Morphotactics of Manipuri Adjectives: A Finite-State Approach

Ksh. Krishna B. Singha
Department of Computer Science, Assam University, Silchar, India
E-mail: ksworks2010@gmail.com

Kh. Raju Singha
Department of Computer Science, Assam University, Silchar, India
E-mail: raju.singha@yahoo.com

Bipul Syam Purkayastha
Department of Computer Science, Assam University, Silchar, India
E-mail: bipul_sh@hotmail.com

Abstract— This paper presents a constrained finite-state model to represent the morphotactic rule of Manipuri adjective word forms. There is no adjective word category in Manipuri language. By rule this category is derived from verb roots with the help of some selected affixes applicable only to verb roots. The affixes meant for the purpose and the different rules for adjective word category formation are identified. Rules are composed for describing the simple agglutinative morphology of this category. These rules are combined to describe the more complex morphotactic structures. Finite-state machine is used to describe the concatenation rules and corresponding non-deterministic and deterministic automaton are developed for ease of computerization. A root lexicon of verb category words is used along with an affix dictionary in a database. The system is capable to analyze and recognize a certain word as adjective by observing the morpheme concatenation rule defined with the help of finite-state networks.

Index Terms— Morphotactics, Natural Language Processing, Artificial Intelligence, Finite-State Automata, Manipuri Adjectives

I. Introduction

In many natural language processing applications, analyzing the internal structure of a particular word is an important intermediate stage. The task is complicated and difficult for languages like Manipuri, an agglutinative language where morpheme concatenations are the primary means for various word forms and word formation processes. Among other word categories, Manipuri adjectives find a really interesting place in the discussion and analysis of the grammar formalism for the language from the computational point of view. First of all, the language does not have a distinct adjective word category. In familiar languages like English, the bases that translate as adjectives fall in the category of verbs in Manipuri language and are derived from these bases following certain strict and predefined rules. On the other hand the language being an agglutinative language, different word forms are formed by concatenating a number of basic meaningful grammatical units called morphemes. The formation of adjectives is guided by various rules following different morphosyntactic structures in the word pattern. The description of complex linguistic phenomena involving morphological alternations and syntactic patterns of adjectives as well as its absence as a free standing word class in the language poses a challenge to develop a computational model for analysis and generation of the adjectives of the language.

Most of the relevant local phenomena encountered in the empirical study of language can be described easily using finite automata [1]. Their uses range from the compilation of morphological and phonological rules [2] [3] to speech processing [4] show the usefulness of finite-state machines (FSM) in many areas. Also finite state techniques are known to be the most suitable technique to represent the morphosyntactic rule of agglutinative languages like Turkish, Manipuri, Finnish, etc. The main goal of our research in this context is the construction of theoretically motivated computational model to represent the morpheme concatenation of Manipuri adjectives using finite-state-techniques. For the purpose linguistic aspects of the Manipuri adjectives are investigated to find out the constituent elements and the rules governing the formation of this word category in the language.

The outline of the paper is as follows: the first section gives a brief description about the features of Manipuri adjectives and the affixes which are required in its formation; the following section describes the various

rules followed while morpheme concatenation that governs the formation of adjectives in this language. The third section describes the finite state machines to represent the morphotactic of this word category. It comprises conversion of the Non Deterministic Finite Automata (NFA) to Deterministic Finite Automata (DFA). In the fourth section we experiment the system with some selected texts and analyze its functionality with variations and suggestions for enhancement of the system. The last section draws conclusion of our work and suggestion for further work.

