Morphological Analysis for Manipuri Nominal Category Words with Finite State Techniques

Ksh. Krishna B. Singha
Research Scholar,
Department of Computer Science
Assam University, India

Bipul Syam Purkyastha, PhD.
Professor,
Department of Computer Science
Assam University, India

ABSTRACT
The paper presents the design and the implementation of a morphological analyzer for Manipuri nominal category words. A method for the analysis of nominal category Manipuri words with a suffix stripping approach in a right to left direction without using any lexicon has been proposed. Manipuri being an agglutinative language and its rule-based nature while morpheme concatenation allows the morphotactics of the different available word forms to be modeled with finite state machines (FSMs). Also the very feature of the word classes which possess the characteristics that they can only be attached with affixes meant for that class only make it possible to analyze a nominal word without a lexicon. This paper discusses the morphological features of Manipuri nominal category, identifying the affixes for this class, and the steps of the new methodology to develop the FSM for nominal category to represent the morphotactics of the language, converting the FSM from non-deterministic finite automata (NFA) to deterministic finite automata (DFA) and thereby cooperating the analysis.

General Terms
Artificial Intelligence, Natural Language Processing, Morphological Analysis, Finite State Transducers.

Keywords
Agglutinative, Morphological Analysis, Finite State Transducers, Manipuri and Nominal Category, Computational Linguistics.

1. INTRODUCTION
In a resource poor language like Manipuri there is hardly any language processing tool. Morphological analysis in Natural Language Processing (NLP) plays a vital role in any application that deals with human language for formal writing and reading. Morphology is the study of form or forms. In linguistics morphology refers to the branch of linguistics that deals with words, their internal structure, and how they are formed. So in short, morphology is a domain of linguistics that studies the internal structure of words. The morphological analysis of a word is the investigation through the identification and study of morphemes, often defined as the smallest linguistic pieces with a grammatical function. It may worth mentioning that some morphemes have no concrete form or no continuous form and some do not have meanings in the conventional sense of the term. It is traditional to distinguish between surface forms of words and their analyses, called lemma. And so computationally the morphological analysis of a word constitutes taking a word form as input and producing the structure of the word by showing the lexical category of the constituent morphemes. Finite-state approaches to morphology, including the readily available implementation known as Two-Level Morphology [11] and [12], have been shown to work in significant projects for French, English, Spanish, Portuguese, Italian, Finnish, Turkish and a wide variety of other natural languages [3]. The concept of a lexical transducer was first discussed in [1]. A lexical transducer is a particular case of a finite state automaton where an inflected surface form is mapped to its lexical forms. The lexical form consists of a canonical representation of the word and a sequence of tags that show the morphological characteristics of the form in question and its syntactic category [4].

E.g. the information that the adjective əfəbə (good) of Manipuri might be represented in the Manipuri lexical transducer by the path in figure 1, where the zeros represent epsilon symbols.

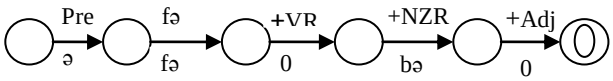

Figure 1. A Path in a Transducer

A set of ordered pairs of strings, the RELATION, can be represented as a finite-state transducer. The arcs of the transducer are labeled by symbol pairs. Each path of the transducer represents a pair of strings in the relation. The basic claim of the finite-state approach to morphology is that the relation between the surface forms of a language and their corresponding lemmas can be described as a regular relation function [2].

