COLLABORATIVE LEXICON DEVELOPMENT FOR BANGLA

Dewan Shahriar Hossain Pavel (pavel@bracuniversity.net), Asif Iqbal Sarkar (asif@bracuniversity.net)*, Faisal Muhammad Shah (faisal505@hotmail.com)**, Dr. Mumit Khan (mumit@bracuniversity.ac.bd)***

Research Programmer, BRAC University, Dhaka, Bangladesh,
Research Programmer, BRAC University, Dhaka, Bangladesh*, Lecturer, People's University Of Bangladesh, Dhaka, Bangladesh **,
Associate Professor, BRAC University, Dhaka, Bangladesh***

Abstract

*This paper addresses the issue of building a Bangla lexicon with a collaborative effort through stand alone application and web based interface. The words in the lexicon will be annotated with a combination of tags addressing Parts-of-speech, syntactic, semantic and other grammatical features. Bangla words have been classified into several different parts – of – speech categories including various major word groups and subgroups. This paper aims to provide an integrated user – friendly software interface to the user to annotate a large existing Bangla word set and proposes a mechanism to collaboratively integrate linguists and other interested people into the lexicon build up process***.** *The effort will be a significant progress towards development of a properly annotated lexicon. The outcome of the effort will significantly help in the processes of Morphological Analysis, Automatic grammar Extraction and machine translation for Bangla.*

Keywords

Corpus, POS (Parts-of-Speech), MVC (Model View controller), Lexical tags, Morphology.

INTRODUCTION

The development of language resources and its availability is a must for enhancing Language processing capabilities and research in this field. A lexicon is an essential language resource. It is the central repository of data for all language processing applications. It contains information for human consumption as well as computer programs. Bangla Lexicon can be used for several purposes including spell checkers and morphological analysis for Bangla language. A Bangla lexicon can be extracted systematically from a Bangla corpus [4] since it is considered a source of all words. But due to the unavailability of a complete Bangla corpus this process of automatic lexicon development [1] did not go too far. This paper proposes another process to manually build up a lexicon, which is essentially a list of all words in the language and tag the words sufficiently with features such as word meaning, Parts-Of-Speech (POS) and all other grammatical features. All these information need to be stored in a database and properly formatted before display to end users. About a hundred thousand Bangla words are already available in text format to start this process. The process itself is quite lengthy and also incomplete if done by a single person or group, because of the unavailability of tag data for Bangla. So the aim of the project is to formalize a procedure for a collaborative effort by different individuals or groups towards producing a tagged Bangla lexicon. This requires a POS tagging interface [2], both web based and stand alone that would provide a common platform for different contributors to enter tag information, semantic and other grammatical information that is available in a dictionary.

1. THE VALUE OF LEXICAL RESOURCES

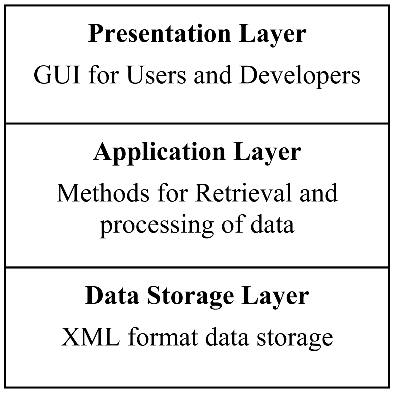
The value of the effort to create a Bangla lexicon is very significant. Developing realistic models of human language that support research and technology development in language related fields and all types of natural language processing activities require masses of linguistic data, lexicon being an important component. Lexicon development is a prerequisite for Electronic dictionaries which have become among the most indispensable language resources for those involved in all aspects of natural language processing (NLP)[3]. Three components are vital for the building of a morphological parser for any language. They are (1)Lexicon (2) Morphotactics (3) Orthographic Rules[5],