II. Manipuri Language and Manipuri Adjectives

Manipuri language (also known as Meitei-lol or (Meitei-lon in the language itself) is the predominant language and the lingua-franca of the state of Manipur in north eastern part of India. The language belongs to the Kuki-Chin subgroup of the Tibeto-Burman group. Though the language has its own script, called Meitei-Mayek, the script used for writing the language currently is Bengali script. Manipuri is a tonal language and agglutinative in nature. By concatenating a good number of morphemes various word forms can be generated, the end product may be inflected or derived word form of the stem or the root word. There are only two available word classes in this language- noun and verb; all the other word classes are derived from the verb roots in the language. Manipuri adjective is one such class where the adjective words are derived from the verb roots with the help of some affixes. In our current work we focus on this word class and its structure components of the language. So to study the internal structure of adjectives in Manipuri, it is very essential to understand the verbal bases available in the language. This is because the language does not have a unique word category for adjectives; and all the words identified as adjectives in familiar languages like English are translated as derived words from bases of verb category using certain selected affixes in Manipuri language.

There is no distinct category of adjective in Manipuri as 'long', or 'hot' in English. It can be elaborated by the following example:

əsaŋbə (long) is derived from the verbal base *saŋ* by affixing prefix *ə* and suffixing *bə*.

ə (Prefix) + *saŋ* (Verb Root,VR) + *bə* (Suffix) /long

There are more examples:

i) P^{h}əjəbə lei / beautiful flower

P^{h}əjə (verb base) + bə (Suffix)

ii) cəttəbə g^{h}əri / Stopped watch

cət(verb base) + *tə*(Suffix) + *bə* (Suffix)

iii) cətkədəbə əŋaŋ / kid who'd be going

cət(verb base)+ *kə* (Suffix) + *də* (Suffix) + bə (Suffix)

As can be seen from above, the adjectives in the language are formed by deriving the verb roots with the help of certain selected affixes (not all the affixes meant for verb bases are required to form adjectives). Whereas the number of prefixes that is affixed to a verb base in the process of adjective formation is limited to a single one (i.e. the formative ə), the number of suffixes can be many. An important point to be noted here is that all the adjectives have the nominalizer suffix *'bə'~'pə'* at the rightmost position of the word.

The concatenation of the affixes follows certain order and must not be violated in order to be a valid Manipuri adjective (or word). A group of affixes, with the help of which Manipuri adjectives are formed has been identified, of which the nominalizer is essential as the rightmost suffix and the formative prefix 'ə' is required when there are no morpheme at all in between verb root and the nominalizer *'bə'~'pə'*. The following table shows the affixes with their category:

Table 1: Affixes used in adjective formation

Affix #	Affix Name (with its allomorphs)	Affix Type	Category
1	অ (ə)	Prefix	Formative
2	ব/প (bə)/(pə)	Suffix	Nominalizer
3	রক / লক (rək / lək)	Suffix	Inceptive
4	ন (nə)	Suffix	Nominative
5	লি / রি (li / ri)	Suffix	Progressive
6	কন / গন (kən/ gən)	Suffix	Habitual
7	র / ল (rə /lə)	Suffix	Prospective
8	খ্র (khrə)	Suffix	Perfective
9	কাই / গাই (kai/gai)	Suffix	Destructive
10	রম/লম (rəm/ləm)	Suffix	Evidential
11	রু/লু (ru/lu)	Suffix	Deitic
12	ক /গ (kə /gə)	Suffix	Potential
13	খি (khi)	Suffix	Perfective
14	খৎ/কৎ/গৎ (khət/ kət/ gət)	Suffix	Directional
15	থোক/দোক/তোক (thok/dok/ tok)	Suffix	Directional
16	শিন/জিন/চিন (sin/jin/cin)	Suffix	Directional (inward)
17	রোই / লোই (roi/loi)	Suffix	Negative Marker
18	ত / দ (tə/də)	Suffix	Negative Marker
19	ত / দ (tə/də)	Suffix	Dubitative
20	ত্র / দ্র (trə /drə)	Suffix	Negative Marker
21	ত্রি/দ্রি (tri /dri)	Suffix	Negative Marker

2.1 Rules for Adjective Formation

The most popular way of forming adjectives is to add the formative prefix ə and the nominalizer suffix to the verbal base of the language. The following describes the rules that are followed while forming adjectives:

Rule 1) Formation of adjective by adding the formative prefix 'ə' and a suffix called nominalizer(NZR) *'bə'~'pə'* to verb root (VR). The general syntax for the rule is-

Prefix + Verb Root (VR) + NZR → Adjective

Some examples are:

ə+waŋ+bə u → əwaŋbə u / tall tree

ə+səŋ+bə Pʰurit → əsəŋbə Pʰurit /green shirt

It is observed that addition of ə without the NZR suffix *bə~pə* is not considered to be a complete form of a Manipuri word.