As described in [4], constructing lexical transducer consists of (1) a finite-state source lexicon that defines the set of valid lexical forms of the language (possibly infinite), and (2) a set of finite state rules that assign the proper surface realization to all lexical forms and morphological categories of the language. In Manipuri the prefix/ suffix concatenation can result in relatively long words, which are frequently equivalent to a whole sentence in English. Here the concatenation of morphemes resulting in different word forms follows strict rules, specified as morphotactics of the language. In addition to this Manipuri suffixes have allomorphs for almost all the suffixes, which should be taken care of for sound harmony for the surface realization of the word form. One of the important morphological feature of the language is that the number of prefix/(es) that can be attached to a root or stem is at the most only one. And more importantly, the class of a stem can be recognized by the type of affix it is attached to, the analysis of the morpheme structure can be done by analyzing the suffix type and

verifying the morphotactics of the language without using a lexicon. Take an example, say
e.g. cət/pə/gi/ni

Here the copula/ni suffix follows the genitive suffix gi and as the morphotactics rule says that this case marker only follows a nominal category, the remaining part of the word is a stem/ root of noun class. Again since the stem has an ending pə (a nominalizer), further analysis reveals that the root cət is a verb root, as a nominalizer only follows a verb root. So the ultimate analysis will look like

cət + VR + NZR + gen + cop

So without knowing the meaning of the stem, anyone who knows Manipuri can easily analyze a Manipuri word. These features of the language which exhibit the agglutinative nature and the morphotactic constraints followed while concatenating morphemes allows Manipuri to be modeled with FSMs easily.

The second section of the paper discusses the morphology of noun class of the language, the subsequent sections prepares a non deterministic finite state automaton network for the nominal category of the language, converting it to deterministic finite state machine to handle morphotactics of the language to morphologically analyze words covering the nominal category of the language and concludes with concluding remarks with discussion and future work.

2. THE MANIPURI NOMINAL CATEGORY AND MORPHOLOGY

Manipuri is an agglutinative language that belongs to the Tibeto-Burman language family of the Kuki-Chin group. Concatenative morphology, also called 'concatenative morphotactics' [5], is used very productively in Manipuri. Different word classes are formed by either prefixing and/or suffixing morphemes to the root/stem. A Manipuri word consists of a root (base form) and a number of suffixes and or prefixes attached to it, each extending its meaning or changing its word class. The following example demonstrates the agglutinative nature of the language where a number of affixes are attached to the root or the stem to form different word forms.

lən — treasure
mə/rən — (his) treasure
mə/rən/gi — for (his) treasure
mə/rəl/siŋ/gi — for (his) treasures
mə/rəl/siŋ/gi/dəmək — for the sake of (his) treasures
mə/rəl/siŋ/gi/dəmək/tə — only for the sake of (his) treasures
mə/rəl/siŋ/gi/dəmək/tə/su — only for the sake of (his) treasures also

Only the noun category has free roots (exceptions being the kinship terms and body parts of animals) while the verb roots are bound roots. Manipuri nominal category can be broadly classified into two classes- i) the noun class (includes common nouns, body parts, kinship terms, wh-words, etc.) ii) and the one that is derived from the verb roots by adding a nominalizer suffix bə~ pə to the verb root as in **cətpə**. The suffixes taken by a stem of nominal category is the same irrespective of whether it is derived from a verb root or it is a stem of a nominal bound root or a free noun class.

Example:

1) Free form: lai/gi/dəmək/tə (for the sake of God only)
2) Bound form: nəbung/gi/dəmək/tə (for the sake of your brother only)
3) Derived from verb root: cətpə/gi/dəmək/tə (for the sake of going only)

Here lai (God) is free noun root, whereas both bung (kinship term for elder brother) of nəbung and cət (verb root for go) of cətpə are all bound roots, but once they are qualified as nominal class by prefixing the pronominal marker nə (see in table 1) and suffixing the nominalizer pə to the verb root, they take suffixes meant for nominal categories only.

Prefixes can be divided into two categories- Pronominal and Non-pronominal. The following table shows the three pronominal markers along with examples:

Table 1. Pronominal markers with example

Prefix	Meaning	Example	Example Meaning
i	First person	i-ma	My mother
nə	Second person	nə-ce	Your sister
mə	Third person	mə-bay	His brother-in-law

The kinship terms do not give a complete meaning without the pronominal markers. In case of animal body parts the third person pronominal marker mə is prefixed, as in mə-ci → horn (its); mə-səm → hair (its). There are two types of non-pronominal prefixes- nominalizing and formative. The former type has two prefixes- khu and mə, while the later one has seven prefixes- ə, i, mə, tə, thə and sə. Some of these prefixes lose its meaning without context while others are meaningful when used at sentence level and/or reduplicated. Only the verb roots take these non-pronominal prefixes.