[8], [9]. There has been a significant development in the process of creating a lexicon in some European and North American countries in their native language and numerous corporate research groups are routinely engaged in medium- to large-scale lexicon development. But there has been no such recognizable attempt to create Bangla lexicon as such. The lexicon, which is a Published language resources benefit a broad spectrum of researchers, technology developers and their customers. The presence of community standard resources reduces duplication of effort, distributes production costs and removes a barrier to entry. As research communities mature, published resources are corrected, improved and further annotated. They provide a stable reference point for the comparison of different analytic approaches in language processing. There have been a few ongoing attempts in this area by some contemporary research groups based in Kolkata, India, but unfortunately in Bangladesh there has not been any fruitful progress in creating a digitized version of a fully annotated lexicon even after realizing the necessity of such a resource. Part-of-speech tagging is the base for further works of natural language processing (NLP), such as word-sense disambiguation, noun-phrase chunking and context-free grammar acquisition. This paper is just the beginning of our progress towards coming up with a strategy towards developing a Bangla lexicon with POS and other tag information.

2. METHODOLOGY

An important aspect in the usability of a Lexicon is its (technical) appearance. Many small, experimental systems have been built on stand-alone personal computers, using tools that were quite appropriate for the scientific project targets but not to support practical applications in a production environment. The Lexicon design as described in this paper is aimed at large-scale collaborative production, providing multi-user access and high performance on an appropriate platform. It can be integrated in external applications when used as a network server, and just as with traditional database systems, will be shielded from casual or non-technical / non-linguist users with appropriate front ends [6].

Bangla lexicon development can be described as the continuous interaction of three layers of functions. They are depicted below:



Presentation Layer

GUI for Users and Developers

Application Layer

Methods for Retrieval and processing of data

Data Storage Layer

XML format data storage

The lexicon development framework involves Model View Controller architecture. The MVC [7] paradigm is a way of breaking an application, or even just a piece of an application's interface, into three parts: the model, the view, and the controller.

Input-->Processing-->Output
Controller --> Model --> View

The user input, the modeling of the external world, and the visual feedback to the user are separated and processed by model. The controller interprets mouse and keyboard inputs from the user and maps these user actions into commands that are sent to the model. The model manages one or more data elements, responds to queries about its state, and responds to instructions to change state.

The model in the lexicon development project consists of the XML files containing lexicon data and the java API which will be used to process user input and output formatting. The lexicon development interface will have a web based view and a standalone view. Both the view will be used for entering data into the lexicon. The web based interface can be used by all people, so the data that is entered through the web interface will not be directly entered into the final lexicon file. Instead they will be stored into a primary XML file for further validation for correctness and redundancy by specialists. After correction the relevant data would be stored into the final XML lexicon file and hence the lexicon will be updated. The intermediate web technology (JSP and Servlet) will be controlling the interaction of the model and view components. The standalone interface is a form based interface that allows a user to enter necessary data directly into a XML file and hence will be used by linguists and specialists. The web based interface consists of a HTML form with the same fields as the standalone interface.

