A Rule Based Approach for Implementation of Bangla to English Translation

Md. Khalilur Rhaman,
Computer Science and Engineering Department
BRAC University, 66 Mohakhali,
Dhaka, Bangladesh.
e-mail: khalilur@bracu.ac.bd

Narzu Tarannum
Computer Science and Engineering Department
North South University,
Dhaka, Bangladesh.
e-mail: narzu.tarannum@gmail.com

Abstract*— An initiative to model a Bangla to English (B2E) translation using Natural Language Processing (NLP) was proposed in our previous research. Here we implemented the model with a lot of modifications. A very successful translator is observed in Anubadok Online which is based on Penn Treebank annotation system and it can only translate English sentences to Bangla. Penn Tree Bank is the collection of English corpus, so Bangle linguistic processing is not observed there. Bangla is an Irregular Language. In our previous research, we proposed a case structure analysis for verb. There are a lot of influences of case in Bangla language. The relationship between verb and case elements is an important issue for Bangla language. But in our current implementation we used rule based approach. For Bangla-English translation first we performed morphological analysis for Bangla then we used rule based analysis where we considered a limited feature of case analysis. After that, using a dictionary we translate bangle words into English. To make an English sentence, we considered English SVO grammatical rules. Our current system is successfully implemented for the translation of Assertive-Affirmative, Negative and Interrogative sentences.

Natural Language Translation, Case Structure Analysis, Morphological Analysis, Lexicon, Syntactic Analysis, Semantic Analysis, Transliterate.

I. Introduction

Considering the context of our country research on Natural Language Translation is very important. In this research we successfully could translate assertive form of Bangla sentence to English sentence. It was not so easy task. There are very limited numbers of researches [1] [2] [7] [8] [9] [11] on this field. Especially the processing of Bangla linguistic is totally ignored almost in all translators. Considering this fact our research group first proposed an Initiative of Bangla-English Natural Language Translation Using Case Structure at ICALIP2010 [20]. Due to the limitations of researches on Bangla NLP, we considered a number of Books [15], [16], [17], [18] and websites [6], [12], [13], [14]. English and Bengali are fairly different languages from each other. It will be necessary to make translation based on a kind of deep structures of sentences which determines sentential forms schematically in respective languages. Now a day, the importance of research on Natural Language Translation for Bangla Language is not ignorable. Machine translations for a number of Languages like Spanish-English, Japanese-English and French-English already have successfully done. Bubblefish translator, Excite Translator and even Google translator are the example of their success. So, we have selected this topic to contribute on our mother language. In this research we figured out and implemented a framework for rule based Bangla-English Natural Language Translator. This framework considered case analysis for Bangla language processing and a corpus for Bangla-English case frame mapping but the implementation is based on rule based approach. In this paper, we are going to discuss on the morphological analysis for Bangla then we will discuss the case analysis after that lexicon will be discussed in details. The interface, algorithms, result and comparison will be shown based on our implementation. We considered 60 inflections of verbs and the number of verb is 46. Our lexicons contain 1480 Nouns, almost all Pronouns, 650 Adjectives which can also be used as Adverb and 19 Prepositions. If it does not find any input word in the database then it directly transliterate considering it as a proper noun.

II. Proposed Architecture

A. Architecture

The architecture is depicted in Fig.-1. First the system receives Bangla text from the user then performs

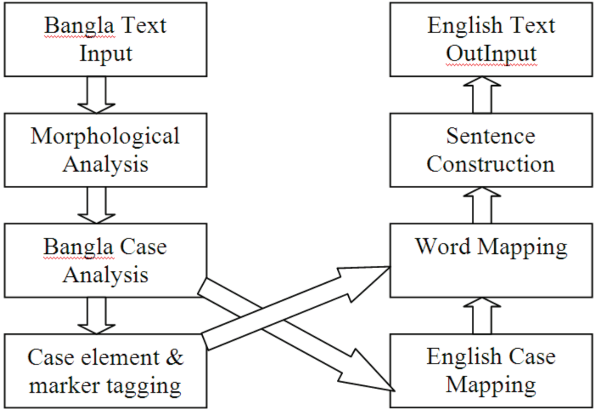


Figure: Architecture

978-0-7695-4959-0/13 $25.00 © 2013 IEEE
DOI 10.1109/ACSAT.2012.98

morphological analysis which will be discussed in section 2.B. After that we used rule based analysis where we considered the features of Bangla case. Case analysis is discussed in section 2.C and most of the features are used in our current implementation. Using a dictionary we translate Bangla case elements and markers into English word. On the other hand we mapped Bangla-Case to English SVO grammatical form. We proposed and implemented a structure for sentence construction.

