ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 10, April 2015

Morphological Analyzer in the Development of Bilingual Dictionary (Kokborok-English) - An Analysis for Appropriate Method and Approach

Partha Sarkar, Dr. Bipul Syam Purkayastha

Abstract: ***The genre of Natural Language Processing is a widespread area which is confronted with many challenges and consecutively it produces new positive dimensions in the fields of computer applications or machine translation. Now a days, the particular area, machine translation, is becoming popular because of its practical uses and applications for various purposes. However, machine translation is very much related with the linguistic issues like morphology, syntax, semantics etc. This paper is particularly focused on the development of Kokborok bilingual dictionary. However, when we tend to discuss the particular area i.e., the development of bilingual dictionary, the area of morphology and morphological analysis become very crucial. This paper is aimed at discussing and analyzing the morphology of Kokborok language, the various NLP methods and techniques related with morphological analyzer to develop a Kokborok bilingual dictionary.***

Keywords: **bilingual dictionary, Kokborok, machine translation, morphology, morphological analyzer, syntax.**

I. INTRODUCTION

In the genre of Natural Language Processing, the field of Morphology in general and Morphological Analyzer in particular always play an important role. If we look at the development in the area of Natural Language Processing, we will find that machine translation always gets confronted with many challenges like problems relating to grammatical ambiguity, word analysis, parts of speech transformations, syntactical problems, bilingual dictionaries etc. However, when we tend to discuss the particular area i.e. the development of Bilingual dictionary, the area of Morphology and Morphological Analysis become very crucial.But before we delve deep into the analysis of the areas 'Bilingual Dictionary'and 'Morphological Analyzer', let us have a bird's eye view on Kokborok language.

II. A BRIEF OVERVIEW OF KOKBOROK LANGUAGE

Kokborok language is regarded as the native language spoken by the Borok people belonging to the state of Tripura. The term `Kok-borok` is actually a compound word consisting of two different words; the word 'kok'literally means language, and the word 'borok'means nation. Here, the second word is used to denote the Borok people who belong to Tripura. In simple words, it can be said that Kokborok literally means `the native language of the Borok people`. Kokborok language belongs to the Tibeto-Burman language family that is a sub-group of the Sino-Tibetan language group. It is the chief language family of East Asian and South East Asian regions. Kokborok language is closely associated with the Bodo language as well as the Dimasa language that are hugely spoken in the state of Assam. Further, Kokborok is also related to Garo language, which is principally spoken in the neighboring country that is Bangladesh and Meghalaya.

