Morphological Analyzer for Kokborok

Khumbar Debbarma[1] Braja Gopal Patra[2] Dipankar Das[3] Sivaji Bandyopadhyay[2]
(1) TRIPURA INSTITUTE OF TECHNOLOGY, Agartala, India
(2) JADAVPUR UNIVERSITY, Kolkata, India
(3) NATIONAL INSTITUTE_OF TECHNOLOGY, Meghalaya, India
khum_10jan@yahoo.co.in, brajagopal.cse@gmail.com,
dipankar.dipnil2005@gmail.com,sivaji_cse_ju@yahoo.com

ABSTRACT

Morphological analysis is concerned with retrieving the syntactic and morphological properties or the meaning of a morphologically complex word. Morphological analysis retrieves the grammatical features and properties of an inflected word. However, this paper introduces the design and implementation of a Morphological Analyzer for Kokborok, a resource constrained and less computerized Indian language. A database driven affix stripping algorithm has been used to design the Morphological Analyzer. It analyzes the Kokborok word forms and produces several grammatical information associated with the words. The Morphological Analyzer for Kokborok has been tested on 56732 Kokborok words; thereby an accuracy of 80% has been obtained on a manual check.

KEYWORDS : Morphology Analyzer, Kokborok, Dictionary, Stemmer, Prefix, Suffix.

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 41–52,
COLING 2012, Mumbai, December 2012.

1 Introduction

Kokborok is the native language of Tripura and is also spoken in the neighboring states like Assam, Manipur, Mizoram as well as the countries like Bangladesh, Myanmar etc., comprising of more than 2.5 millions[1] of people. Kokborok belongs to the Tibeto-Burman (TB) language falling under the Sino language family of East Asia and South East Asia[2]. Kokborok shares the genetic features of TB languages that include phonemic tone, widespread stem homophony, subject-object-verb (SOV) word order, agglutinative verb morphology, verb derivational suffixes originating from the semantic bleaching of verbs, duplication or elaboration. Kokborok is written in the script similar to Roman script.

In general, morphological analysis is the first step to analyze the source language whereas morphology is the field of linguistics that studies the structure of words. The Morphological Analyzer takes one word at a time and produces its structure, syntactic and morphological properties or sometimes the meaning of a morphologically complex word (Dhanalakshmi et al., 2009). The morphological structure of an agglutinative language is unique and capturing its complexity using machines and generate in presentable format is a challenging job. Various approaches are used for building morphological analyzers such as Brute force method, root driven approach, affix stripping etc. (Rajeev et al., 2007; Parakh and Rajesha, 2011).

However, a morphological analyzer is an essential and basic tool for building any language processing application for a natural language e.g., Machine Translation system. Morphological Analyzers are essential technologies for most text analysis applications like Information Retrieval (IR) and Summarization etc. The most obvious applications are found in the areas of lexicography and computational linguistics.

For example, with respect to the word "dogs", we can say that the "dog" is the root form, and‘s’ is the affix. Here the affix gives the number information of the root word. Thus, morphological analysis is found to be centered on the analysis and generation of the word forms. It deals with the internal structure of the words and how those words can be formed. Morphology also plays an important role in applications such as spell checking, electronic dictionary interfacing and information retrieving systems, where it is important that words that are only morphological variants of each other are identified and treated similarly. In natural language processing (NLP) and especially in machine translation (MT) systems, we need to identify words in texts in order to determine their syntactic and semantic properties (Parakh and Rajesha, 2011). Morphological study helps us by providing rules for analyzing the structure and formation of the words.

Several Morphological Analyzers have been developed in different languages using both rule based and statistical methods. Moreover, different approaches to Morphanalyzer for English have already been developed such as in (Minnen et al., 2001). On the other hand, many Morphological Analyzers for Indian Languages have also been developed such as in Hindi (Goyal and Lehal, 2008), Bengali (Das and Bandyopadhyay, 2010), Malayalam (Rajeev et al., 2007), Manipuri (Choudhury et al., 2004; Singh and Bandyopadhyay, 2005) and for four of the languages, viz., Assamese, Bengali, Bodo and Oriya (Parakh and Rajesha, 2011). Manipuri is quite similar to Kokborok as it falls under Sino language family and an accuracy of 75% was achieved in (Choudhury et al., 2004). To the best of our knowledge, no previous work has been done on

[1] http://tripura.nic.in/
[2] http://en.wikipedia.org/wiki/Kokborok

developing Morphological Analyzer for Kokborok language, though a stemmer has been developed for Kokborok language (Patra et al., 2011) and its reported average accuracy is 82.9%.