Rule 2) There are exceptional cases where the formative prefix *ə* is not required to be added to the verb base, however addition of the NZR suffix *bə~pə* is essential. See the following example:

nuŋsi + bə mənəm → sweet smell

həraw + bə numit → joyful day

It can be noted here that the addition of formative prefix *ə* is possible only to monosyllabic base and not to polysyllabic bases. e.g.

* ə + *pʰəjə + bə* / (*pʰəjə + bə* is ok)

* ə + *warə + bə* / (warə + bə is ok)

* ə + *thəkli + bə* / (thəkli + bə is ok)

The polysyllabic bases are either composed of verb bases attached with some selected suffixes (say aspect markers) or compounds (opaque) like

Pʰə(Verb Root) + *sa* (Verb Root) + *bə* (NZR)

Here Pʰə(good) and sa (make) are two different verb roots and the compound word Pʰəjə is derived from these two roots [5] and attaching the suffix NZR to the compound, forms the adjective Pʰəjəbə.

Take another example:

sək(Verb Root) + tə(NM) + bə(NZR) isei / the song that has not been sung

This rule can again be classified as the following:

Rule 2.1) Suffixing the negative marker (NM) tə~də to the verb base and followed by the NZR (bə~pə) the formative prefix (ə) is no more required to form adjectives.

Examples are:

ca(VR) + də(NM) + bə (NZR)   cak / uneaten rice

ca(VR) + roi (NM) + də (NM) + bə (NZR) cak / rice that won't be eaten

whereas

*ə + ca(VR) + də(NM) + bə (NZR)

is an invalid word.

Rule 2.2) The formative prefix (ə) is not required if the verb root is immediately followed by an aspect marker (AM) which is again followed by a nominalizer (NZR). Such as

i (VR) + *ri* (AM) + *bə* (NZR) *məyek* / written script

Rule 2.3) When the verb root is attached with the nominative suffix nə (MN) and is followed by the nominalizer then the prefix ə is not required to form adjectives.

Say for example:

tum (VR) + *nə* (MN) + *bə* (NZR) *ka* / Sleeping room

is valid form of a Manipuri word and is qualified as adjective, however

* *ə*(Prefix) + *tum* (VR) + *nə* (MN) + *bə* (NZR) ka

is not a valid form of a word.

Rule 2.4) Analyzing each syllable of verb roots which are polysyllabic is an indispensible task and also an exceptional case as far as the formation of adjectives in this language is concerned. It is observed that most of the polysyllabic roots are probably compounds, made up of two or more verb roots.

Take the case of following examples where the formative prefix is no more required to form an adjective.

nuŋsi (VR) + *bə* mənəm / sweet smell

*ə + *nuŋsi* (VR) + *bə* mənəm / invalid word

Here the verb root contains two syllables nuŋ and si and are two verb roots nuŋ (internal) and si (keep), conveying the meaning that if we love somebody we keep him/her internally. So the polysyllabic root nuŋsi can be treated as opaque compounds.

Another example here is

nuŋon (VR) + *bə* (NZR) paw /heartening (or emotional) news

nuŋon also consists of two syllables *nuŋ* (internal) and *on* (move or change) and act as compounds.

Syntax is

Compound + NZR → Adjective

It is worth mentioning here that in case of rule 2, the number of suffixes (listed in table 1) intertwined between nominalizer suffix and the verb bases may be more than one.

In addition to above mentioned adjectives, there are some more adjectives which do not require any affixes to be attached with any root in order to qualify as an independent Manipuri adjective. These are some color terms such as *Higok* (blue), *Leiŋaŋ*, etc.

