CHAPTER I
INTRODUCTION

In addition to fulfilling the information needs by individuals, dictionaries serve a collective need for recording and documenting language through generations and cultures. MMD serves as editorial lexicography, corpus based lexicography, and WordNet inspiration based lexicography, electronic online-based lexicography, computational lexicography and collaborative lexicography. Recent developments in Natural Language Processing (NLP) have highlighted text processing and language engineering, Machine translation, corpus based language processing, text summarization, Information Retrieval, and Information Extraction. Dictionaries are essential tools for language learners; however, as the technology is changing rapidly, the trend for using electronic dictionaries and online dictionaries among people is also increasing. Multilingual online dictionaries often include a database of cross-referenced unilingual dictionaries with the use of Interlingua such as ontology (Onyshkevich and Nirenburg, 1994) or a pivotal language (Boitet et al., cf.). Information and communication technology has been developing enormously throughout the last decade. MMD (Multilingual Manipuri Dictionary) deals with development of Multilingual dictionary; using English, Hindi and Manipuri Language where English or Hindi will be the source language to translate the target language i.e. Manipuri. The MMD dictionary storage is based upon a trie (M-ary tree) data-structure. The design of the dictionary will be described, and the way in which the data-structure is implemented will also be discussed. The target language will use Meitei Mayek script, which will be the first development in the context of Manipur with full enhancement in NLP using Meitei Mayek. The dictionary storage is based upon a trie (M-ary tree) data structure. The MMD is designed with inspiration features of WordNet, which shows how to build dictionary based on corpus and the way in which the data-structure is implemented will also be discussed.

1. 1 Dictionary

The term Dictionary derives from the Latin word ***dictionarium*** (diction meaning ‘word’). Dictionaries are data structures that support search, insert and delete operations. Dictionary is a simple tool, which helps us to pronounce, respell, and the grammar we need to collaborate within, in a communication. Dictionaries which may be Monolingual, Bilingual or Multilingual -are used for a wide variety of purposes and by people with various backgrounds in the language. From the point of view of time the dictionaries can be either diachronic (dynamic) or synchronic (static), the former dealing with words across time and the latter at a particular point of time. It is difficult to have a published dictionary that suits the dictionary needs of a language learner as well as the information requirements of a more experienced speaker of the language. This has led to the creation of many different dictionary products aimed at fulfilling these different requirements, such as primary school editions, and beginner’s guides.

In Automata theory Dictionary can be defined as a Nondeterministic Finite Automaton (NFA), or nondeterministic finite state machine, that is capable of transition to zero or two or more states for a given start state and input symbol. This distinguishes it from a Deterministic Finite Automaton (DFA), in which all transitions are uniquely determined and in which an input symbol is required for all state transitions (Martin, John, 2010). Like DFAs, NFAs only recognize regular languages. NFAs were introduced in 1959 by Michael O. Rabin and Dana Scott, (Rabin, M. O.; Scott, D., 1959). An NFA, similar to a DFA, consumes a string of input symbols. For each input symbol, it transforms to a new state until all input symbols have been consumed. An *NFA* is represented formally by a 5-tuple, $(Q, \Sigma, \Delta, q_0, F)$, consisting of

a finite set of states Q

a finite set of input symbols Σ

a transition relation $\Delta : Q \times \Sigma \rightarrow P(Q)$.

an *initial* (or *start*) state $q_0 \in Q$

a set of states F distinguished as *accepting* (or *final*) *states* $F \subseteq Q$.

Here, P (Q) denotes the power set of Q. Let $w = a_1a_2 \ldots a_n$ be a word over the alphabet Σ. The automaton M accepts the word w if a sequence of states, $r_0, r_1, \ldots, r_n$, exists in Q with the following conditions:

1. $r_0 = q0$

2. $r_{i+1} \in \Delta(r_i, a_{i+1})$, *for* $i = 0, \ldots, n-1$

3. $r_n \in F$.

A Dictionary can be categorized as five types, as the figure below showns the types and sub types of Dictionaries.

