Kokborok Morphological Analyzer using Stemmer

Abhjiit Debbarma
Ramkrishna Mahavidyalaya
Department of Information Technology
Kailashahar, Tripura, India

ABSTRACT
Morphological analyzer tries to analyze the morphemes of the given word. Given a word it will analyze the different aspect of grammars, affixes and root words. This work tries to study the Morphology of Kokborok language word and proposed to develop a light Kokborok Morphological Analyzer using a Stemmer. Kokborok is a member of a Tibeto Burman language family. It is the official language of the state of Tripura situated in North Eastern region of India. The morphological analyzer focuses on Numbers, Genders and Tense aspect of Kokborok words. With the limitation in size of database for an under resource language like Kokborok, the analyzer gives encouraging results.

General Terms
Morpholigical analyzer, Kokborok, Stemming Kokborok words.

Keywords
Kokborok, NLP, Stemmer, Morphological analyzer.

1. INTRODUCTION
Morphological analyzer gives an output based on analysis of morphemes of a given word. It is the segmentation of words into separate components or morpheme. Morphological rules are applied to obtain a satisfactory result. On applying rules we obtain various information of a word. The word ***'cats'*** can be divided into two units, root word '***cat***' and suffix '***s***'. We can thus get additional information from its analysis of suffix, that it is plural and consist of many cats. Stemming of word into separate components of morpheme will ease the work of Morphological analyzer. We propose here a morphological analyzer using a stemmer. Stemming word makes the basic into morphological analysis of word in a language. Knowing the root word helps widely in information retrieval study. Like most Indian languages Kokborok is highly inflectional language. Porter [2] algorithm discusses about suffix stripping to find the root words. This work tries to study the word analysis and develop a Kokborok Morphological Analyzer. A similar work has been done for Hindi, Gujarati, Bengali, Punjabi, Tamil, Kanada [6] languages. There has been no reported work for Kokborok language. As per our information this is the first attempt made to develop the morphological analyzer for Kokborok.

2. KOKBOROK
Kokborok is the language of the Tripuri people spoken in the state of Tripura situated in the North Eastern part of India and also adjoining areas of Chittagong Hill Tract in Bangladesh.
This language belongs to the Tibeto-Burman language family [4] and shows close affinity with other North Eastern language like Boro, Garo etc. Earlier Kokborok has its own script, used by the Tripura Royal Family, but presently it uses either the Bengali script or Romans Script for writing of which the later has greater acceptance among the educated Tripuri intellectuals. Kokborok has been recognized as the official language of the state of Tripura.

The vowels consist of the following letters
'A' as in Father, ama
'E' as in End, bereng
'I' as in Inside, bini
'O' as in Hot, bolong
'U' as in Put, buwa
W' as in kwtal, (IPA ω)

The consonants of Kokborok are as:

B C D G H J K L M N P R S T Y along with combined consonants CH, KH, PH, TH, NG.

3. KOKBOROK ANALYSIS
3.1 Verb analysis of suffixes
Present Tense: Suffix ***O*** is used to indicate present tense when added to the root word of a verb.
Past Tense: When ***Kha*** is used with the root word, it indicates past tense.
Future Tense: ***Nai*** is used to indicate future tense when it is added to the root word.
Present Continuous Tense: the Suffix ***wi*** is used for present continuous tense.
Ending in ***kha***: When the verb ends in ***kha*** the verb becomes past tense.
Nwng***kha*** (drank)
Ending in ***nai***: when the word ends in nai, it becomes future tense.
Nwng***nai:*** (will drink)
Ending in ***O:*** When ***o*** is suffix at the end the verb remain in present tense.
Nwng***o***
Other suffix are like root + imperative suffix as in chah (eat) + di = chadi ((you) eat)
Negation: it adds ***'ya'*** to the root word as in nwng(drink) + ya = nwngya (will not drink)

3.2 Person analysis
Kokborok has three persons like other languages. First Person, Second Person, Third Person. The persons are used in terms of relations as in father, mother, brother etc. When '***a'*** is added before the noun, it becomes first person sometimes without any prefix being added to the noun it is understood to be first person, and when '***nw'*** is prefix before the noun it becomes second person and for third person '***bu'*** is added before the noun.

Example:
pha => father:
Apha => my father
Nwpha => your father
Bupha => his/her father.

3.3 Number analysis

The numbers are denoted in Kokborok by using '***song'*** and '***rog***' suffix and plural forms of the word are formed. The plural marker '***song'*** is used to form plural number for relatives and kinship relation and for other plural forms of the words are denoted by uses of '***rog'.*** Example:
kiching (friend) + song = kichingsong (friends)
manwi (thing) + rog = manwirog (things)

3.4 Genders

Gender component of Kokborok words are visible when nouns have ***jwk*** as their suffixes. When ***jwk*** is found as the ending word it represents the feminine gender else it is understood as masculine gender. Sometimes the suffix '**la**' is used to denote masculine gender.