III. Finite-State Morphotactics

The linguistic components of a word in agglutinative languages are morphemes that are concatenated together one after another following certain strict rules. The description of the arrangement of morphemes in words is called morphotactics, e.g. the properties of the nominalizer *bə~pə* that it is a suffix and it only follows a verb base. With different combinations and number of morphemes attached to the base produces various inflected and/ or derivational word forms. The formation of Manipuri adjectives is also not an exception, where adjective words are formed with the help of various affixes attached to the verb bases. Obviously it follows certain predefined rules while formation of this word category. In order to develop a computational model for analysis and generation of Manipuri adjectives, it is required to identify the morphemes and their position within the word structure. From a computational efficiency point of view, it is important to recognize the sequence and the order of morpheme in the adjective words through a pure finite-state mechanism rather than through more complex parsing procedures. The morphotactics of the adjective word formation in Manipuri language are represented by the rules defined in section II. Our approach captures the rules using finite-state automata (FSA).

3.1 Representation of Manipuri Adjective Morphotactics using Finite-State Automata

The various rules comprising the adjective word category of the language are represented using finite-state automata. The following finite-state automaton represents the morpheme ordering i.e. morphotactic of rule 1) of section 2:
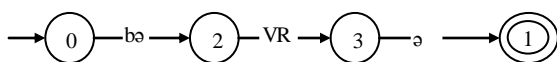

Fig. 1: FSA for Rule 1

The numbered circles represent states and the labeled arcs represent transitions from one state to another. Here the start state is the circle numbered with 0. The double circle denotes the final state. The machine approaches from the right side of the input word, i.e., the rightmost morpheme will be checked and matched

first. The labels with each arc suggest that a transition is possible only when the labeled string is matched with the input text. So, when a bə is read by the machine, a transition is made from the initial state to the next state, i.e. state 2. The automata will then move to the next state after examining the second input, i.e. a verb root. The final state will be reached if the input to the next state is a string ə. At any stage, if the input does not match with the predefined string, the input string will be rejected.

While the prefix and suffix required in the first rule were fixed, the Rule 2 in section II consists of a selected number of suffixes that may come in between the verb root and the nominalizer. Here the normal adjective formation process of adding the formative prefix ə and a nominalizer suffix bə~pə is not followed in deriving adjectives from verb roots. The prefix ə is no more required to form adjectives if there are suffixes in between the verb root and the nominalizer suffix. It can be represented using the following finite-state network.
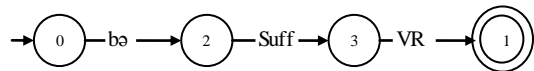

Fig. 2: FSA for Rule 2

Here the *Suff* stands for any suffix which is qualified to be concatenated in between the verb root and the nominalizer.

Though the number of suffix (es) which appear in between the verb root and the nominalizer is only one [5,6,7], more number of suffixes in between are also observed as in the following:

Sat(VR) + kai(Suff) + rə(Suff) + bə(NZR) leirang / an already bloomed flower.

When we enter the above input text in our automata of figure 2 (for rule 2), the machine approaches in the following way: the presence of the nominalizer suffix is checked from the rightmost position of the input text. After matching the suffix, a transition is made and reaches state 2. From here it finds the suffix *rə*, it searches for a match in the affix database and reached the next state. Matching for suffixes is done for the next state also till the end. As the last label on the arc is not a suffix it does not matches and so it will search for a match in the verb root database. The matching is done and reaches the final state. Now as the machine is in the final state and the input string matched, the input text is recognized to be a valid Manipuri adjective word.