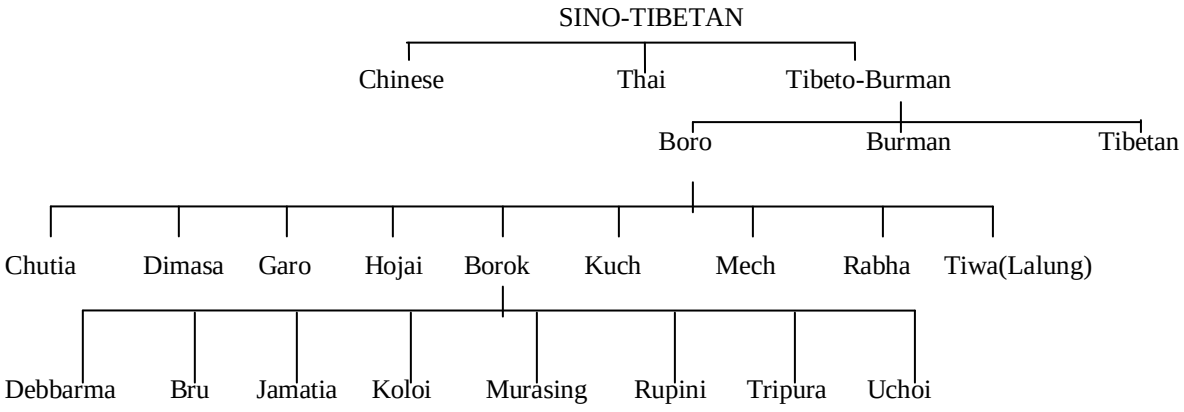


Fig: 1 Sino-Tibetan language group

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 10, April 2015

Kokborok language has been present in several forms. It is said that Kokborok language has 36 dialects. The Borok nation actually comprises of several communities as well as sub-communities in Tripura, Mizoram and Assam. They have their own dialects namely Puran, Tripura, Reang, Jamatia, Noatia, Murasing, Ulsoi (also called Usoi), Kalai and Rupini which differ from one another. The Kokborok script was known as `*Koloma*`. The history of the Borok Kings was originally written down in Kokborok language by using the Koloma script in a book called *Rajratnakar*. Later, two Brahmins, Sukreswar and Vaneswar translated it into Sanskrit and then again translated the chronicle into Bengali in the 14th century. However, the chronicle of Twipra in Kokborok and Rajratnakar are no longer available. Kokborok was relegated to a common people's dialect during the rule of the Borok Kings in the Kingdom of Twipra from the period of the 14th century till the 20th century. In the year 1979, Kokborok language was given the status of the official language of Tripura. Kokborok shares the genetic features of TB languages that include phonemic tone, widespread stem homophony, subject-object- verb (SOV) word order, agglutinative verb morphology, and verb derivational suffixes originating from the semantic bleaching of verbs, duplication or elaboration. With the passage of time, Kokborok has come in contact with Bangla, Arabic, Persian, English languages, and has taken many words.The Kokborok words have complex agglutinative structures. Due to the socio-economic and language barrier, many of the native people do not access information conveniently. This paper aims at analyzing the possibility of different methods, as discussed below, to develop an appropriate Kokborok morphological analyzer to analyze the input Kokborok word and for each word to produce the root word(s) and identify the associated lexical categories of the root like 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) to develop the Kokborok to English bilingual dictionary.

III. BILINGUAL DICTIONARY

Now a day, in the area of machine translation, the development of bilingual dictionaries in many regional languages has become very popular. However, it is important here to mention the meaning of bilingual dictionary. The word 'bilingual' refers to the ability to use two languages equally and in this connection a bilingual dictionary refers to a dictionary giving equivalent words in two languages. Bilingual dictionaries are seldom diachronic and usually alphabetic in arrangement.A bilingual dictionary consists of an alphabetical list of words or expressions in one language (the source language) for which exact equivalents are given in another language (the target language). The purpose is to provide help to someone who understands one language but not the other. Thus, bilingual dictionaries are different from monolingual dictionary not only in number of languages but also in purpose. The proper role and aim of a bilingual dictionary is to help the user to work with foreign lexical means. In simple words, the bilingual dictionary should provide precise equivalents of particular items of the vocabulary of the source language in the target language. Nevertheless, it is not indispensable that a bilingual dictionary elaborate the semantic structure of the words in the source language. It is true that bilingual dictionaries are vital resources in many areas of natural language processing. In any machine translation system, the dictionaries are of immense importance. However, in the process of developing the bilingual dictionary, two aspects should be taken care of, their content and their organization. The content of the dictionaries must be adequate in both quantity and quality: that is, the vocabulary coverage must be extensive and appropriately selected and the translation outputs should be carefully chosen for a satisfactory outcome. The size and quality of dictionary limits the scope and coverage of a system, and the quality of translation that can be expected. The dictionary entries are based on lexical stems of specified category and proper monolingual analysis. MT systems are linked to electronic dictionaries. Such electronic dictionaries can be of immense help even if they are supplied or used without automatic translation of text.

IV. MORPHOLOGICAL ANALYZER

Morphological Analysis is a part of NLP research which studies the structure of words and word formation of a language. Words in a language can be divided into many small units. In morphology the smaller meaningful units are called morphemes. The problem of recognizing different morpheme in a word is known as morphological parsing or morphological analysis. For example in English language, the morphological analysis of the word 'Boys' is [Boys=Boy +s {Root: Boy (category- noun)} + {'s' (indefinite plural marker)}]. In the above example, the word "Boys" is a combination of two morphemes- 'Boy' and 's' (root word "Boy" and suffix "s"). If a morphological analyzer analyses a word, it should give both root word and affix added with it and some other information like tense, case marker, gender, number, person and other relevant morphological information etc. as output. There are two broad classes of morphemes: stems and affixes. The stem is the 'main' morpheme of the word, supplying the main meaning, while the affixes add 'additional' meanings of various kinds. Affixes are further divided into

- Prefixes: These are placed in front of thestem. Example: Unhappy =Un+ happy

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 10, April 2015

- Suffixes: These are added after the stem. Example: Boy= Boy + s
- Infixes: These are inserted in the middle ofthe word. English language has no true infixes.