B. *Morphological Analysis*

In our implementation, we extracted the prefix and suffix markers from the root word according to the lexicon of noun affix and verb markers. We did not considered feature unification but the proposed structure shown in Fig.-2 discussed in [1] can be easily implemented in our system. This structure shows the Part of Speech representation for each case.

C. *Case Analysis:*

Case Analysis is a relationship between a verb and constituents, which is often a noun or pronoun or an adverb, in a sentence is called case relation. A collection of such relations with a common verb is called case structure.

1) *Karta (Nominative case):*
Which corresponds to English's subjective case, indicates the subject of a finite verb. That which independently (unaidedly) does (performs) its own deed (function) is called the doer (agent or Karta).

2) *Karma (Accusative case):*
Whatever modification is obtained by the subject (the doer) is called its function (Karma), which corresponds to English's objective case, indicates the direct object of a verb.

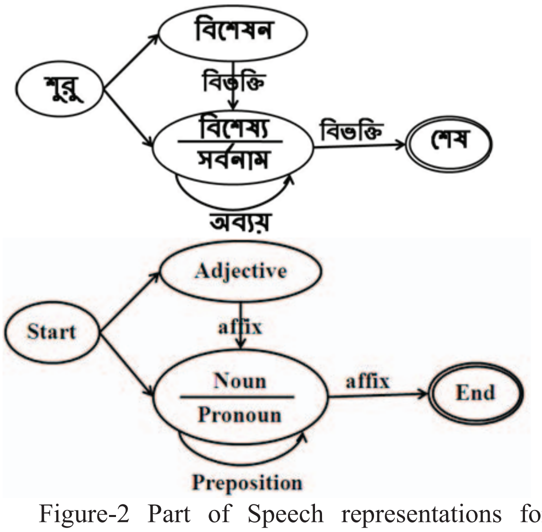


Figure-2 Part of Speech representations for “karak” (Case).

3) *Karan (Means of that deed):*
The substantial cause of that particular deed by which it is done or originated is called the means of that deed (the Karan).

4) *Sampradan: (The receiver of the action):*
That for which that particular deed is performed or done, is called the receiver (Sampradan karak).

5) *Apadan (Ablative case):*
The ablative case indicates movement from something, and/or cause. The permanent substance out of which that particular function or deed is done or obtained is called the (Apadan).

6) *Adhikaran (The base of the deed):*
The permanent cause and the same permanent substance is called the base of the deed (Adhikaran).

In order to process Bangla sentences to English translation, in this research we considered mainly four cases which is depicted in figure 3.

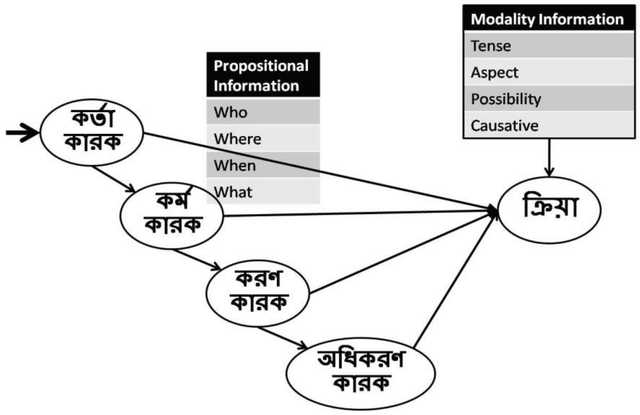
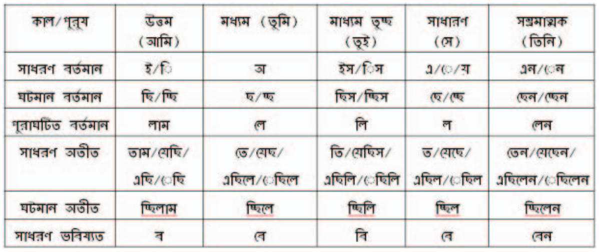