Tok (cock) + jwk = tokjwk (hen)
Tok + la = togla (cock)
Sa (child) + la = sala (son)
Sa (child) + jwk = sajwk (daughter)

4. STEMMER

We used Kokborok stemmer as developed in [5]. The stemmer was developed based on the grammar [4] using the longest suffix removal method by and crafted rule applied on the word. On successfully stemming and deriving its suffixes, the suffixes are put on to the morphological analyzer for analysis.

Table 1: Kokborok Stemmer

Words	Root word
Phai (to come)	phai
Phainai (will come)	Phai
Phaikha (came)	Phai
Phaima (to come)	Phai
Phaidi (come)	Phai
Phaio (come)	Phai
Phaiwi (coming)	Phai
Bwkha	Bwkha
Nogo	Nok
Bwchabo	Bwchap

5. MORPHOLOGICAL ANALYSER

The morphological analyzer starts working on receiving the suffix output of the Kokborok Stemmer. The morphological analyzer is an important aspect in the Natural Language Processing. It is an important tool in many application of NLP.
We introduce a set of rules for the morphological analyzer based on the word analysis that we have discussed earlier. The stemmer forms the basic tools that will help us in our analyzer.
The morphological analyzer looks for the affixes to derive to a meaningful conclusion of the word. The suffixes when present in a word, gives sufficient information of the word. Among the many meaning of suffixes some are discussed in given table below.

Table 2: Selected Kokborok Suffix

Suffix	Meaning
Song	Plural (kinship)
Rog	Plural (general)
Kha	Past tense
Nai	Future tense
O	Present tense
Wi	Present continous
Ya	Negation
Khu	Third person
Bw	Third person(kinship)
Nw	Second Person (kinship)
A	First Person
Da	Questionnaire

The morphological analyzer works as, given a word to analyze, it goes through different step before the analyzer gives output.

Step 1: We input the Kokborok word for analysis
Step 2: The word goes through the Kokborok Stemmer that was previously developed [5].
Step 3: Affix output is obtained for analysis
Step 4: The output Prefix/Suffix are then check with our rules to get required information of the word.
Step 5: Morphological Analysis end.

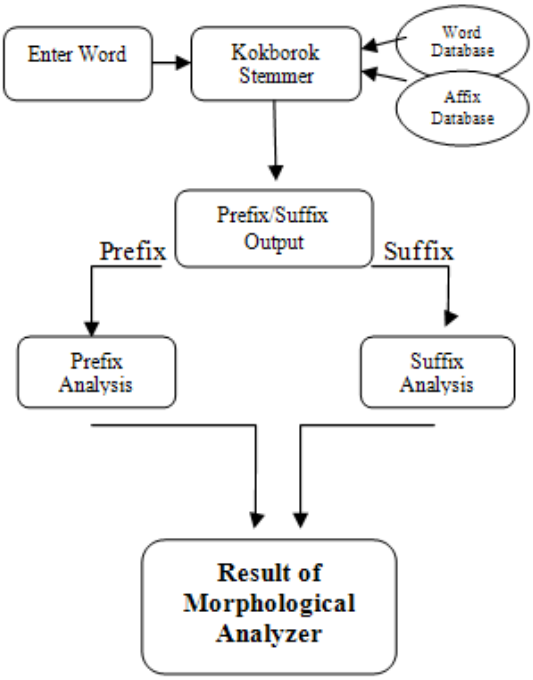

Figure 1: Kokborok Morphological Analyzer

6. RESULT & ANASYSIS

As there is no much available computational resource for Kokborok language, for the analysis of this work, we have used the database of Kokborok words as in [5]. The database consists of 7924 Kokborok words which were collected from various online website including social website like Facebook. The words are being tested to obtain the correct information of their representation. We got correct analysis for 5071 words, which is 64% correct result, whereas 2753 words yielded wrong information. The wrong information resulted mainly due to misspelling of the words in the database. 36% of the incorrect analysis occurred due to absence of rules and improper stemming of the words.

7. FUTURE WORK

We have tried our best in developing a simple Morphological Analyzer for Kokborok language. However due to shortage of sufficient database or word corpus, we could not verify our works as a complete morphological analyzer. We intend to extend the work firstly increasing the database size; secondly, deeply going through the linguistic aspect of the language, which will help us in getting better and efficient result. The morphological analyzer is tested manually for accuracy based on our database which is very small to develop a complete working morphological analyzer. However with the drawbacks in having a computational linguistic resources we have tried our best to have a simple working morphological analyzer. We would be looking forward to enhance our work further to develop a complete morphological analyzer as the future work.

8. REFERENCES

[1] A Ramanathan, Durgesh D Rao "A lightweight Stemmer of Hindi", Proceeding of EACL. 2003.

[2] M. Porter, "An algorithm for suffix stripping" Procedings of SIGIR. 1980.

[3] A Debbarma, M Nagamani, B Krishna, "Speech Recognition for Kokborok Language", International Conference on Biomedical Engineering and Assistive Technologies (BEATs 2010) 17-19 December, 2010.

[4] Kumud Kundu Chowdhury, "Kokborok the promising language of North East", Tripura, India

[5] A Debbarma, "Kokborok Stemmer and its application in spell checking", International Conference ICFC2012, Bangalore (Accepted)

[6] LTRC, IIIT Hyderabad, http://ltrc.iiit.ac.in