Manipuri common nouns are mostly free standing. The third person pronominal prefix mə and certain formative prefixes are found to be attached to this category occasionally. Pronouns occur either in free or bound forms; free when they are used alone while bound occurs when they are attached to some element like plural marker, kinship terms, body part terms, as shown in the following table 2.

Table 2. Free and bound forms of Pronouns

Person	First Person	Second Person	Third Person
Free form	əy	nəŋ	mə
Bound form	i	nə	mə
Extended form	əy/hak	nə/hak	mə/hak
Plural form	əy/khoy	nə/khoy	mə/khoy
Attaching Locative(də)	əy/ŋondə	nə/ŋondə	mə/ŋondə
Attaching Genitive(gi)	əy/gi	nəŋ/gi	ma/gi

It can be observed from the above table that the first person pronoun take the free form irrespective of the context. In case of kinship terms only the bound form is used.

The wh-words of Manipuri expresses a person's lack of knowledge about a particular element and is denoted by using the element kə resembling the English wh occurring in who, what, where, when, etc. This set of words has their own function and has a set of suffixes that follows the wh-element, the most prominent one being the suffix "no" such as in kəna/no (who is it, etc.). The following table shows the derived forms of these words.

Table 3. Some wh-words with derived forms

Wh-words (derived forms)	English meaning
kəna	who
kəri	what
kərəm/bə	which (one)
kərəm/nə	how
kəday	where
kədom/da	where to (direction)

The wh-elements take no prefixes though; it takes all the suffixes of the nominal category with an optional suffix "no" at the rightmost position. Say, e,g. kədom/də/gi/no.

While the nominal category takes a single prefix, the number of suffixes it can take is many. The following table 4 shows the suffixes it can be concatenated with strict morpheme ordering rules.

Table 4. List of suffixes for nominal category

Name of Suffix	Purpose
(nə)- Nominative/ Instrumental	**Case markers**
(pu)/(bu)-Accusative	
(gi)/(ki)-Genitive	
(də)/(tə) - Locative	
(dəgi)/(təgi) - Ablative	
(gə)/(kə)-Associative	
(siŋ)/(khoy)/(yam)	Number
(si)	Proximate (Demonstrative particle)
(du)/(tu)	Remote (Demonstrative particle)
(də)/(tə)	Emphatic/ Particle
(ti)/(di)	Only
(su)	Also
(dəmək)/(təmək)	Only
(ni)	Copula
(mək)	Personification
(ra) /(la)	Interrogative

3. CHALLENGES IN MANIPURI NOUN MORPHOLOGY

Being an agglutinative language, the number of suffixes that can be attached to the root of a nominal category of Manipuri word is high and thereby the complexity of the process increases. The constraints on the morpheme order, the morphophonemic changes at the morpheme boundary, homophonous morphemes, etc. makes the task of identifying the morphemes difficult. Moreover, all the subcategory of the Manipuri nominal category cannot be affixed with the suffixes meant for it at once. Say for example, the bound forms of the root of body parts cannot be attached with the suffixes of the nominal category. First of all they will have to be prefixed with the third person pronominal marker. See the example:

e.g. mə/ci/gi → mə: pronominal marker

→ ci: horn (bound root)

→ gi: genitive (case suffix)

is ok, whereas

* ci/gi is not an accepted form.

Similarly in case of personal pronouns, the suffixes are attached to the free form like əy/gi (and not i/gi), nəŋ/gi and ma/gi, etc. Here again an exceptional case has been observed that the locative də is used as əy/ŋon/də, nə/ŋon/də and mə/ŋon/də when it immediately follows the root. Also the peculiarities of the wh-elements in Manipuri pose modeling the morphology of Manipuri nominal category a very challenging task. The affixation process can only be initiated once they gain a particular form in each case of the sub categories.