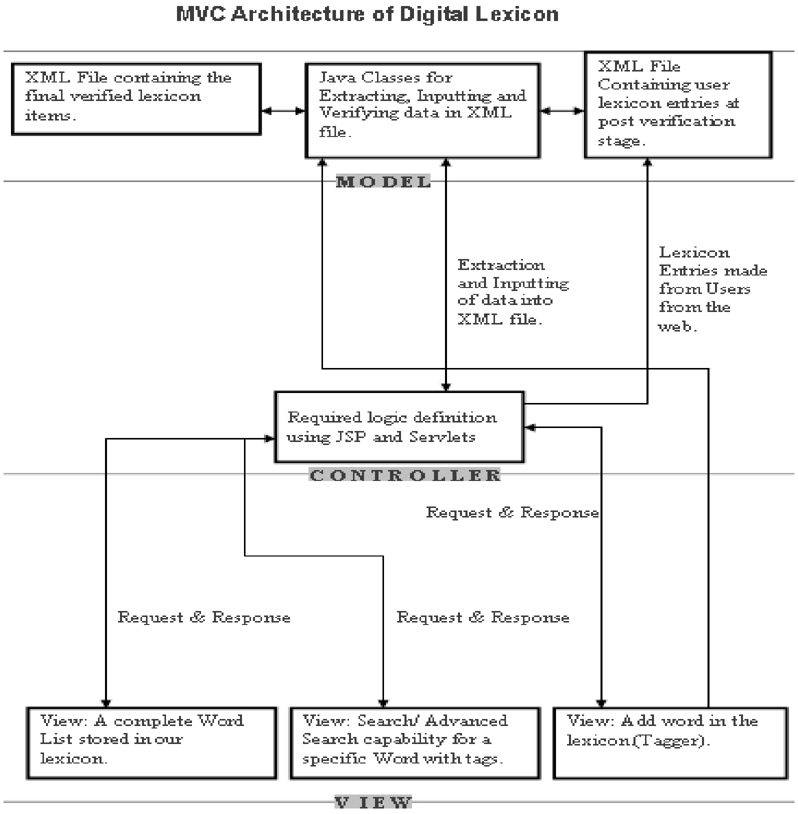
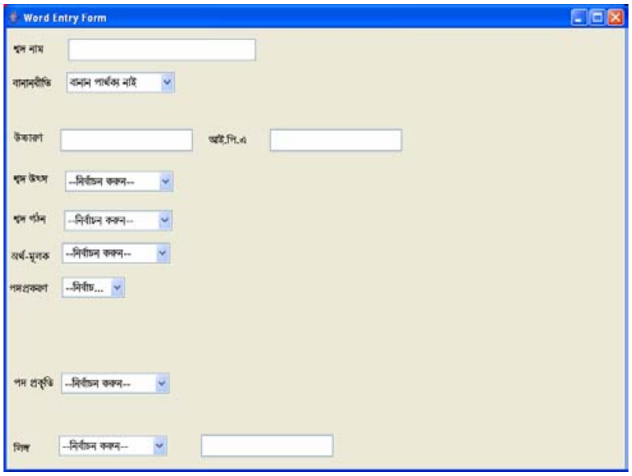

MVC Architecture of Digital Lexicon
XML File containing the final verified lexicon items.
Java Classes for Extracting, Inputting and Verifying data in XML file.
XML File Containing user lexicon entries at post verification stage.
MODEL
Extraction and Inputting of data into XML file.
Lexicon Entries made from Users from the web.
Required logic definition using JSP and Servlets
CONTROLLER
Request & Response
Request & Response
Request & Response
View: A complete Word List stored in our lexicon.
View: Search/ Advanced Search capability for a specific Word with tags.
View: Add word in the lexicon(Tagger).
VIEW

Fig 1: MVC architecture of the Lexicon Development.

The web interface will help different users to contribute to the lexicon development. Each word might have numerous entries from different users sharing their own knowledge. All the entries will be taken into account and will be readily stored in the primary database and later scrutinized for redundancy or errors by some language specialists in the project. The interface allows people to leave incomplete information of a word and store it, which can be viewed by other users and they can add to the existing information. The web interface allows multiple ways of viewing the lexicon and search options that would make the job of the lexicon contributors much easier.



Word Entry Form

শব্দ নাম

বানানরীতি: বানান পার্থক্য নাই

উচ্চারণ | আই.পি.এ

শব্দ উৎস: --নির্বাচন করুন--

শব্দ গঠন: --নির্বাচন করুন--

অর্থ-মূলক: --নির্বাচন করুন--

পদপ্রকরণ: --নির্বাচ...