3.2 Automata Construction

Now an NFA is constructed to represent the adjective formation rules (both rule 1 & 2). For our convenience, a Finite-State Machine (FSM) is constructed first, to identify each path that takes the machine to a final state

from its initial state. Every transition from one state to another is labeled with an affix from table 1 above. It may be noted here that every transition that reaches the final state of the machine is possible when the nominalizer suffix bə~pə is matched with the input. The following figure 3 illustrates the ordering of the morphemes, i.e. the morphotactical features in the word structure of Manipuri adjectives.
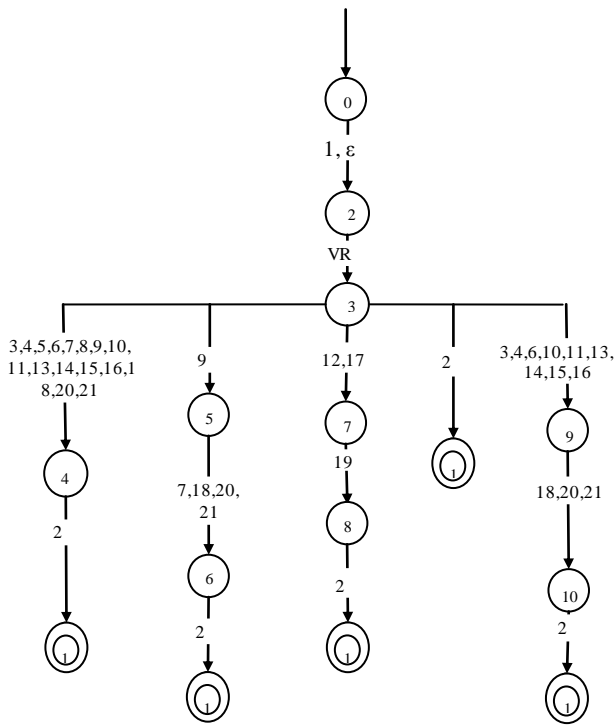

Fig. 3: Left to Right FSM for Manipuri Adjective words

The affixes of table 1 are assigned with a numeric value which corresponds to the affix number of the table. Here a single numeric value refers to all the allomorphs of the same affix. The numbered circles represent states and the arcs connecting the circles are labeled with affixes which are the numeric values assigned to it as per the table in section 2. The circle with the numeric 0 is the initial state of the FSM and the double circles (with numeric 1) are the final states. The ε (epsilon) symbol is used to represent an empty transition. An empty transition is said to occur when the machine transits from one state to another without consuming any input. In the figure a group of affixes for a single transition means the transition can be made with each labeled affix of the arc, individually.

The figure 4 below is the corresponding NFA for the above FSM where there is more than one transition from a state for a single suffix, as well as one ε transition.
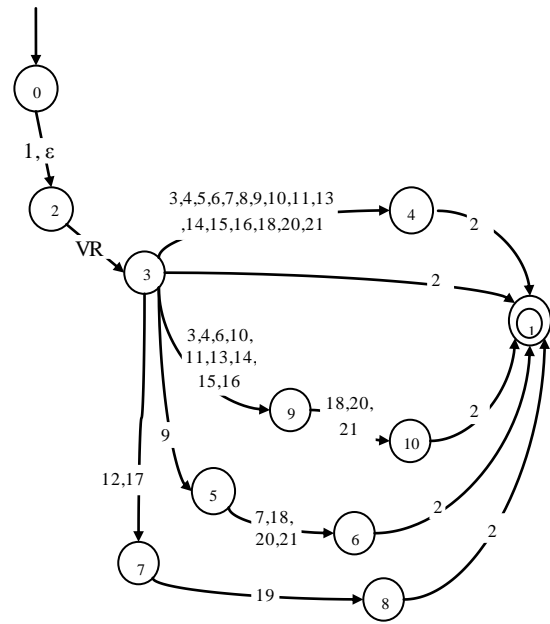

Fig. 4: Left to Right NFA

Programming NFAs are difficult logically that for some pathological cases they may take exponential time to figure out that there is no match. As in our FSM of figure 4, at state 3 the FSM has two options to make a transition for the single suffix #9. Each of them will lead the automaton to two different paths and so it may take ambiguous decision and therefore may take longer time to reach to the correct match. As we know that every NFA has an equivalent corresponding DFA, one best solution here is to convert the NFA to a deterministic finite-state automaton (DFA) before its computational implementation is done. A DFA offers performance guarantees which is characterized by the fact that- there is one and only one transition from a state for a single input label and there are no empty transitions. We use the "subset construction" algorithm [8] to convert the NFA to its corresponding equivalent DFA. The idea behind this algorithm is that every DFA state corresponds to a set of NFA states which are connected by a ε-transition and can be reached by a single input. In other words, after processing a sequence of input symbols the DFA is in a state that actually corresponds to a set of states from the NFA reachable from the starting symbol on the same inputs.