4. FINITE STATE TRANSDUCERS AND MANIPURI MORPHOLOGY

A finite state transducer is a finite state network, represented with the help of states and directed arcs which connects the states. Each arc is labeled with a symbol pair, representing input-output string pair. *Path* is a sequence of transition(s) over arcs to a particular state. Transitions between states are possible only if the required input matches with the label at the corresponding arc. The input string is said to be recognized when a mapping is possible from the input to the corresponding output string for the arc in the network. So the finite state transducer provides a set of outputs from an accepted input and thereby expressing *relations* between the input and output languages. A finite-state automaton equates to regular languages, and finite-state transducers equate to regular relations.

Finite-state transducers constitute appropriate representations of Natural Language phenomena. Indeed, they have been shown to be sufficient tools to describe morphological and phonetic forms of a language ([1] and [6]). Transducers can then be viewed as functions which map lexical representations to the surface forms, or inflected forms to their phonetic pronunciations, and vice versa ([7]. The analysis of Manipuri word with the help of finite state transducers is considered well suited because of the very nature of the morpheme concatenation to form different word forms, the agglutinative nature, and the constraints imposed on morpheme ordering while concatenation.

5. MORPHOTACTICS OF THE NOMINAL CATEGORY

The affixation of suffixes to a stem/ root cannot be haphazardly but follows some strict definite ordering rules. As the suffixes play a major role in identifying the class of the root of a word, our approach is to analyze the words by examining its suffixes from the rightmost position and segmenting the morphemes in a right to left direction. For our convenience, the nominal suffixes in table 4 are assigned with a numeric value as shown in the table 5.

Table 5. Nominal Suffixes assigned with a number

Suffix #	Suffix #	Suffix #	Suffix #
1- nə	5-dəgi	9- du	13- dəmək
2- bu	6- gə	10- də(particle)	14- ni
3- gi	7- siŋ	11- di	15- mək
4- də	8- si	12- su	16- ra

The following finite state network is a nondeterministic finite state machine (NDFSM or NFA) representing the morphotactics of Manipuri nominal category with respect to suffixes. In the figure, single circles represent a state and the double circle represents a final state(s). The numbers inside the circle is the state number, and here the circle with a value 0 is the start state of the FSA. Each transition from one state to another is labeled by suffix (es) (represented here by a numeric value) it can consume. The ε (epsilon)s represent empty transitions. the ε (epsilon)s represent empty transitions. For every string in the language of the network, there is a path, a sequence of zero or more arcs, from the initial state to a final state such that the arc labels along the path constitute a match for the string. A transition occurs when both the input symbol and the label morpheme of the arc matches. There are three conditions the state of the FSM network can be in- i) when the word has been consumed (i.e. there are no more letters left on the input string) and we are at a final state (word accepted and thus recognized). ii) When we can match a letter with a label symbol and so make a transition to a next state. iii) When we can't match a label symbol and the current letter in the input string, and so can go no further (word not in the language and thus rejected.