Figure-3 Bangla Case representation

D. *Verb Inflections*

This is the most important part of our dissertation. In our verb lexicon we will consider every property of a verb. A number of researches [2], [3], [4], [5] are considered to find out the layout of this structure. Each verb will be analyzed by a linguistic expert than it will be represented in lexicon. In Bangla it is observed that the entire cases are highly bounded with the verb by some defined case marker. We will handle these things with case analysis but for each verb there are some restrictions of categories. These restrictions will be mentioned here. The categorization is also a very important issue which was introduced from wordnet [10] concept. In our previous modeling we proposed a hierarchical representation of category. In current implementation, categorization is implemented very limitedly. In Bangla language it is possible to extract aspects, modality, person and tense information from the inflections of verb. So, our proposal was to include aspect and modality information inside this verb representation. In our current implementation we extract person and tense by using an external table which is shown in figure-4. Here we consider 60 inflections of a verb. It seems that this kind of analysis is impossible for a number of verbs. But there is a light at the end of this darkness is the number of verb root is not too much. It might be possible to process almost all sentences by analyzing only 300 root verbs in bangle vocabulary. The aspect and modality information are not so limited but very similar for Bangla verbs.



কাল/পুরুষ	উত্তম (আমি)	মধ্যম (তুমি)	মাধ্যম তুচ্ছ (তুই)	সাধারণ (সে)	সম্ভ্রমাত্মক (তিনি)
সাধরণ বর্তমান	ই/ি	অ	ইস/িস	এ/ে/য়	এন/েন
ঘটমান বর্তমান	ছি/চ্ছি	ছ/চ্ছ	ছিস/চ্ছিস	ছে/চ্ছে	ছেন/চ্ছেন
পুরাঘটিত বর্তমান	লাম	লে	লি	ল	লেন
সাধরণ অতীত	তাম/য়েছি/ এছি/েছি	তে/য়েছ/ এছিলে/েছিলে	তি/য়েছিস/ এছিলি/েছিলি	ত/য়েছে/ এছিল/েছিল	তেন/য়েছেন/ এছিলেন/েছিলেন
ঘটমান অতীত	চ্ছিলাম	চ্ছিলে	চ্ছিলি	চ্ছিল	চ্ছিলেন
সাধরণ ভবিষ্যত	ব	বে	বি	বে	বেন

Figure 4 Tense and person representation table.

Algorithm-1 shows the Verb Analysis. Propositional information is already analyzed in morphological analysis section. In this algorithm, it will just check the validity of category. Than find the tense, type and person information by using the forms of verb lexicon.

Algorithm-1: for finding Verb Inflection

1.	Read Verb text
2.	Open Verb Lexicon
3.	Compare verb stems
4.	IF not valid "ERROR MESSEGE" and Exit
5.	ELSE Open Verb_form Lexicon
6.	Compare Verb_marker with Verb_form[m][n]
7.	Tense= Verb_form[m][]
8.	Person= Verb_form[][n]
9.	IF number of Verb > 1 than Verb_type= Intransitive (False)
10.	ENDIF

E. *Syntactic and Semantic Analysis*

For Bangla language we considered first noun as subject. But if there exist a "r/র" as a suffix than we considered it as compound noun for example "ছাত্রীর বাবা". If there are more than one verb than we considered it as compound verb and the last verb as main verb "খেতে ভালবাসে". Our current version will consider rest of the nouns combined with adjective as object. For translation for compound noun, nouns will be concatenate with "'s" (Student's father). And for compound verb, verbs will be concatenate with "to" (likes to eat).

III. Implementation

A. *Implementation Platform*

We implemented the B2E-0.0.1 by using JAVA. We implemented entire algorithms and searches by only JAVA. Although it is easier to find data using SQL of database software but considering the efficiency factor we used JAVA. We saved the lexicon in .txt files. At the time of execution, we first read the lexicon from .txt files to memory using array variables then we did entire job from memory.

