Finite State Automata as Analyzers of certain Grammar Class Rules in the Khasi Language

Aiusha V. Hujon
Department of Computer Science, St. Anthony's College Shillong, Meghalaya, India
Email: avhujon@gmail.com

Abstract—In this paper the concept of Finite states is applied to design Recognizers for certain morphemes in Khasi language. Individual Grammar Class is implemented using Finite State Automata. Then Finite state automata are implemented to design a Morphological Analyzer for Nouns and Verbs. These Finite state automata can be concatenated one after the other to recognize Sentence Rules. Hence these can be used to design a Morphological Analyzer for legitimate sentences in the Khasi language.

Keywords— Morphological analyzer, Finite state automata, khasi language, word analyzers, Austroasiatic languages, word recognizer, Natrual language processing.

I. Introduction

The Khasi Language is an Austroasiatic language. According to classifications of Austro-Asiatic Language Family, Austroasiatic language is divided into Munda and Mon-khmer, Mon-khmer is divided into languages in the North and East. There are three languages in the North; the Khasi, the Khmuic and the Paluangunic. Khasi is a verb medial language; its basic word order pattern is Subject Verb Object (SVO)[1]. It is spoken by 1.6 million native speakers (2001 census), residing in the Northeastern region of India, particularly in the state of Meghalaya. Many Morphological Analyzer and Generator have been constructed for many languages in India. But till date a working Morphological analyzer and generator for the Khasi language has not been constructed, although some research is still going on in this area.

II. Application of FSA to construct certain Khasi words Analyzers

A Finite state automaton (FSA) can be used to determine valid strings in a particular language [3]. Natural languages have their own valid words which are form using letters from their respective character sets. We can use FSAs as a morphological analyzer to determine whether an input string of letters makes up a legitimate Khasi word or not. Consider a few words in Khasi which are Nouns as given in Table I. A Finite State Automaton for each word listed in Table I is constructed with path sequence from the initial state to the final state. The Finite State Automaton given in Fig. 1 can determine whether a string of letters is a Noun in the Khasi language or not. Many more automata can be constructed like this for all the Nouns in the Khasi Language.

Similarly Verbs can be recognized by an FSA as shown the Fig. 2 as FSA 'B'. There are certain Verbs in Khasi that can be converted into a Noun by adding a morpheme "jing" as a prefix. Such as the verb "Khang" (close) can become a Noun when prefix with "jing", i.e., "jingkhang" (door) which is a Noun, "bam" (eat) can be come "jingbam" (food), "thiah" (sleep) can become "jingthiah" (bed). For these kinds of Verbs the FSA for recognizing these verbs can be concatenated with the FSA for the morpheme "jing". The diagram in Fig. 2 shows how this can be done. Let the FSA 'A' recognize the prefix word "jing" and FSA 'B' recognizes certain Verbs like "thiah", "khang", "bam".

TABLE I

A Few Khasi words which are Nouns	
Khasi	Meaning in English
Briew	man
Mrad	animal
Khiew	pot
Shuki	chair
Ing	house
Ksew	dog
Miaw	cat
Dieng	tree
Pliang	plate
Pela	Cup

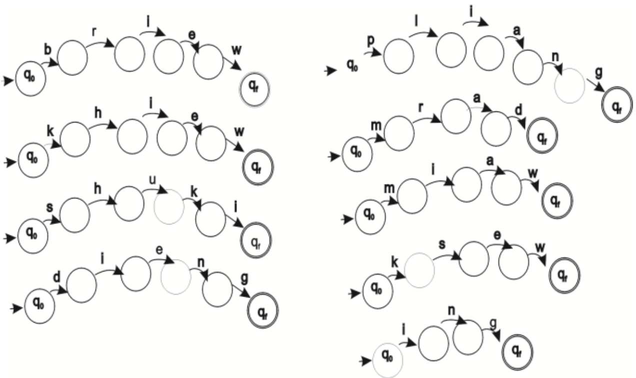

Fig 1. Finite state automata for certain Nouns in Khasi

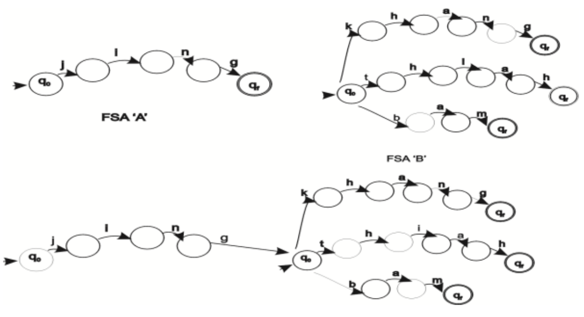

Fig 2. FSA recognizing certain Nouns that are prefix with "jing" and certain Verbs in Khasi

Using the same concept, we can build an FSA for each Morpheme in Khasi. There are many other words that are form by using prefixes like “jing” in khasi, e.g., like “pyn” which can also concatenate with “bam” (eat) to form “pynbam” (to feed/feeding). FSA for each class of words like Noun, Verb, Adjective, Adverb, Determiner, etc., can be constructed as shown in Fig. 2.