This paper focuses on designing of a database driven affix stripping based Morphological Analyzer in Kokborok. In general, the Kokborok words have complex agglutinative structures. In the present work, a Kokborok morphological analyzer has been developed to analyze the input Kokborok sentence and for each surface level word to produce the root word(s) and associated information like lexical category of the roots, the prefix and/or the suffix using three dictionaries namely root, prefix and suffix. The morphological analyzer uses the Kokborok root words and their associated information, e.g., part of speech information, category of the verbal bound root (action/ dynamic, static) from the Kokborok to English bilingual root dictionary.

The rest of the paper is organized in the following manner. Section 2 provides details of Kokborok word morphology whereas Section 3 provides an elaborative description of the Morphological Analyzer. Next, Section 4 describes the implementation of Morphological Analyzer while Section 5 presents the results and analysis. Finally, conclusions and future directions have been presented.

2 Kokborok Word Morphology

Kokborok language is highly agglutinative and rich in morphology. The verb morphology is more complex compared to noun morphology. Kokborok words can be easily formed by affixations.

2.1 Verb Morphology

Most verbs have a monosyllabic root, and the main method for processing verb phrases is to add suffixes to the root. Kokborok verbs always occur in bound form to which multiple affixes are added to give the tense, manner of action. The suffixes can be classified in three layers at least (Jacquesson, 2008):

- The immediate layer, just after the root, concerns for instance locative markers: the action may reach far away, or go from up to down etc.
- The medium layer after the locative information concerns actancy: this is the kingdom of factitive, passives, reciprocals etc.
- The outer layer is the so called Tense Aspect Modality (TAM) area, where indications of Tense, Aspect and Mode are given.

2.1.1 Inflectional Morphology

Inflectional morphology derives words from another word from acquiring certain grammatical features but maintaining the same part of speech or category. There are a number of inflectional suffixes indicating tense of the verb of a sentence. Inflectional morphology is more productive than derivational morphology. First, inflectional morphology is paradigmatic, i.e., every Kokborok verb exhibits a paradigm with each inflectional marker as illustrated in the Table 1.

Inflectional affix	Inflection type	Verb	English meaning
O	Aorist	Chaho	eats
Anə	Future	Chahanə	Will eat
Na	Verbal noun	Chahna	eat

TABLE 1 – Inflectional Paradigm of Verb Chah.

2.1.2 Derivational Morphology

The derivational morphology can be divided into three different levels viz., first level derivation, second level derivation and third level derivation as given in Tables 2, 3 and 4. There are non-category changing derivational suffixes and category changing derivational suffixes.

Suffix	Meaning	Use and Meaning
-sa-	Upwards <Up>	Look up
-khlai-	Downwards<Dw>	Look down
-laŋ-	Away from speaker<Lat>	take away
-gra-	First in order<Pri>	Go first

TABLE 2 – First Level Derivation.

Suffix	Meaning	Use and Meaning
-sa-	Upwards <Up>	Look up
-khlai-	Downwards<Dw>	Look down
-laŋ-	Away from speaker<Lat>	take away

TABLE 3 – Second Level Derivation.

Suffix	Meaning/ feature	Use and meaning
-o-	Aorist<Aor>	Go
-anə-	Near future<Ftp>	Will go(may be next year)
-nai-	Future<Fut>	Will go (on the verge of going)
-kha-	Past<Pf>	Went
-li-ja	Negative <Pf>Neg	Did not go
-kho-	Still<Pf>	Still
-ja-	Negative	Not go
-glak-	Negative	Will not be going

TABLE 4 – Third Level Derivation.

2.2 Noun Morphology

Monosyllabic nouns are relatively rare in Kokborok where bisyllabic formations are dominant. This is due to the widespread process of compounding, either true compounding when two lexical roots form a new word.

2.2.1 Inflectional Morphology

The sole nominal inflectional category is case marking. The category is highly productive, both formally and semantically. The following Table 5 shows the paradigmatic nature of case marking.

Inflectional Affix	Type	Surface Words
-ni	Genitive & ablative	*Nokni,musukni* 'from house,cows'
-no	Accusative & dative	*Chwngno ,bono* 'us,to him'
-o	Locative e& illative	*Or-o ,bisiŋ-o* 'here, inside'

TABLE 5 – Nature of Case Marking.

2.2.2 Derivational Morphology

Derivational morphology is not productive in that there are apparently arbitrary restrictions on which suffixes may occur with the different categories of nouns. There are no categories of gender and number in Kokborok. No accord of any kind on this respect. Gender is marked as number only when needed not when items have to be feminine or masculine, singular and plural.