We adopted a method to segment the input word into its constituent morphemes from the rightmost position towards left. If the input text is an adjective, either the verb base or the formative prefix ə is expected to be found at the leftmost position. As all the adjectives ends with the nominalizer suffix it is worthy to do the segmentation from the right. The following figure shows the inverted as well as converted Right to Left DFA of the above Left to Right NFA.
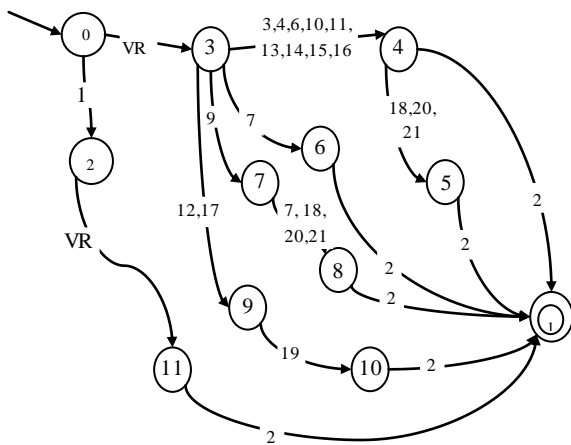
Fig. 5: Left to Right DFA

IV. Experiment and Analysis

Applications related to text are required to have very detailed information about the structure of the word. These include information about the constituent morphemes (root, affix, particle, etc.), their lexical category and corresponding surface forms, the property of the morphemes i.e. inflectional or derivational, the category of the word, and last but not the least the ordering of morphemes in the word structure, called morphotactic of the language. The use of morphotactic is mainly in the morphological analysis of words that caters to the needs of variety of applications like spell-checking, machine translation, information retrieval, etc. The validity of a word form is verified by the morphotactic rules and hence played an important role in the morphological analysis of languages with complex morphology and agglutinative in nature. So the performance of an NLP application that uses morphological analyzer will be highly influenced by correct analysis and verification of morphotactical features of the various word forms.

The suffixes which are identified for the purpose of adjective formation are stored with a unique identity of numeric values in the database. Allomorphs of these suffixes are also stored in the same table along with the main suffixes and are referred to with the same identity as its parent suffix. To test the system, we have a database of about 300 verb roots. We choose unique adjectives out of the verb roots to verify the validity of the morpheme ordering in the word structure.

In order to elaborate the working of our system, let us consider the adjective word

*cagəndəbə / no*t so used to eat

Here the morphemes are

ca → Verb Root

gən → Habitual suffix

də → Negative Marker and/or dubitative suffix

bə→ Nominalizer suffix

It may be noted here that while verifying this word the system encounters two similar suffixes viz. *də* morpheme in the database. The ambiguity created by the existence of two suffixes with same pattern is handled by the DFA by identifying the preceding morpheme i.e. the habitual suffix *gən*; because here the negative marker suffix #18 can follow the habitual suffix whereas the dubitative suffix #19 cannot.

The performance of the system is good though the coverage should be increased to enhance the performance of the system. In Manipuri formation of adjective words with the help of opaque compounds are very common, like *Səkhenbə*, *Khoiraŋbə*, Nuŋnaŋbə, etc. Also gender markers are also observed in the adjective markers which occur with certain selected nouns such as *Huranbi*, *Luciŋbi*. However, our database does not have compound words (opaque compounds). But when they are attached with affixes and follows the morphotactic rules as per our FSM, our system fails to recognize the word form as adjective, as the compound word acting as verb root is not found in the database.

e.g. səή/ban/nə/bə is an adjective word.

