysis of natural languages. A brief note on Xerox finite-state tool xfst is also included in this chapter.

Chapter 5: Chapter five is all about the morphological structure of Manipuri language which would find an answer to the first research objective- linguistic insight of Manipuri for morphological analysis for NLP applications. A thorough study of the morphological elements, such as affixes, roots and word formation processes using affixation is discussed here. Word classes along with idiosyncratic features unique to some sub categories of word classes are briefed. The morphotactics description of major word classes are discussed for each morpheme. The morphophonemic alternation rules that govern the various spelling change rules and their corresponding regular expressions are specified as per xfst regular expression notations.

Chapter 6: The accounts of the implementation of morphological analysis of Manipuri using xfst tool is detailed here in this chapter.

Chapter 7: Evaluation of the developed model with sample data and analysis of the evaluation outcome is described in chapter seven. This chapter answers the second and third research objectives- efficiency, sufficiency and coverage.

Chapter 8: Chapter eight is the conclusion on the findings of the research work. It elaborates whether the objectives set for the research work has been answered or not with a hint for future directions for research in this field for the language.

References, Abbreviations, Multichar Symbols, Annexes, List of Papers published and Conferences and seminers, forllows chapter 8.