- **Gender:** the male role is marked by suffix *–la* in *tokla* (cock). It is unlikely that *joŋgla (*frog*)* can be explained by this suffix. The suffix *–jək* denotes the feminine gender as in *sajək* (daughter).
- **Size:** However the suffixes *–ma* and *–sa* forms an antonymic couple specialized in big and small.since *–sa* also means offspring or young as in *toksa*(chicken), *təima*(river), *təisa*(stream).
- **Number:** plural is in *–rok* and mostly for animates and also used in pronouns. For example *cherai* (child), *cherairok* (children).

3 The Morphological Analyzer

The proposed architecture of Morphological Analyzer is shown in Figure 1. The Morphological Analyzer is composed mainly of three modules: Tokenizer, Stemmer and Morphological Analyzer.

- **Tokenizer**: This module breaks the Kokborok sentence in to its constituent words or tokens for analysis.
- **Stemmer:** it strips the word in to root and affixes.
- **Morphological analyzer**: it analyzes various types of word structures and combines the features associated with the affixes and tags the word.

3.1 Dictionary Development

3.1.1 Affix Dictionary

Altogether 91 affixes are there out of which 76 are suffixes and 19 are prefixes. The various affixes are associated with words belonging to different part of speech to give resultant words with a particular meaning. The statistics of affixes are given in Tables 6 and Table 7.

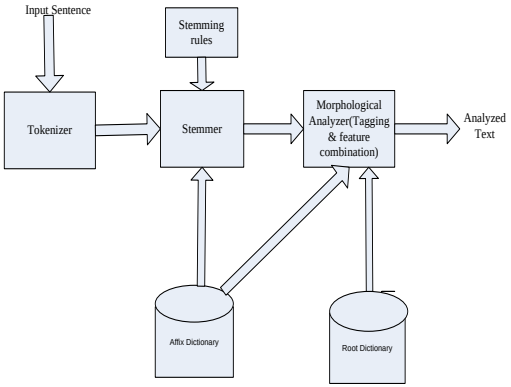

FIGURE 1 –Architecture of Morphological Analyzer for Kokborok

Type	No. of prefix	No. of suffix
Derivational	9	15
Inflectional	10	61

TABLE 6 – Statistics of Prefix types.

Lcat	No. of prefix	No. of Suffix
Noun	13	10
Verb	6	44
Adjective	0	22

TABLE 7 – Statistics of Suffix types.

Table 8 shows the affix dictionary entries where lcat, pers and emph are the features associated with each affix.

Kokborok_prefix	Icat	pers	Emph
Masema	Verb	2	Emph

TABLE 8 – Affix Dictionary Entries.

3.1.2 Root Dictionary

Altogether 2000 Kokborok root words have been collected, digitized and stored along with the associated information in root dictionary and the dictionary stores attribute of each Kokborok root words such as the part of speech (POS) and its English meaning. Root Dictionary has been developed by stemming the corpus collected from the Bible and the story books and the stemmed data are checked manually. Then we have assigned the POS and the English meaning manually.

The stemming algorithm used for the development of root dictionary is given below and Table 9 shows the root dictionary entries.

Kokborok	POS	English
Achuk	Verb	Sit

TABLE 9 – Root Dictionary Entries.

3.2 Stemming algorithm for generating Root Words

The algorithm is designed to remove both multiple suffixes as well as prefixes from the inflected words. It has been observed that the boundary of root words in Kokborok change after addition of suffixes (Patra et al., 2012). Thus we have added some rules in the algorithm as boundary changes after addition of suffixes. The algorithm is given below.

3.2.1 Prefix Stripping Algorithm

1. repeat the step 2 until all the prefixes are removed
2. read the prefix,
   if matched then store it in array and decrease the length of string
   else read another prefix.
3. If length of string >2 then go for suffix stripping, else exit

3.2.2 Suffix Stripping Algorithm

1. repeat the step 2 until all the suffixes are removed
2. read the largest suffix,
   if matched then check for rules.
   then store it in array and decrease the length of string
   else read another suffix.
3. exit.

We have achieved an accuracy of 85.5% by maximum suffix striping algorithm, where we striped maximum suffix first. There is no case of under stemming seen as we striped largest suffix first. The overall statistics of accuracies on major categories like verb, noun, adjective and adverbs are given in Table 10. In this case out of the total error, there are 69.3% mis-stemming and 30.7% over-stemming. For example,
Over- stemming: *sumano*(input) →*suma+no*(output)
Desired output: *suman+o*
Mis-stemming: *tongo*(input)= *tonk +o* (output)
Desired output: *tong+o*

Categories	Accuracy
Noun	72%
Verb	79%
Adjective	87%
Adverbs	96%

Table 11 – Results of Morphological Analyzer.