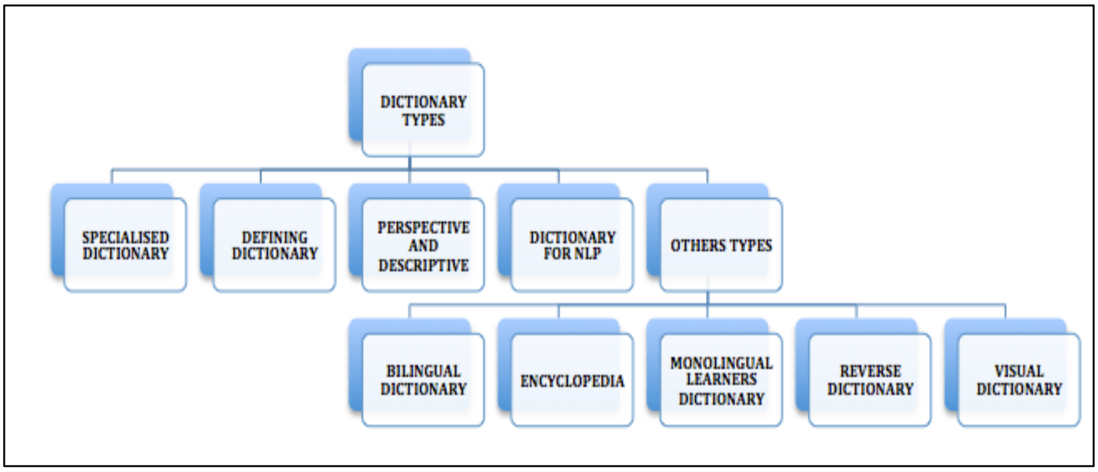

Figure 1: Types of Dictionary

1. 2 Multilingual Dictionary

Multilingual dictionary means dealing with multiple languages. Multilingual dictionary is closely related to Bilingual Dictionaries and Trilingual Dictionaries. In a multilingual dictionary, we look up a word or phrase in one language and are presented with the translation in several languages. The MMD dictionary structure is inspired by the WordNet basic design and is corpus based. The dictionary's source knowledge base and the software system provide interfaces to producing an on-line dictionary, a printed-paper dictionary, and several electronic resources useful in Natural Language Processing. Multilingual dictionaries can be arranged alphabetically or words can be grouped by topic and domainwise. When grouped by topic it is common for multilingual dictionaries to be illustrated. It is a specialized dictionary, which is used to translate words or phrase from one language to another. It also indicates the microstructure; POS, gender, verbs type, gloss, concept, source, declension model and other grammatical clues to help a non-native speaker use the word. A Multilingual Dictionary may have as its function to help user translate text from one language into another or its function may be to help understand foreign-language text. The main function of Multilingual Manipuri Dictionary (MMD) is to present the history of a lexical item of Manipuri Text using Meitei Mayek as Manipuri script. The features of MMD dictionary record the development of a lexical item in terms of both the form and the meaning of the particular lexical unit, the origin of words by tracing the present day words to their oldest forms.

The relationship between NLP and MT has always been close. Perhaps too close, and much of NLP research and the evolution of NLP techniques have been tailored to the MT application. The relationship between NLP and information management based on content has not been quite as symbiotic compared to NLP and MT. The kind of NLP that has been developed for application like MT, until recently, has had little interface on information retrieval.

The dictionaries - Monolingual, Bilingual or Multilingual are the standard ways of collecting and presenting lexicographic knowledge about one or more languages. The electronic dictionaries are not merely a straightforward extension of their printed counterparts, but they entail additional purely computational problems. Current electronic dictionaries do nothing more than simply search and retrieve dictionary data, presenting information in a plain format similar to the paper version it was adapted from. While these systems may save the user the time in turning pages, they lose the functionality of paper dictionaries in allowing users to browse through the other words of the language, or see the entries nearby.

Past research into Multilingual dictionaries (not for Manipuri Language) seems to fall into two very distinct categories of - dictionary databases or Machine Readable Dictionaries and the use of dictionaries for language learners. As identified in (Kegl, J, 1995), it is surprising that despite the enormous potential of having dictionaries in electronic databases there has been almost nothing effort towards combining these two areas of research and using advances in electronic dictionaries and language education to benefit speakers of the language.