When this text is entered, it will segment the word starting from bə, proceeds with nə. As per the second rule FSM, the remaining segment should be in the root directory to qualify to be a valid adjective and would search for it. Sənban is an opaque compound word and since it is not listed in the root database our system cannot recognize the word form as adjective. Listing all these compounds in the database may resolve this issue however, but if the system is extended for the purpose of morphological analysis, it should be handled in a way that the compounds are broken into its constituent morphemes. But the point here is that world languages tend to borrow words from other languages of high influence; and in a lesser popular language like Manipuri, the possibility of loan words and coining of new words is very high. So it is an open ended question whether to keep a list of all roots/compounds in the database or just the affixes to make room for the loan words and the newly coined words. Our idea here is that it should be that application dependent.

Our system uses Bengali script for the text input where occurrences of conjunct characters are usual in writing systems of these scripts. One issue here is that the system cannot identify these characters and cannot segment if it occurs at the morpheme boundary. Again solution here may be to identify all such characters and list them in the database along with its constituent characters. Regular expressions are a good option to identify the text patterns.

V.   Conclusion and Future Work

The morphotactical rule of the Manipuri adjective word formation derived from the verb bases are presented here with the help of finite-state techniques. A morphotactic rule recognizer system with the help of hand crafted rule automaton has been developed to implement the system which will verify the valid adjective word forms of the language. The system aims to identify and recognize the constituent morphemes in a word and decide the validity of the word form as adjective. To do this a word is segmented into its root and affixes. The segmentation is processed from the rightmost position as all the Manipuri adjectives ends with the nominalizer. Addition of new suffix to this system is allowed with an option to update the existing FSMs and the corresponding rules associated with it.

Our work can easily be extended to perform morphological analysis of Manipuri words. This can be achieved by developing finite-state transducers (FST) for the various word categories available in the language to represent morphotactical rule as well as mapping the surface forms of the morphemes to its corresponding lexical forms.

References

[1]   Mohri, Mehryar. Finite-State Transducers in Language and Speech Processing[J]. Computational Linguistics, 1997, 23(2): 269–312

[2]   Kaplan, Ronald M. and Martin Kay. Regular models of phonological rule systems[J], Computational Linguistics, 1994, 20(3): 331-378

[3]   Karttunen, Lauri, Ronald M. Kaplan, and Annie Zaenen, Two-level morphology with composition[C], In Proceedings of the Fifteenth International Conference on Computational Linguistics, 1992, Aug 23-28, 141-148

[4]   Mohri, Mehryar, Fernando C. N. Pereira, and Michael Riley, Weighted automata in text and speech processing, In Proceedings of the 12th biennial European Conference on Artificial Intelligence (ECAI-96), Workshop on Extended finite state models of language. Budapest, Hungary, 1996. John Wiley and Sons, Chichester

[5]   S. Imoba Singh, L. Sarbajit Singh. Manipuri Adjectives- A New Approach[J], Linguistics of the Tibeto-Burman Area, Volume 25.2, Fall 2002.

[6]   Singha, Ch. Yashwanta, Manipuri Grammar[B], Rajesh publications, New Delhi, 2000.

[7]   Bhatt, D.N.S, Ningomba, M.S., Manipuri Grammar[B], LINCOM Europa, 1997

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

Authors' Profiles

Ksh. Krishna B. Singha: Post-graduated Research Scholar and a student for doctor degree for Computer Science in Assam University, Assam. India, research interest in Natural Language Processing and Fuzzy Logic.

Kh. Raju Singha: Post-graduated Research Scholar and student for doctor degree for Computer Science in Assam University, Assam, India, research interest in Natural Language Processing and Data Communication & Networking.

Bipul Shyam Purkayastha: Professor of Department of Computer Science in Assam University, Assam, India, interested in natural language processing and automata theory.