We have observed more number of errors in case of proper nouns or nouns, because the occurrence of proper nouns or nouns is more in the sentences. Some words have alphabet pairs similar to the affixes leading to over stemming for example:
Mis-stemming: *Kothmano → Kothman+o* (after stemming which is incorrect)
Desired output: *Kothma+no*
In case of verbs, error occurs due to the order of stemming of word for example
Mis-stemming: *Malwi → ma+lwi*(error if prefix stemmed first)
Desired output: *Mal+wi*(correct)
There are less number of errors in Adjectives and those are due to presence of alphabets similar to affix in the word. For example *bw+rwichwk* (error since bw is in prefix list), but "bwrwichwk" is a single words. Numbers of errors in case of adverbs are negligible.

4 Implementation Details

The Morphological Analyzer has been used in the present work for identification of word class features and sentence type. Broadly, there are five types of words that can be handled by this Morphological Analyzer.

Free words are formed without any affixation or compounding. E.g. Borok (people). Words with multiple prefixes and suffixes, such word occurs in the pattern given below. Where P, RW, S stands for prefix, root word and suffix respectively.

P + RW	For e.g., ***bu****pha* (my father)
RW+S	For e.g., *Khumbar****no*** (to Khumbar)
P+RW+S.	For e.g. ***Bu****kumui****ni*** (His/Her Brother In Law"s)
P+RW+S+S…	For e.g., *Ma(P)+thang (to go)+lai(S)+nai(S)* →***Ma****thang****lainai***(need to go)
RW+RW…	For e.g., *Khwn (Flower)+Lwng(Garden)* →*Khwmlwng*(Flowergarden)
RW+S+RW+S.	For e.g., *Hui(RW)(to hide)+jak(S)+hui(RW)+jak(S)+wi(S)* → *Hujakhujakwi* (Without Being Seen)

Compound words are formed by compounding among various words belonging to different part of speech. The various pattern of compounding is given below.

Noun + Noun
Noun + Adjective
Pronoun + Noun + Noun

4.1 Algorithm for implementing the Morphological Analyzer

1. Give input Kokborok sentences to the tokenizer module.
2. Tokenize each sentence to words
3. Repeat from step 4 to 7 until each word is analysed
4. Check if the word is in dictionary
5. If match not found stem the word in to root and affixes
6. Check for the pattern given above with the help of root and affix dictionary. If match is found apply rules, combine grammatical features and tag the word. If not found then send the word to complex word handler.

7. Complex word handler will stem the word with the help of root dictionary. If match found then will tag accordingly, otherwise tag it a unknown words.
8. Exit.

The flow chart of the Morphological Analyzer is shown in Figure. 2.

5 Evaluation

In the segmentation of words, we tested two methods: (i) First affix isolation, then detection of root and (ii) First detection of root, then isolation of affixes. In the former case there is overhead due to repeated access to the root dictionary. On the other hand, the later approach needs a single pass in the root dictionary. The first approach handles the orthographic complexity well and the second strategy is much faster in comparison with the former. The Morphological Analyzer has been tested using the corpus. The unanalysed words have been tagged by the analyzer as unknown (unk) and after manual check it has been found out that maximum number of unknown words belong to proper noun, thus later on it was tagged as NNP. The Morph Analyzer was tested again and it has seen that it is giving a better result. Errors have been calculated on the basis of words wrongly tagged, unanalyzed words and words tagged as unk out of total input words. Correctness of Kokborok Morphological Analyzer is shown in Table 11. There are some unknown words which could not be analyzed based on rules available and due to unavailability of root word dictionary, are effectively reducing the performance of Morphology Analyzer. The words in Kokborok can be easily formed by affixations and compounding, so the number of unknown words are relatively large. The accuracy of the Morphanalyzer can be further improved by introducing more numbers of linguistic rules and adding more root words to the dictionary.

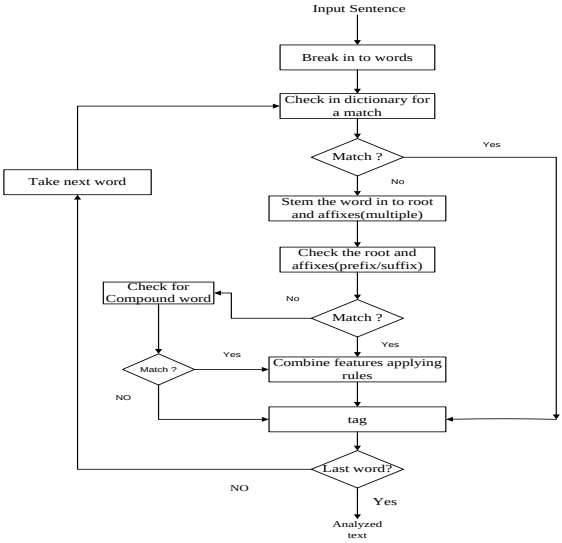

FIGURE 2 – Flowchart of Kokborok Morphological Analyzer.