A morphological analyzer, in view of the above context, is a program for analyzing the morphology of an input word; it detects morphemes of any text. In other words, a Morphological Analyzer is the computational implementation of human ability to analyze a language. The main task of a morphological analyzer is to translate word forms into a string that represents its morphological makeup, such as 'eats'-{eats: *(eat+V+3p+Sg)* a verb in the third person singular present tense}.The job of the analyzer is to output all tag sequences dependable onthe grammar and the input word. The convenient way is to build up morphological transducers so that the *input* (or domain) side remains as the analysis side, and the *output* (or range) side contains the word forms. The general format of the morphological analyzer is Word → stem/root + suffixes. In real life, morphological analyzers tend to provide much more detailed information than this. Thus, the function of Morphological Analyzer isto identify newly encountered words, to extract roots for comparison of content and to determine parts of speech.

V. INFLECTIONAL MORPHOLOGY

In morphology, the study of the affixation processes that distinguish the forms of words in certain grammatical categories is called inflectional morphology. In most languages, inflectional morphology marks relations such as person, number, case, gender, possession, tense, aspect, and mood, serving as essential grammatical glue holding the relationships in constructions together. Following below is a brief discussion on various types of inflectional affixes. Diversity of verb morphology in Kokborok is very significant. For example if we consider 'cha' as root word than after adding 'kha' we get 'chakha' which means work has been done in past. Similarly after adding 'witongo' means work is being done in present and by adding 'witongmani' means work was being done in the past. Kokborok has a very strong and structural inflectional morphology for its nouns based on case. Case of noun may be nominative ("chwla", man), accusative ("chwla-no", to the boy), and genitive ("chwla-ni", of the boy), locative ("nog-o", in/at house) and so on. Gender and number are also important for identifying proper categories of nouns. Number may be singular ("ri", clothe) or plural ("rirok", clothes) gender of nouns can be masculine ("takhuk", brother), feminine ("bukhuk", sister), common ("cherai", child) and neuter ("swikong", pen) etc. we consider six different types of nouns and show possible representation of their inflectional suffixes in the UNL format. Some examples of analysis of nouns are shown in Table-I, based on number.

Table I: Example of Verb Morphology

Person/tense	Verb as appears	Inflectional suffix
Present	chao/khaio	o
Present continuous	chawitongo/khaiwitongo	witongo
Past	chakha/khaikha	kha
Past continuous	chawitongmani/khaiwitongmani	witongmani
Past perfect	chamani/khaimani	mani

Table II: Examples of Noun Morphology

Number	Root word	Word as appears in a sentence
Singular	chwla (boy)	Chwlano(to the boy)

Table III: Examples of Adjective Morphology

Root word	Word as appears in a sentence	Inflectional prefix

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 10, April 2015

sok (to rot)	kosok (rotten)	ko

V. DERIVATIONAL MORPHOLOGY

Morphological derivation is the process of forming a new word on the basis of an existing word. It often involves the addition of a morpheme in the form of an affix which can be prefix, suffix or infix. Derivation stands in contrast to the process of inflection, which only produce grammatical variants of the same word. Following below is a brief model of the inputs of some well-known Kokborok words with suffixes:

Table IV: Sample Suffix List of Kokborok Language

Suffix	**Word**	**Word**
Brebre	Khabrebre	Twibrebre
Lolo	Peklolo	Phulolo
Lwlwk	khalwlwk	komolwlwk
Bai	kokbailambai	hoebai
Dudu	khadudu	phududu
Roro	buraroro	Romroro
Thothok	chathothok	khathothok

A. Addition of Prefixes in compound Kokborok words

In Kokborok, some compound words are formed from root word with the addition of prefixes. And the prefixes changes according to the person. We have taken root word "Chwi-Chu" (Grand Mother and Grand Father).

VI. METHODS OF MORPHOLOGICAL ANALYSIS

There are different methods for the morphological analysis of natural language processing. Following below is a brief description of the different methods commonly used in the area of morphological analysis.