পদ প্রকৃতি: --নির্বাচন করুন--

লিঙ্গ: --নির্বাচন করুন--

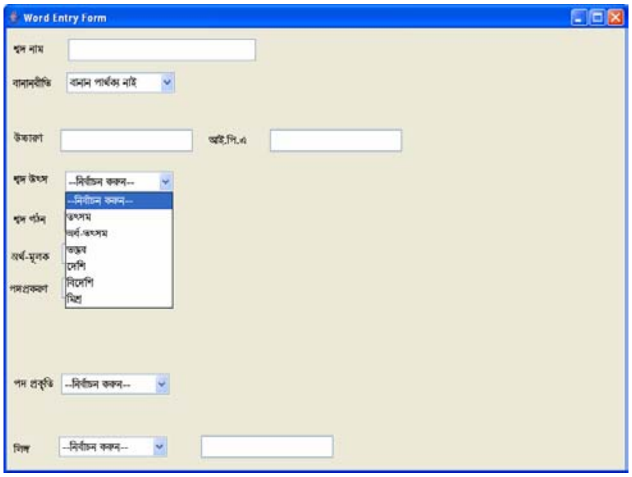
Fig 2: Standalone tool for populating the lexicon.

The Bangla tag set for the lexicon development project will initially contain the following list of basic tags along with their detailed classifications.

শব্দ= head-word
উচ্চারণ= pronunciation
শব্দার্থ = Meaning of Word
উদাহরণ=example usages
সমার্থ=synonym
শব্দের উৎস= source of word
পদ প্রকরণ= parts of speech

বচন = Number
লিঙ্গ=gender
প্রত্যয়=suffix
বাচ্য=voice
সন্ধি=shondhi (phonological or orthographic spelling rules)
সমাস=compound-word

Fig 3: Bangla Tagset for the Lexicon Development.



Word Entry Form

শব্দ নাম

বানানরীতি: বানান পার্থক্য নাই

উচ্চারণ | আই.পি.এ

শব্দ উৎস: --নির্বাচন করুন--

- --নির্বাচন করুন--
- তৎসম
- অর্ধ-তৎসম
- তদ্ভব
- দেশি
- বিদেশি
- মিশ্র

শব্দ গঠন

অর্থ-মূলক

পদপ্রকরণ

পদ প্রকৃতি: --নির্বাচন করুন--

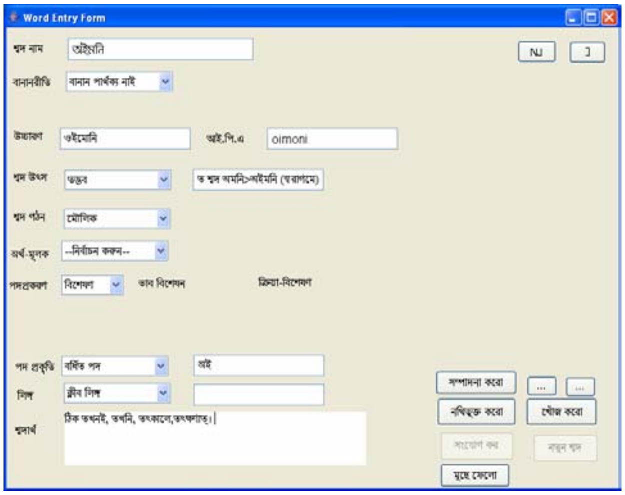
লিঙ্গ: --নির্বাচন করুন--

Fig 4: Selection of Specific Tag Information

The interface allows users to choose from the different classifications of the tags corresponding to each word. For example the source of word tag has six options to choose from.