III. Sentence Rule and FSAs

The word form in Khasi is mainly Subject-Verb-Object (SVO). Most of the sentences in Khasi are of the following forms, there are many others which are not possible to describe in this paper. The following are just a few of these forms.

1) Noun Verb e.g., Nga rwai (I sing)
2) Noun Verb Noun e.g, Nga bam ja (I am eating food)
3) Determiner Noun Verb Adverb e.g., U John u long u briew uba bha (John is a good man)
4) Determiner Noun Verb Adverb Preposition Verb Adverb e.g., U jah jlang khlem iohi shuh (He disappear without seeing him again)
5) Determiner Noun Adjective e.g., Ka kot jingkhein (the arithmetic book)

The FSAs that are shown in Fig. 1 and Fig. 2 are sub-FSAs and can be used to construct a full FSA for recognizing legitimate sentences in the Khasi Language. For each of the sentence forms above, a corresponding FSA can be constructed as in the shown in Fig. 3. The labels for each edge are given the name of each grammar class, like Noun or Verb, but instead practically each of the corresponding FSAs for Noun, Verbs, Adjectives, etc., can be inserted where ever the label appears in the FSA for recognizing valid sentences in Khasi.

As an example the words “Nga” a Noun, “Bam” a Verb, and “Kwai” another Noun can be formed into a sentence by ordering them using the sentence rule Noun Verb Noun, i.e., “Nga bam kwai” which means “I’m eating betelnut” in English. An FSA is shown in Fig. 4 which accepts this sentence in Khasi.

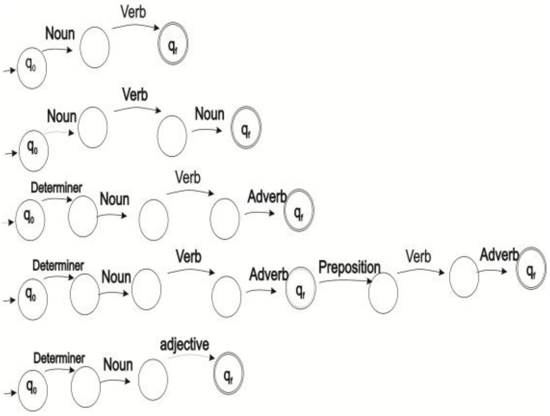

Fig 3. Finite state automata for accepting Sentence Rules in Khasi

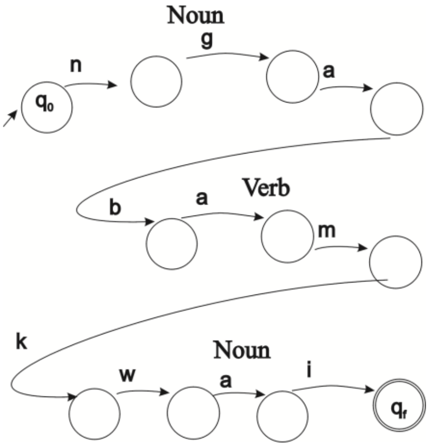

Fig 4. Finite State Automaton To Accept The Sentence “Nga Bam Kwai” (Noun Verb Noun) In Khasi

IV. Conclusion

Since there are no such Analyzers and generators for the Khasi Language, this paper presents a very preliminary investigation to design Morphological Analyzers and Generators particularly for Khasi Language using certain legitimate words forms as examples to construct Finite state automata to recognize it. These Analyzers are very initial considerations and could be used to design Morphological Analyzers to process the Khasi language and could be applied to design Morphological analyzer and Generator in certain areas of Machine translation for the Khasi Language.

References

[1] Diffloth, 1974 cited in Encyclopedia Britannia and Diffloth, 2005
[2] Daniel Jurafsky, James H. Martin, Speech and Language Processing, 4th Impression, Pearson Education and Dorling Kinderseley Publishing, 2008
[3] John E. Hopcroft and Jeffery D. Ullman, Introduction to Automata Theory, Languages, and Computation, Narosa Publishing House Pvt. Ltd. New Delhi, 2002
[4] H.W Sten, Ka Grammar Khasi, Gratus Publication,Shillong,