A. Brute Force Method

Brute force method is developed from the genre of artificial intelligence and problem solving concepts in mathematics denoted as brute force search. Brute force stemmers mainly explore the relations between root forms and inflected forms through the presentation of a table. The table is looked up to find a matching inflection to stem a word. If a matching inflection is found, the associated root form is returned. Brute force approaches sometimes leg behind for the simple reason that in this approach no algorithm is applied that would more quickly meet the solution. The algorithm is accurate provided that the inflected form already exists in the database. Moreover, Brute force algorithms are initially very difficult to design given the immense amount of relations that must be initially stored to produce an acceptable level of accuracy. However, brute force algorithms are easy to improve in that decreasing the stemming error is only a matter of adding more relations to the table. Someone with only a minor experience in linguistics is capable of improving the algorithm, unlike the suffix stripping approaches which require a good knowledge. This term is also same as pattern matching algorithm in which word is matched word by word. The result depends on the matching of word. In this context, it is unrealistic to expect that all word forms can be found and manually recorded by human action alone. A considerable size of Kokborok dictionary, having the common words used in day-to-day life, has at least 10,000 words and to analyze this amount of words with their constituent affixes is a lengthy and time consuming process and there always remains a chance for manual errors.

B. Root Driven Method

In the root driven approach, the stem of the word is to be found in a lexicon before starting the morphological analysis. The major drawback of this approach is the lengthy searching process required to find the stem. In this process, the word itself and its subparts, which have been obtained by removing the letters one by one from the end of this word, are looked up in a lexicon to find all the possible stems. The real stem is discovered after the morphological analysis made by using these possible stems. Even though there are different search methods improving the performance like letter tree encoding, the examining of each subpart is obviously a very time

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 10, April 2015

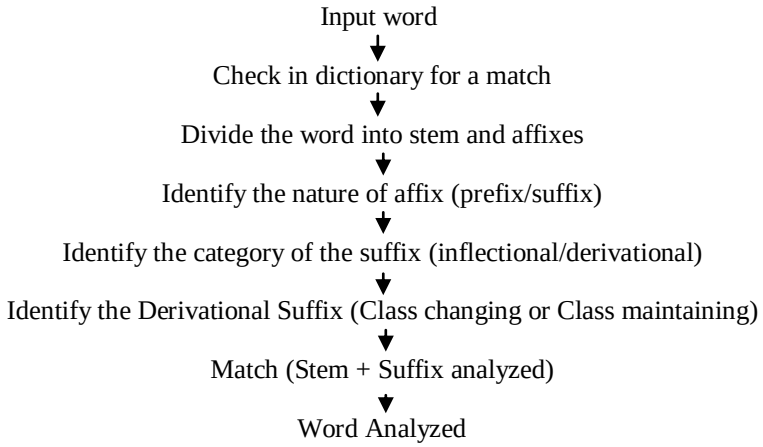
consuming process especially for the languages where the words can appear in very long forms. On the other hand, in the affix stripping approach, the searching process is relatively fast as the search is only done for affixes.

C. Suffix Stripping Method

This method aims at identifying individual suffixes from a series of suffixes attached to a stem/root, using morpheme sequencing rules. This approach is highly efficient in case of agglutinative languages. However, in languages that display tendency formorpho-phonemic changes during affixation (such as Dravidian languages), this method will require an additional component of morpho-phonemic rule besides the morpheme sequencing rules. The suffix stripping algorithm is a method of morphological analysis which makes use of a root/stem dictionary for identifying legitimate roots/stems, a list of suffixes, comprising of all possible suffixes that various categories can take, and the morpheme sequencing rules. This method is economical. Once the suffixes are identified, removing the suffixes and applying proper morpheme sequencing rules can obtain the stem. In order to identify the valid roots/stems, the dictionary of root/stem needs to be as exhaustive as possible. Considering this fact, the analyzer is designed to provide three types of outputs. The first one is obtained on the basis of a complete match of suffixes, rules and the existence of the analyzed stem/root in the root dictionary. The second one is obtained on the basis of either a matching of the suffixes and rules, even if the root/stem is not found in the dictionary or a matching of the suffixes, but not any supporting rule or existing root in the dictionary. At the end, there are some words which remain unanalyzed due to either absence of the suffix in the suffix list or due to the absence of the rule in the list. Keeping in view of all the methods discussed above, it is appropriate to use the suffix stripping method to analyze the Kokborok words in a bilingual Kokborok-English dictionary. The method is convenient. Even if the word is not found in the dictionary, still the analyzer can analyze the inflected form of the word into suffixes and stem.