B. *Algorithms*

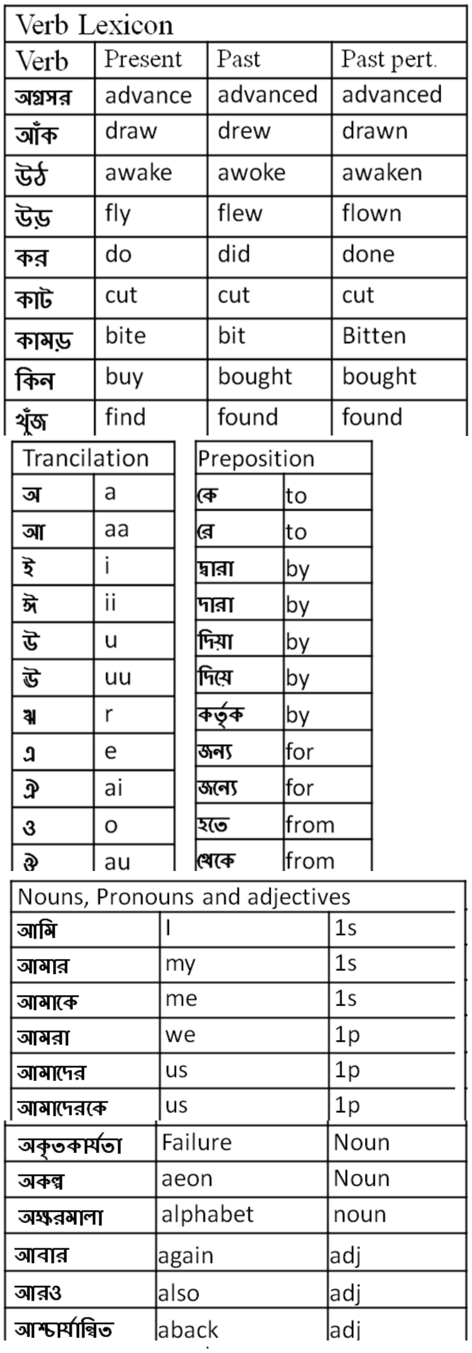
We could not implement exactly how we discussed in our previous model [20] due to some complexity. The whole translation Algorithm-2 is in bellow:

Algorithm-2: Translation Algorithm

1	Initialize Verbs, Nouns and Trancilation file
2	Read Input text and split into Input array word by word
3	FOR i=0, i<input word count, i++
4	Search ith input word in verb array
5	IF match
6	Verbs[]=attributes of that verbs
7	ELSE Search ith input word in verb array
8	Nouns[]=attributes of that verbs
9	ELSE Nouns[]= Transliteration of input word
10	i++;
11	END WHILE
12	IF verbs NOT FOUND Verbs[]="is"
13	FOR i=0, i<number of verb, i++
14	Search suffix of ith verb with inflections
15	Verbs-eng= translation of that verb(present, past, past participle) considering tense and person
16	END FOR
17	Compound Subject recognition
18	Translation=Subject
19	Compound Verb or intransitive verb recognition
20	Translation= Translation + Verb
21	FOR i= number first object word, i<number of last object word, i++
22	Translation= Translation + preposition + noun
23	ENDFOR
24	Print Translation

C. *Structural Design of Lexicon*

The structural design of lexicon is shown in figure5. For verb lexicon we mapped three forms for each English verb with a single Bangla verb. For trancilation and preposition we used direct mapping from Bangla to English bu. For noun, pronoun and adjective lexicon we directly mapped from Bangla to English but for noun we used 3 forms of English persons and singular and plural number which will be used in language generation.



Verb Lexicon			
Verb	Present	Past	Past pert.
অগ্রসর	advance	advanced	advanced
আঁক	draw	drew	drawn
উঠ	awake	awoke	awaken
উড়	fly	flew	flown
কর	do	did	done
কাট	cut	cut	cut
কামড়	bite	bit	Bitten
কিন	buy	bought	bought
খুঁজ	find	found	found

Trancilation	
অ	a
আ	aa
ই	i
ঈ	ii
উ	u
ঊ	uu
ঋ	r
এ	e
ঐ	ai
ও	o
ঔ	au

Preposition	
কে	to
রে	to
দ্বারা	by
দারা	by
দিয়া	by
দিয়ে	by
কর্তৃক	by
জন্য	for
জন্যে	for
হতে	from
থেকে	from

Nouns, Pronouns and adjectives		
আমি	I	1s
আমার	my	1s
আমাকে	me	1s
আমরা	we	1p
আমাদের	us	1p
আমাদেরকে	us	1p
অকৃতকার্যতা	Failure	Noun
অকল্প	aeon	Noun
অক্ষরমালা	alphabet	noun
আবার	again	adj
আরও	also	adj
আশ্চর্যান্বিত	aback	adj

Figure-5 Structural design of lexicons

D. *Bangla2English Mapping*

In current state we could not find any case structure for English that we can map from Bangla case frame. So, we considered SVO (subject verb object) structure for English. Where we will map "karta karak" to subject, "karma karak" to object and verb to verb. All other cases will be added at the end of the SVO with proper adjective and preposition. Once the categorization is completed in Bangla analysis, it would be easier to find the position with proper marker.