XML Schema of the lexicon

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" targetNamespace="http://www.bu.ac.bd" xmlns="http://www.bu.ac.bd" elementFormDefault="qualified">
<xs:element name="Shobdo">
<xs:complexType>
<xs:sequence>
<xs:element name="item" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
<xs:element name="word_name" type="xs:string"/>
<xs:element name="word_serial" type="xs:string"/>
<xs:element name="word_spelling" type="xs:string"/>
<xs:element name="word_bangla_pronunciation" type="xs:string"/>
<xs:element name=" word_ipa_pronunciation" type="xs:string"/>
<xs:element name=" word_source" type="xs:string"/>
<xs:element name=" word_formation" type="xs:string"/>
<xs:element name=" word_category" type="xs:string"/>
<xs:element name=" word_POS_name" type="xs:string"/>
<xs:element name=" word_POS_category" type="xs:string"/>
<xs:element name=" word_POS_subcategory" type="xs:string"/>
<xs:element name=" word_characteristics" type="xs:string"/>
<xs:element name=" word_gender" type="xs:string"/>
<xs:element name=" word_meaning" type="xs:string"/>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>
```



Word Entry Form

শব্দ নাম: অইমনি

বানানরীতি: বানান পার্থক্য নাই

উচ্চারণ: ওইমোনি আই.পি.এ: oimoni

শব্দ উৎস: তদ্ভব ত শব্দ অমনি>অইমনি (স্বরাগমে)

শব্দ গঠন: মৌলিক

অর্থ-মূলক: --নির্বাচন করুন--

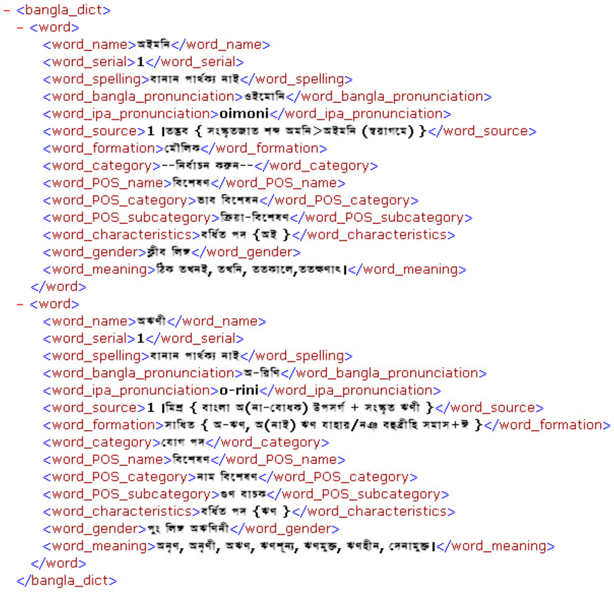
পদপ্রকরণ: বিশেষণ ভাব বিশেষণ ক্রিয়া-বিশেষণ

পদ প্রকৃতি: বর্ধিত পদ অই

লিঙ্গ: ক্লীব লিঙ্গ

শব্দার্থ: ঠিক তখনই, তখনি, তৎকালে,তৎক্ষণাত্।

Fig 5: Sample entry of a Bangla word.