1.3 Rationale of the Study

With the advent of computers and the Internet, a new possibility to enhance vocabulary learning is brought into the field of language learning. Acquiring or learning new words while reading a text is an important practical method of vocabulary enlargement. The availability of authentic materials on the Internet and the access of online dictionaries provide two helpful conditions for learners to learn vocabulary. Firstly, the World Wide Web is not only one of the most efficient channels for global communication but also a huge and abundant language-learning source for users or learners also. Secondly, online dictionaries appear with computer technology combined with dictionary information. Integrated with computer technology, NLP, dictionaries assume more importance for text comprehension and vocabulary learning with its technical benefits. Online dictionaries are no longer as troublesome as paper dictionaries are with the characteristics of being able to show the explanations of a new word promptly. They overcome the disadvantages of a paper dictionary in the sense of saving the time used for searching for the word in a thick dictionary, which has several hundred pages or more. Now, in its new form, an online dictionary has become an important instrument for learning a language, especially for learning vocabulary. The searching process for a word in an electronic dictionary is greatly shortened by the advantage of computer speed. Apart from time consuming, looking up for a word in a paper dictionary is a process of switches, first switching from a reading material to dictionary and then from the dictionary to the reading material. It is a disruptive process. Now with the help of e-dictionaries, learners' thought flow is no longer disrupted as much as before, especially with the function of instantly obtaining the explanation when putting the cursor on a word. With the merits of saving time and not disrupting the thought flow as much as paper dictionaries do, the e-dictionaries make it possible for learners to read more fluently; therefore, they increase the learners' chance of acquiring the looked up words while reading. Leffa (1992) compared the efficiency of an e-dictionary and a conventional dictionary in a translation task and found that the computer dictionary enabled the students to "understand 38% more of the passage, using 50% less time". Many studies (Hulstijn, 1993; Knight, 1994; Chun & Plass, 1996, Chun & Plass, 1997; Hulstijn, Hollander & Greidanus, 1996; Hulstijn & Trompetter, 1998; Laufer & Hadar, 1997; Laufer & Hill, 2000; Chun and Payne, 2004, Peter, 2007; Peters, Hulsijn, Seru & Lurjeharms, 2009, etc) show that looking up an e-dictionary (containing computerized glosses) has a positive effect on word learning. This provides evidence to the value of e-dictionary use for vocabulary learning while reading a text, especially an authentic one on the Internet or on a computer. However, using an e-dictionary alone may accompany shallow processing of word information since the flow of reading is not disrupted much (Laufer & Hill, 2000). When the e-dictionary was used alone to help learners read a text, it was found that the retention of new words was not as high as when the e-dictionary was combined with one enhancement technique or two or three enhancement techniques (Hulstijn, 1993; Laufer and Hill, 2000; Peters, 2007; Peters et al., 2009). Word relevance is the most frequent task researchers used to make learners focus on new words to be learnt. In Hulstijn's (1993) study, "relevance of words to reading comprehension questions is found to increase the chance of dictionary consultation". Laufer and Hill (2000) point out the indispensability of a word relevance task for studies on e-dictionary, i.e., "The task which cannot be carried out without the knowledge of the words targeted for investigation". The RC task with the factor of word relevance (called task-induced word relevance by Laufer and Hill) makes learners pay attention to the relevant new words and look up for relevant words in order to answer the questions. In fact, the effects of word relevance tasks are more than making the learner consult a dictionary, what is more significant is "retention was very high on the immediate vocabulary tests" (Peters, 2007). The function of the vocabulary task is similar to that of word relevance. Both of them can make learners look up more and process the target words in an elaborate way. Therefore, the merits of word relevance and the vocabulary task lie in creating chances to make learners "do with words". The vocabulary task is more word-directed than word relevance.

This study is to design a well-maintained multilingual dictionary with combination of e-dictionary and hard copy dictionaries of Manipuri. The model may be useful as CALL (Computer Assisted Language Learning).

1.4 Levels of structure

A dictionary can be structured at three levels and these levels can be classified as below:

1. The Dictionary as a whole has framing structure, which comprises a set of main sections.
2. A subset of these sections comprises mostly an entry list. The structure of each of these sections is a macrostructure (or the macro structure of the dictionary).

3. Each of the entries in a macrostructure has an internal structure of its own, which is the microstructure of the entries of that list (thus, if the dictionary contains only one word list, that is the microstructure of the dictionary).