E. *Language Generation*

We used following DFA to generate English sentence that is depicted figure-6. The DFA starts with the subject and according to rule of possessive noun it will add "'s" then add another noun. After concatenating adverb it add verb. Concatenating proverb with verb like "going to write" DFA goes to add the object part. Object can be constructed using a number of nouns, pronouns and adjectives which is denoted by marker in the figure.

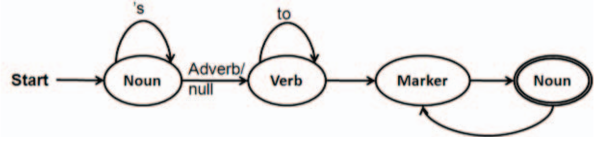
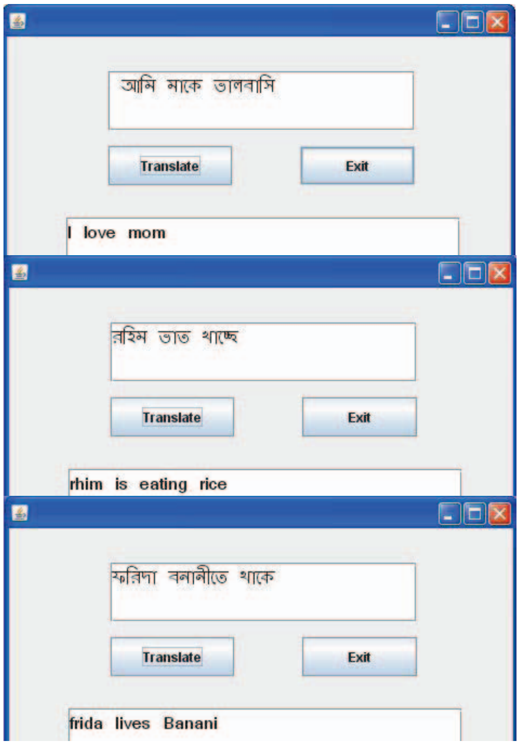


Figure-6 DFA of language generation

F. *Demo of Output*

Demo output is showing in figure-7. Where first two is showing correct output last two showing strange output.


আমি মাকে ভালবাসি

I love mom

রহিম ভাত খাচ্ছে

rhim is eating rice

ফরিদা বনানীতে থাকে

frida lives Banani

Figure-7 Demo output

IV. PERFORMANCE

Our current version considered 46 verbs till today. But we have a very good lexicon for Noun (1480), adjective and adverbs (650). Table-1 is showing the status table for our system. From these tables we can understand that our system is well implemented for assertive simple sentences and moderately implemented for compound sentence which was discussed in our theoretical model [20] because of time and linguistic challenges. Even our system can process compound noun but it consider only "r/র" by which more than one sentence combine together. So, for other case it will wrongly translate the sentence and also if any "র" exist after any proper noun then it will consider it as compound noun. The same thing will happen with compound verb. One more limitation is ambiguous word. Some word can be a noun or verb. We considered only 46 verbs so, it is not a big issue for our current system but in future when we will increase the number of verb would be a big challenge. Our system only can translate single sentence. But it is not a big challenge to overcome this limitation. We can easily split sentences by punctuation symbols. Translation for multiple words for a single word is still a challenge for us. Lastly, we wanted to implement our system using more emphasis on case structure but in implementation became rule based approach.

Status Table (Version: B2E-0.0.1)

	Subject	Verb	Verb Inflect.	Object	Assertive	Imper.	Interro.	Exclam.
Affirmative	W	W	W	W	W	N	N	N
Compound	M	M	W	M	M	N	N	N
Complex	N	N	M	N	N	N	N	N
Compound – Complex	N	N	M	N	N	N	N	N

Table-1
W: Well implemented
M: Moderately implemented
N: Not/Not-well implemented

V. Conclution

This is the first initiative to translate Bangla sentence considering case. Most of the works in Bangla language have done by using phrase structure analysis. And even a complete Bangla to English translator has not yet been found. So, the accuracy could not be compared with other system. According to our status table, our implemented system is well only for affirmative sentence but the scalability of our system is very high. This model supports the facility to include any feature of word or any kind of grammatical or non-grammatical information of a sentence. If we can successfully implement categorization/ classification, we can reduce the word sense ambiguity. In morphological analysis we considered single word-stem that only considered "র". But we have given the guide line to use compound structure. Categorization is playing a vital role in our model. Since a huge survey is necessary it can be included in future work. We need to include a linguistic professional for information verification and at least satisfactory number of verbs (200) is necessary to prove the efficiency of this system. For case frame mapping of verb, phrase structure grammar analysis and rules for language generation we need more support from both Bangla and English linguistic expert. We strongly believe that our model will be able to handle all of the linguistic challenges and this initiative will open a new era of research in Bangla language.