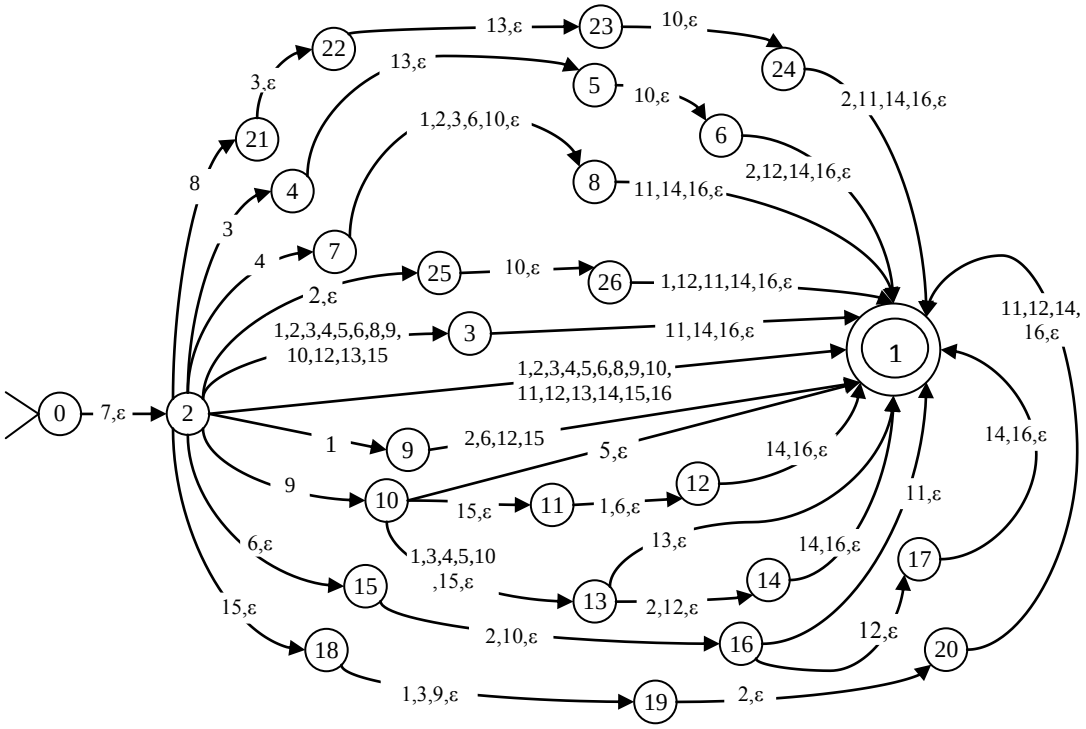

Figure 2: Right to left NFA for Nominal category

.

5.1 Conversion of the NFA to DFA

The empty or ε-transitions and multiple transitions for a single input make the NFAs hard to implement using computer programs. However, it is true that there exists at least one path through the NFA for a string that is in the language defined by the machine, but not all paths directed through the machine for an accept string lead to an accept state. Deterministic means that at each point in the processing there is always one unique action to be taken. Therefore in a DFA, for each single input label there will always be a single and unique transition thereby making programming easy. To convert the NFA in figure 2, the strategy is to remove all the ε-transitions and to make sure that there will be at the most one transition for a single input label. For this purpose we use an algorithm called "subset construction" [9]. The basic principle behind this algorithm is that every NFA has an equivalent DFA which describes the same language and each state of the DFA machine corresponds to a set of states of the NFA. The overall idea is that all the states, which are connected by an ε-transition and reachable by a single input on the current state will be combined into a single DFA state [10]. There are three operations that can be applied on NFA states-

i) ε-*closure*(*s*)- set of NFA states reachable from state *s* on ε-transition,
ii) ε-*closure*(*T*)- set of NFA states reachable from some *s* in *T* on ε-transition,
iii) *move*(*T*, *a*)- set of NFA states to which there is transition on input *a* from some state *s* in the set.

The ε-*closure*(*s*) operation computes exactly all the states reachable from a particular state on seeing an input symbol. When such operations are defined the states to which our automaton can make a transition from set *T* on input *a* can be simply specified as: ε-*closure* (*move*(*T*, *a*)). The DFA states

which has the initial state (here 0) of the NFA will be the initial state of the DFA. Similarly the DFA states which has at least one final state in the set of NFA states (which makes the corresponding DFA state) will be the final states of the DFA.

The following figure 3 shows the converted DFA from the NFA in figure 2.