1.5 The challenges in Dictionary Development

Humans are able to deal with incorrect spelling when reading text so spelling correction would be a major challenge while developing a dictionary. One of the major problems of the development of dictionary in computational dictionary and general dictionary is that we have to add all the micro and macro structures of the dictionaries to fulfill the users. In detailing the use of a lexicographical workstation for the creation of dictionaries, Weiner (Weiner, 1994) discusses the initial purpose of the Oxford English Dictionary (OED) and the eventual diversion from their goal:

"*…To create a record of vocabulary so that English literature could be understood by all. But English scholarship grew up and lexicography grew with it…inevitably parting company with the man in the street*".

For a dictionary, however Monolingual or Bilingual or Multilingual it may be accurate record of a living language, the process of revising and restructuring the lexical database never ends. The computerization of this process has made this very tedious task simpler and more efficient for lexicographers. An interesting point made by Weiner is that there is a rich network of information available during the composition of dictionaries that cannot be expressed "entry by entry, in alphabetical order". While the network is implied in the paper editions by cross-references made to related entries on other pages, an 'e-dictionary' can allow these interrelated textual categories to be more easily understood and accessible.

How well a Multilingual dictionary application satisfies a diverse range of users depends on how well the developers use the potential of working in an electronic medium. Despite some significant research in the area of computerizing dictionaries, there has been little effort in addressing the challenge of taking this to real speakers and language learners. This has been a focus of this thesis, to be satisfied not with simply a better-structured dictionary database, but also to address the issues of whether this allows the dictionary to be any more usable by real users. In utilizing the potential of an electronic online dictionary we may come closer to the initial aims of the MMD by customizing the overall experience of vocabulary to the user, so that learning can take place for all.

1.6 Why Multilingual Dictionary for Manipuri language

The importance of a good multilingual dictionary as an aid to overcome the gravitational pull of the mother tongue cannot be exaggerated. In fact, the lexicographical art in Manipur is in its primal stages. Manipuri dictionaries, which can match the high standards of multilingual dictionaries available in developed languages, are still a long way off. English-Manipuri dictionaries have mostly been compiled and edited by non-English speaking people as well as un-trained lexicographers. Unfortunately, trained compilers are of very limited in numbers in Manipur. Therefore, the qualitative values of multilingual dictionary are very poor. The problems in making a multilingual dictionary broadly comprise

a) Phonological,

b) Grammatical, and

c) Semantic.

In all existing Manipuri English Dictionaries, the pronunciation of the English entry words are transcribed in Manipuri with Bengali script, which creates many problems to the user (Manipuri speaker) as Bengali has a distinct character of its own. As a matter of fact, Manipuri has a traditional writing system with indigenous alphabets called 'Meitei/Meitei Mayek's (Manipuri script). With the advent of Hinduism (early 18th century) the Bengali alphabet has been adopted in lieu of the traditional one in education etc. According to different scholars the number of script for Manipuri is varied. Grierson has mentioned 35 alphabets in his well-known book 'Linguistic Survey of India' (Vol. III Part III) while some Meitei scholars are of the opinion that it is 27 in number (Grierson, 1973). However, in 1980 the Government of Manipur has recognized 27 for the purpose. Despite the traditional/indigenous writing system the Bengali alphabet is being continued for all practical purposes. In 1992, the Government of India has also included the Manipuri language in the 8th Schedule of Indian Constitution as 'Manipuri written in Bengali script'. It can be mentioned that Meitei script for written Manipuri (Meiteiron) is in the implementation stage. The Government of Manipur have replaced the Bengali script with Meitei script from Class

I to IX and also assured that Bengali script would be completely replaced in a phased manner (Singh, L. Sarbajit, 2002). Information about the pronunciation becomes a primary need for a dictionary as the dictionary making has been increased by new attitude towards speech. "*The printed word is no longer the only means of mass communication; the spoken word has become as important in the age of radio, telephone, phonograph, television, tape recorder, videotape recorder, cinema, and Telstar*." (Al-Kasimi, 1983). Every lexicographer has provided the pronunciation of headword at least and the transcriptions used in the dictionaries are based on purpose of the dictionary especially in bilingual dictionary. However, the application of transcription is the most important task for dictionary makers. As pronunciation is to be included in a bilingual dictionary, it is necessary to choose an appropriate type of transcription. Again Piotrowski pointed out: "when selecting the appropriate transcription, lexicographers must consider their position between two attitudes usage of a transcription widely known on an international scale, and keeping to the traditional one used in their country. International transcription makes it easier for the user to use other dictionaries (monolingual one, for example), while following traditional is reassuring for the user and relates the dictionary to other books published in his or her country" (Piotrowski, 1987). Berkov puts a case clearly: lexicographers in their choice of the system of transcription are restricted only by the convenience of the user (ibid).