```
- <bangla_dict>
 - <word>
     <word_name>অইমনি</word_name>
     <word_serial>1</word_serial>
     <word_spelling>বানান পার্থক্য নাই</word_spelling>
     <word_bangla_pronunciation>ওইমোনি</word_bangla_pronunciation>
     <word_ipa_pronunciation>oimoni</word_ipa_pronunciation>
     <word_source>1 ।তদ্ভব { সংস্কৃতজাত শব্দ অমনি>অইমনি (স্বরাগমে) }</word_source>
     <word_formation>মৌলিক</word_formation>
     <word_category>--নির্বাচন করুন--</word_category>
     <word_POS_name>বিশেষণ</word_POS_name>
     <word_POS_category>ভাব বিশেষণ</word_POS_category>
     <word_POS_subcategory>ক্রিয়া-বিশেষণ</word_POS_subcategory>
     <word_characteristics>বর্ধিত পদ {অই }</word_characteristics>
     <word_gender>ক্লীব লিঙ্গ</word_gender>
     <word_meaning>ঠিক তখনই, তখনি, ততকালে,ততক্ষণাৎ।</word_meaning>
   </word>
 - <word>
     <word_name>অঋণী</word_name>
     <word_serial>1</word_serial>
     <word_spelling>বানান পার্থক্য নাই</word_spelling>
     <word_bangla_pronunciation>অ-রিণি</word_bangla_pronunciation>
     <word_ipa_pronunciation>o-rini</word_ipa_pronunciation>
     <word_source>1 ।মিশ্র { বাংলা অ(না-বোধক) উপসর্গ + সংস্কৃত ঋণী }</word_source>
     <word_formation>সাধিত { অ-ঋণ, অ(নাই) ঋণ যাহার/নঞ বহুব্রীহি সমাস+ই }</word_formation>
     <word_category>যোগ পদ</word_category>
     <word_POS_name>বিশেষণ</word_POS_name>
     <word_POS_category>নাম বিশেষণ</word_POS_category>
     <word_POS_subcategory>গুণ বাচক</word_POS_subcategory>
     <word_characteristics>বর্ধিত পদ {ঋণ }</word_characteristics>
     <word_gender>পুং লিঙ্গ অঋণিনী</word_gender>
     <word_meaning>অনৃণ, অনৃণী, অঋণ, ঋণশূন্য, ঋণমুক্ত, ঋণহীন, দেনামুক্ত।</word_meaning>
   </word>
 </bangla_dict>
```

Fig 6: XML Document of sample words in Bangla lexicon

3. FUTURE WORK

This project aims at developing a Bangla Online Digital Dictionary with a significant number of words and annotations. Bangla Morphological analysis is dependent on a complete lexicon of root words, prefixes and suffixes. So the digitized version of the annotated lexicon will greatly help in this process. The lexicon will also be a key component for Machine Translation and Text-to-Speech, so the aim would be to integrate the lexicon for these purposes.

ACKNOWLEDGEMENT

This work has been partially supported through PAN Localization Project (www.PANL10n.net) grant from the International Development Research Center, Ottawa, Canada, administered through Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan. The research team would also like to thank Mr. Kamrul Hayder for his valuable contribution with digital copy of a hundred thousand word list and for his time and effort in formalizing the list of tags.

CONCLUSION

Lexicon is considered as one of the basic resource for language analysis and research for many languages. This reflects both ideological and technological change in the area of language research. The use of lexicon in Bangla language for various technological developments as well as for various linguistic studies in Bangla language can open up many new avenues for us. This lexicon will be useful for producing many sophisticated automatic tools and systems, besides being good resources for Natural Language Processing. The collaborative effort will be useful for easy acquisition of different levels of information against all lexical entries. The tag set has not been finalized yet. But significant effort is going on to formalize a standard list of tags which would also be a significant achievement.

REFERENCE

[1] Towards Full Automation of Lexicon Construction, Richard Rohwer,, Computational Lexical Semantics Workshop at HLT-NAACL (Human Language Technology Conference/North America Chapter of the Association for Computational Linguistics) 2004.

[2] Computer Assisted Bangla Words POS Tagging, Goutam Kumar Saha, Amiya Baran Saha and Sudipto Debnath

Proc. International Symposium on Machine Translation NLP & TSS (iSTRANS-2004), New Delhi 2004.

[3] An electronic dictionary as a basis for NLP tools: The Greek case, Ch. Tsalidis , A. Vagelatos and G. Orphanos.

Neurosoft S.A., 24 Kofidou Street, GR-14231 Athens, Greece, R.A. Computer Technology Institute, TALN 2004, Session Poster, Fès, 19–21 avril 2004.

[4] J. Hasan. “Automatic dictionary construction from large collections of text”. Master's thesis, School of Computer Science and Information Technology, RMIT University, 2001.

[5] Daniel Jurafsky and James H. Martin, "Speech and

Language Processing: An Introduction to Nataural

Language Processing, Computational Linguistics, and

Speech Recognition ", Prentice Hall, (2000).

[6] Conceptual Modeling And The Lexicon, Jeroen Hoppenbrouwers, PhD Thesis Paper, 1997, Center for Economic Research, Tilburg University.

[7] http://ootips.org/mvc-pattern.html

[8] Morphological Parsing of Bangla Words Using PC-KIMMO, Sajib Dasgupta Dr. Mumit Khan, Department of CSE, Department of CSE,BRAC University, Bangladesh. BRAC University, Bangladesh, ICCIT 2004, BRAC University, Bangladesh.

[9] Feature Unification for Morphological Parsing in Bangla, Sajib Dasgupta Dr. Mumit Khan, Department of CSE, Department of CSE, BRAC University, Bangladesh. BRAC University, Bangladesh, ICCIT 2004, BRAC University, Bangladesh.