	Based on unanalyzed words tagged as UNK	After UNK words tagged as NNP
Total input words	56732	56732
Analyzed words	42549	45386
Unanalyzed words	14183	11346
%age of analysed words	75%	80%
%age of errors	25%	20%

TABLE 11 – Results of Morphological Analyzer.

Conclusion and Future Work

In the present work, the development of a Kokborok Morphological Analyzer has been described. The analyzer uses three dictionaries of morphemes viz., root, prefix and suffix. The root dictionary stores the related information of the corresponding roots. The stemmer performs with an accuracy of 85.5% considering the inflectional and derivational suffixes. The Analyzer can classify the word classes and sentence types based on the affix information. In Kokborok, word category is not so distinct except Noun. The verbs are under bound category. The verb morphology is more complex than that of noun. The distinction between the noun class and verb classes is relatively clear; the distinction between nouns and adjectives is often vague. Thus, the assumption made for word categories depend upon the root category and affix information. Currently, we use a sequential search of a stem from the root dictionary because of its smaller size. Further a part of root may also be a prefix which leads to wrong tagging. In the stripping of the morphemes the various morphemes pattern combinations are tested. The morphology driven Kokborok POS tagging is very much dependent on the morphological analysis and lexical rules of each category.

The Natural Language Processing tools need more text corpus with better transfer rules and techniques to achieve quality output. The performance of the various Kokborok NLP tools that have been developed in the present work need to be improved by experimenting with various machine learning approaches with more training data. Future works include the developments of automatic Morphological Analyzer using some machine learning algorithms. The exploration and identification of additional linguistics factors that can be incorporated into the Morphological Analyzer to improve the performance is an important future task.

References

Choudhury, S., Singh, L., Borgohain, S. and Das, P. (2004). Morphological analyzer for Manipuri: Design and Implementation. *Applied Computing*, 123-129.

Das, A. and Bandyopadhyay, S. (2010). Morphological Stemming Cluster Identification for Bangla. *Knowledge Sharing Event-1: Task*, *3*.

Debbarma, Binoy and Debbarma, Bijesh (2001). Kokborok Terminology P-I, II, III, English-Kokborok-Bengali. Language Wing, Education Dept., TTAADC, Khumulwng, Tripura.

Debbarma, K., Patra, B. G., Debbarma, S., Kumari, L. and Purkayastha, B. S. (2012). Morphological analysis of Kokborok for universal networking language dictionary. In

Proceedings of 1st International Conference on Recent Advances in Information Technology (RAIT), pages 474-477. IEEE.

Dhanalakshmi, V., Kumar, M. A., Rekha, R. U., Kumar, C. A., Soman, K. P. and Rajendran, S. (2009). Morphological Analyzer for Agglutinative Languages Using Machine Learning Approaches. In *Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom'09)*, pages 433-435. IEEE.

Goyal, V. and Lehal, G. S. (2008). Hindi Morphological Analyzer and Generator. In *Proceedings of First International Conference on Emerging Trends in Engineering and Technology (ICETET'08).* pages 1156-1159. IEEE.

Jacquesson, F. (2008). A Kokborok Grammar. Published by Kokborok tei Hukumu Mission.

Minnen, G., Carroll, J. and Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering, 7*(3), 207-223.

Parakh, M. and Rajesha, N. (2011). Developing Morphological Analyzers for Four Indian Languages Using A Rule Based Affix Stripping Approach. In *Proceedings of Linguistic Data Consortium for Indian Languages, CIIL, Mysore*.

Patra, B. G., Debbarma, K., Debabarma, S., Das, D., Das, A. and Bandyopadhyay, S. (2012). A light Weight Stemmer for Kokborok. In *Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012)*, pages 318-325, Yuan Ze University, Chung-Li, Taiwan.

Rajeev, R. R., Rajendran, N. and Sherly, E. (2007). A Suffix Stripping Based Morph Analyzer for Malayalam Language. *Morph Analyzer Science Congress*.

Singh, T. D. and Bandyopadhyay, S. (2005). Manipuri morphological analyzer. In *Proceedings of the Platinum Jubilee International Conference of LSI*. University of Hyderabad, India.