It can be said that the people around the globe recognizes Manipur from the benefit of our rich culture. So, it has been the main aim to spread our cultural ideas around the globe. Again, as we are in the 21st century, people have been living surrounded by the Computer media networks and the Internet, which has become the fastest information linking media for accessing information. In developed Countries the study of Dictionary system of natural language has been taken up very early. Such system had been very useful among different races of people in understanding the Language of one another and also in sharing knowledge among different people.

1.7 Motivation

- The idea of Manipuri culture or any other knowledge can be communicated to other speaking different languages by means of translation with a computer.
- The cultural knowledge or any other development knowledge of different

countries having different languages will be better understood by the people of Manipur when they are translated into Manipuri language.

- Development in linguistic science.
- Development of NLP in Manipuri language.
- Linguistic affinity.
- Multilingual search portals can be established to produce the same results, no matter which language is used for retrieval.
- Providing with the knowledge base required for true semantic analysis.
- A high degree of objectivity based on large volumes of text.
- Fundamental content that is highly generalized across different languages and fields.
- Service in language grid.
- Document specific representations no longer affect the search. This is extremely important in the case of multilingual representations.
- The management of large amounts of information and knowledge is of ever increasing importance in today's large organizations.
- Today's search tools perform rather poorly in the sense that information access is mostly based on key- word searching or even mere browsing of topic areas.
- This unfocused approach often leads to undesired results. The problem becomes even worse, if the result searched for only appears in a foreign language document.

With the ongoing ease of supplying information online, especially in corporate intranets and knowledge bases, finding the right information becomes an increasingly difficult task. Today's search tools perform rather poorly in the sense that information access is mostly based on keyword searching or even mere browsing of topic areas. This unfocused approach often leads to undesired results. The problem becomes even worse, if the result searched for only appears in a foreign language document. However, since the document has been annotated with the ontological semantics, this will not affect the search results. Secondly, since the ontology used for annotating the document is domain specific, the semantic meanings and interpretations of keywords are bound to that domain and therefore the retrieval is likely to be more efficient. A term can have several meanings in different domains. Thirdly, document specific representations no longer affect the search. This is extremely important in the case of multilingual representations. Keywords of several languages are mapped to the same concept in ontology and are therefore given the same meaning. Multilingual search portals can be established to produce the same results, no matter which language is used for retrieval. Presently, subject specialists in a time consuming process carry this out. With today's vast amount of available information on the WWW, automatic support is needed to efficiently manage this task. Oncologist plays a critical role in supporting the machine readable semantics needed to facilitate automation. Before such powerful Semantic Web applications can be built and used within certain domains of knowledge, the basic requirement, a machine readable vocabulary represented by domain ontology has to be established. The creation of oncologist is a time consuming task and often carried out in an adhoc manner. Only a few methodologies exist, and even less automated tool support is available. Constituting the knowledge base for future Semantic Web applications, domain oncologists have to be created continuously in all possible areas and communities. The need for a reusable methodology is evident.

1.8 Objectives

1. To explore a rich classical language in term of its technical terms involved and developed by an unknown ancient civilization to the academic world.
2. To render great help for interpretative purpose in modern electronic media or inter language communication in the present context.
3. Intelligent Information Retrieval: To identify necessary information through inference according to given information.
4. To ensure it is a valuable dictionary of Manipuri language technically both for the young learners and scholars.
5. To encourage the development of freely accessible multilingual lexical resources by way of online collaboration of work on the Internet.
6. To reduce the cost factor.
7. To provide Web platform to gather on Internet community around lexical services.
8. To improve Dictionary quality.
9. To promote the contribution of Information and Communication Technology for the development of MMD.