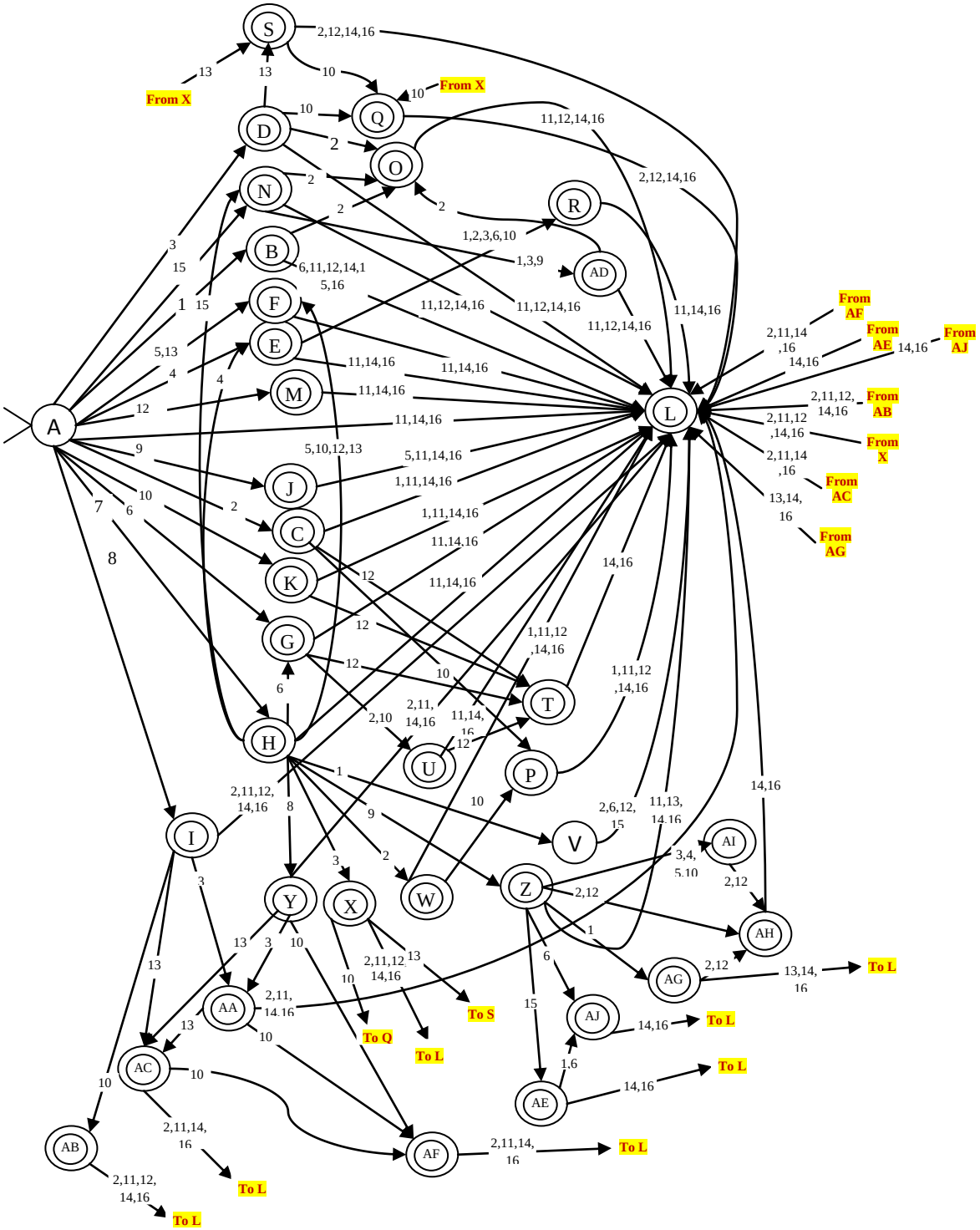

Figure 3: Nominal suffixes Right to Left DFA

6. CONCLUSION

To develop the proposed morphological analyzer, a hand crafted morphotactics with the help of finite state techniques has been developed. The most important thing here is that an attempt has been made to analyze a Manipuri nominal category word without the help of lexicon because of the fact that class of the stems can be decided by the type of suffixes that immediately follows the nominal stem. The rules that define the various word forms of the language using suffixes of the nominal category are captured in the finite state automata. To analyze a nominal word the system checks the suffixes starting from the right, finding a suffix it checks with the rules of the morphotactics, with a match it maps to the corresponding lexical category of the suffix. This iterative process continues till it finds the stem of the word or a suffix that does not match according to the rule of the morphotactics. Here it uses a nominal root table in the database, a table for suffixes along with the allomorphs and its lexical information and the morphotactics as described by the finite state automata.