D. Method Developed

Keeping in view of the above mentioned method, it can be assumed that the stem-suffix segmentation process can be worked out to develop a suitable morphological analyzer of Kokborok language. In the method developed for morphological analysis of Kokborok words through stem-suffix segmentation, the input word will be segmented first from stem and its consequent affix. After that, the affixes will be identified as prefixes or suffixes. In case of suffix, the suffix attached will be further analyzed to find out whether the suffix is inflectional or derivational. If the suffix is derivational, the derivational suffix segmentation will also indicate the class changing or class maintaining nature of the derivational suffix attached. After considering all these segmentation, it can be said that the word is analyzed into its consequents morphemes.



VII. CONCLUSION

The morphological structure of a language indicates what method to be followed among plethora of different methods. Kokborok language is highly agglutinative and rich in morphology. The verb morphology is more complex compared to noun morphology. Most verbs have a monosyllabic root, and the main method for processing verb phrases is to add suffixes to the root. There are a number of inflectional suffixes indicating tense of the verb of a sentence. As in case of nouns, there are no categories of gender and number in Kokborok. It is said that Kokborok, inflectional morphology is more productive than derivational morphology. In the stripping ofthe morphemes the various morphemes pattern combinations are tested. To conclude it can be said that experimenting and combining with other machine translation methods, Kokborok morphological analyzer could help in providing information about automatic spelling and grammar checking, parts of speech identification and many other promising areas.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 10, April 2015

VIII. FUTURE ENHANCEMENT

Kokborok as a language and the users of this language is growing day by day. Thus, the development of Kokborok bilingual dictionaries is getting prominence. However, in the development of Kokborok bilingual dictionary, structuring the entries of Kokborok morphemes for rule generation will be carried out. At the same time, the method and morphological rules for Kokborok prefix, its connection with root, verbal suffix and primary suffix will be experimented and analyzed in future. In connection, the availability of Kokborok infixes and the possibility of their inclusion in the machine translation will be analyzed.

REFERENCES

[1] Penelope Sibun, and Jeffry C Reynar, Language Identification: Examining the issues

[2] A. Goyal, "Named Entity Recognition for South Asian Languages Jan 2008," in Proceedings of the IJCNLP-08 Workshop on NER for South and South-East Asian Languages, Hyderabad, India.

[3] E. Antworth(1990) PC-KIMMO: A two-level processor for morphological analysis, Dalls, TX:Summer Institute of Linguistics.

[4] Daniel Jurafasky and James H Martin (2002) Speech and Language Processing-An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.

[5] GMinnen, J Carroll, D Pearce - Natural Language …, 2001 - Cambridge Univ Press

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

[7] E Loper, S Bird - … for teaching natural language processing and …, 2002 - dl.acm.org

[8] A. Pirkola. (2001). Dictionary – based cross- language information retrieval: Problems, methods and research findings. Information Retrieval.

[9] J. Mehrad. & M. Naseri. (2008). Natural language processing and information retrieval.

[10] M.F. Porter. (1980). An algorithm for suffix stripping.

[11] M. Collins, 2003 "Head-driven Statistical Models for Natural Language Parsing", Computational Linguistics.

[12] http://en.wikipedia.org/wiki/Kokborok_grammar

[13] http://www.ijcaonline.org/proceedings/icrtitcs2012/number9/10313-1448

[14] http://en.wikipedia.org/wiki/Morphology_%28linguistics%29

[15] http://www.aclweb.org/anthology/W12-5004.

[16] http://www.academia.edu/7096634/Morphological_Analysis_of_Kokborok_for_Universal_Networking_Language_Dictionary

[17] http://www.banglajol.info/index.php/DUJL/article/download/6902/5483.

AUTHOR'S PROFILE

Partha Sarkar received MCA degree from Bangalore University 2007.Currently he is an Assistant Professor in ICFAI University, Tripura and pursuing his PhD degree in Computer Science Department from Assam University, Silchar. His research interests include Artificial Intelligence, Natural Language Processing.

BipulSyamPurkayastha received PhD degree in Mathematics from North Eastern Hill University, Shillong in 1997.Currently he is a Professor in Computer Science Department in Assam University, Silchar. His research interests include Artificial Intelligence, Natural Language Processing.