10. To promote the advancement of Information and Communication Technology in MMD.
11. To promote the study and practice of information and Communication Technology in the countries educational institutes, Information processing centers and other IT development center.
12. To promote research in Information and Communication Technology and to assist in the dissemination of the findings of such research in Multilingual dictionaries.
13. To provide forum for the discussion of ideas and issues related with profession.
14. To benefits its member with professional privileges and recognition.
15. To promote the computer literacy throughout the country.

1.9 Data and Methodology

The data for the present study will be collected in the form of written materials like:

- Newspaper
- Journals.
- Novels
- Short-stories
- Dramas
- Text-books
- Word-books
- Library works
- The existing bilingual dictionaries.
- The corpus (mono corpora and parallel corpora)
- WordNet (Manipuri WordNet, IndoWordNet)
- Loan words
- Word transliteration
- Coined words
- Online submission from users (verification and validation by Administrator)
- Participants

- Writing task

1.10 Main Contribution

The main contributions of this thesis will be improvements in Natural Language Processing (NLP) Systems, Knowledge structure, IE (Information Extraction), IR (Information Retrieval). It will promote the contribution of Information and Communications Technology for the development of MMD. The Main approach to multilingual dictionaries is that it could be practically used by learners of the Manipuri language and main contribution is for computer literacy throughout the country, and for use in educational institutes, information processing cultures and other IT development centers. Multilingual Dictionaries development, which may be online, or offline with the concept of WordNet and its features, which are not available in any other Dictionary will be one major contribution. Lastly corpus analysis for many purposes like word frequency word count finding unique word using Perl program for converting TTF code to WX notation and vice versa, developing Dictionaries with Corpora of mono and parallel work. A major computational challenge is how to design the Multilingual dictionary structure in order to make its maintenance manageable and efficient.

1.11 Thesis outlines

Chapter II. Reviews the past work in the field of Monolingual, Bilingual and Multilingual dictionary mainly on target language i.e. Manipuri language and its uses in teaching and for other purposes. The chapter will discuss about the existing English Dictionary, Hindi Dictionary and Multilingual Dictionary features with Manipuri Dictionary available in hard copies. So far there is no such online Manipuri Dictionary available, and the script used in this MMD i.e. English, Hindi and Manipuri script, and about its history and development. It will also discuss about the language affinity, features and characteristics of the language plus the issues of the Manipuri language. And lastly, it will analyze the modern work of dictionary development and type of dictionaries.

Chapter III. Begins with the topic discussing about WordNet i.e. Princeton WordNet, IndoWordNet and its features, advantages and limitations. Further, it will discuss about advantages of multilingual dictionaries and application areas of multilingual dictionaries with the number of words or the data collection from WordNet. Lastly, it will discuss the main differences between WordNet and Multilingual Manipuri Dictionaries (MMD) regarding advantages and limitations of both

Chapter IV. Discusses the corpus and the methodology of corpus development with the methods of data input; later it will discuss about the various principles corpus development and basic principles to be used in MMD corpus. It will also discuss the types of corpus, corpus cleaning or corpus sanitation process in linguistic way, computational process both and the basic application areas of corpus. It will also discuss the types of corpus, corpus cleaning or corpus sanitation process in a linguistic way, computational process and the basic application areas of corpus. Corpus processing with the help of Sketch engine with CQL (Corpus Query Language) will be highlighted with many features of text processing. The chapter will highlight the conversion of raw data of corpus to WX notation and vice versa, POS tagging of Corpus for Domain wise with BIS tagset guidelines.

Chapter V. This chapter will focus on the design and searching of MMD. Further, it will discuss about the target language in addition to trie structure generation, trie search and GI dictionary structure. This chapter will also discuss the technology used to develop MMD, the languages used in the front end, the database used in backend and the server types used by administrator. Here I will discuss the outcomes or the results of MMD, i.e. the pages or the form, where the administrator works and the pages where the general user browses the pages.

Chapter VI. This chapter will discuss and summarize the areas of future works in this field, which will unfold many exciting yet untapped areas in Multilingual Dictionaries and future prospects of Multilingual Dictionaries

1.12 Conclusion

The chapter discusses about the definition of Dictionary in general and Automata theory concept, Multilingual Dictionary, types of Dictionaries, needs of Multilingual dictionaries. In addition, the chapter also focuses on the motivations, objectives, data and methodology of Multilingual Manipuri Dictionary (MMD). Lastly, it also discusses about the main contribution thesis, and its brief summary.