7. DISCUSSION AND FUTURE WORK

Literature available on the researches for the development of morphological analysis for agglutinative languages like Finnish, Turkish, etc. established the fact that finite state networks are best suitable mechanism to express the morpheme concatenation rule of these languages.

The proposed morphological analyzer can analyze a Manipuri nominal word which has attained the free standing form of nominal category. That means before receiving the suffixes the input words should have qualified as a full-fledged nominal word. The DFA in figure 3 can be reduced to its minimized form by using different minimization technique algorithms, thereby reducing the programming complexity overhead. The result shown by this analyzer is quite appealing; however, the deficiency is that it is very hard to make any change in the FSA for any morphotactic rule. Also, despite the good coverage, there are still a number of grammatical forms that have not been implemented into the FSA.

The major work yet to be done is to develop FSAs for each subcategories of the nominal category to capture their irregular forms before they can be subjected to the main FSA of the nominal category for analysis of the word. The subsequent works follow to develop transducers for other word classes like verbs, adverbs, adjectives, etc. of the language in order to develop a morphological analyzer for Manipuri language, which will facilitate the morphological analysis of Manipuri words.

8. REFERENCES

[1] Karttunen, Lauri, Ronald M. Kaplan and Annie Zaenen (1992). Two-Ievel Morphology with Composition. In the *Proceedings of the fifteenth International Conference on Computational Linguistics. Coling-.92.* Nantes, 23-28/8/1992. Vol. 1 141-48. ICCL.

[2] Karttunen, Lauri, Applications of Finite-State Transducers in Natural-Language Processing, Xerox Research Centre Europe.

[3] Kenneth R. BEESLEY, Xerox Research Centre Europe, Arabic Morphology Using Only Finite-State Operations

[4] Karttunen, Lauri, Constructing Lexical Transducers, Rank Xerox Research Centre, Grenoble

[5] Kenneth R. Beesley and Lauri Karttunen. 2003, Finite State Morphology. Center for the Study of Language and Information, April.

[6] Kay, Martin, and Ronald M. Kaplan. 1994, Regular Models of Phonological Rule Systems. *Computational Linguistics*

[7] M. Mohri, 1994a, Compact Representations by Finite State Transducers, *In Proceedings of the* 32nd *Annual Meeting*, Las Cruces, New Mexico. Association for Computational Linguistics

[8] ARONOFF, M. & FUDEMAN, K., 2004, *What is Morphology?* Blackwell Publishing: Cornwall.

[9] A. V. Aho, R. Sethi & J. D. Ullman, *Compilers: principles, techniques, tools* (Reading, MA: Addison-Wesley, 1986).

[10] Gülşen Eryiğit and Eşref Adalı,2004, AN AFFIX STRIPPING MORPHOLOGICAL ANALYZER FOR TURKISH, Proceedings of the IASTED International Conference ARTIFICIAL INTELLIGENCE AND APPLICATIONS, February 16-18 2004,Innsbruck, Austria

[11] Evan L. Antworth. 1990. PC-KIMMO: a two-level processor for morphological analysis. Number 16 in Occasional publications in academic computing. Summer Institute of Linguistics, Dallas

[12] Kimmo Koskenniemi. 1983. Two-level morphology: A general computational model for word-form recognition and production. Publication 11, University of Helsinki, Department of General Linguistics, Helsinki

[13] Ch. Yashwanta Singha, 2000, Manipuri Grammar, Rajesh Publications.

[14] D.N.S. Bhatt & M.S. Ningomba, 1997, Manipuri Grammar, LINCOM EUROPA