VI. Future Scope

The limitations are the scope of future development. The prime concern is to implementation and test the efficiency with the simple sentences considering at least 200 verbs. Then anyone can improve any part of the model. Here are some specific scope guidelines:

As we mentioned before more morphological analysis is necessary, that is very closely related with linguistic information. Improvement of case analysis would be an ongoing process. The inflexions of words can be analysis by finite automata for Bangla language. We have to standardize the category. By adding some affixes we can also improve the scalability that will increase the semantic efficiency. Vocabulary and Linguistic information improvement is also a continuous process. This kind of improvement will affect the accuracy because more information means more ambiguity. So, new challenges will be arise to improve the model. This system only considered assertive simple sentences. So, Affirmative, Imperative, Exclamatory, compound, complex, complex-compound and many other forms of sentences are left for future. Finally, it was the initiative only from Bangle to English. But a number of languages are waiting for translation. A huge impact of case is found in most of the Asian languages like Japanese, Hindi, Urdu etc. So, case analysis can open a new area of research in Natural Language Processing. The modification of Language is a continuous process. So, our research would have no perfect end.

References

[1] N. Khan and M. Khan, "Developing a Computational Grammar for Bengali Using the HPSG Formalism," Proc.International Conference on Computer and information Technology (ICCIT), 2006.

[2] Mahmud and M. Khan, "Building a Foundation of HPSG-based Treebank on Bangla Language," Proc. International Conference on Computer and information Technology (ICCIT), 2007.

[3] Hirosato Nomura, "Modeling and Representative Framework for Linguistic and Non-linguistic Knowledge in Natural Language Understanding," Proc. of Germany-Japan Science Seminar, 1986.

[4] Bruce, Bertram C., "Case Systems for Natural Language," Artificial Intelligence, vol.6, no.4, pp.327-360, 1975.

[5] Shimazu, S. Naito, H. Nomura, "Japanese Language Semantic Analyzer based on an Extended Case Frame Model," Proc. of International Joint Conference on Artificial Intelligence. 1983.

[6] Omicron lab: 5/5/2012 http://www.omicronlab.com/avro-keyboard.html

[7] Center for Research on Bangla Language Processing (CRBLP): 5/5/2012, http://crblp.bracu.ac.bd/

[8] S. Dusgupta and M. Khan, "Feature Unification for Morphological Parsing in Bangla", BRAC University Journal, 2010

[9] S. Dusgupta and M. Khan, “Morphological Parsing of Bangla Words Using PC-KIMMO,” BRAC University Journal, 2010

[10] Wordnet: 5/5/2012, http://wordnet.princeton.edu/

[11] English to Bangla Translator: 5/5/2012 http://anubadok.sourceforge.net/.

[12] E2B documentation: 5/5/2012, http://bengalinux.sourceforge.net/cgi-bin/anubadok/index.pl/.

[13] Ankur for free dictionary: 5/5/2012 http://www.bengalinux.org/english-to-bengali-dictionary/dumps/

[14] Ankur for free dictionary: 5/5/2012 http://www.bengalinux.org/english-to-bengali-dictionary/dumps/

[15] Book: Daniel Jurafsky & James H. Martin, “Speech and Language Processing” Upper Saddle River, N.J. : New Delhi : Prentice Hall ; 2000 Pearson Education.

[16] Book: Muhammad Enamul Haque, Shib Proshonno Lahori, Shorochish Sharkar, “Bangla Academi Baboharik Bangla Ovhidhan,” Bangla Academi, 2005

[17] Book: Mahbubul Alam, “Vhasha Shauravh Bakoron o Rachana,” Bangla Academi, 2007

[18] Book: P.C. Dash, “Applied English Grammar and Composition”

[19] Book: Abdul Khalek, Jobed Ali, “Bangla Bakoron o Rachana”, Globe Library Prv. Ltd, 2007.

[20] Narzu Tarannum, Md. Khalilur Rhaman, “An Initiative of Bangla-English Natural Language Translation using CASE,” proc. International Conference on Audio, Language and Image Processing, 2010.