

## 1. Project goal:

Compilation of the prothom-alo corpus, converting the corpus to Unicode format as well as producing a document on the approach to compilation of a Bangla corpus.

## 2. Corpora and its importance in language research:

A corpus, most simply can be defined as a collection of texts, which can be of a particular language or of more than one language. It is as important a resource as any other resources in linguistic research. Natural language processing has always been an interesting research area and computational linguistics is one important part of it. The key resource to any linguistic research is a trained, annotated corpus that can enhance the language processing capability such as automatic part-of-speech tagging, information extraction etc. As an example many lexicographers have found that they can more effectively create dictionaries by studying word usage in a very large linguistics corpora. Corpora have significantly affected research in linguistic discipline and have succeeded to open a new area of research.

The corpus being such an important resource has made linguistic researchers produce corpora of their language. English language has many corpora varying in size, genres and purposes. The first English corpus is the Brown corpus which was created by W. Nelson Francis and Henry Kucera in the early 1960's. There are many other English corpora available such as The British National Corpus, London Lund corpus, Penn TreeBank corpus, the International corpus of English and many more . Unfortunately in Bangla we do not have any corpus available.

The goal of the project was to convert the available text collected from the on line version of Prothom-alo and convert it to Unicode format to make it usable for further research.

## 3. Compilation procedure :

The corpus has been created in two phases. These phases are:

- 1) Collecting raw text from “Prothom-alo” website.
- 2) Converting to Unicode.

### 3.1 Collection of text:

The raw text for the corpus was collected from the prothom-alo home page- [WWW.Prothom-alo.net](http://WWW.Prothom-alo.net). This was done using a web crawler program that surfed through the website of prothom-alo and downloaded all the news available for the year of 2005- including magazines, periodicals published by them. The crawler crawled for one night to collect all the text, which were in html of course. After that using the Linux script all the files were converted to text files.

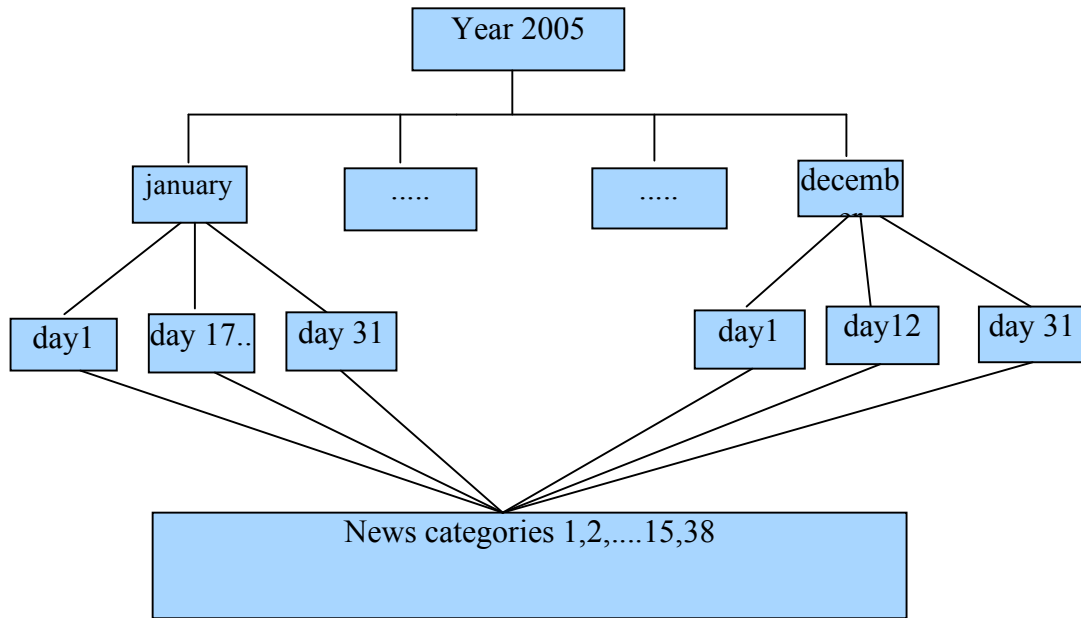


Fig: Prothom-alo corpus

### 3.2 Unicode convertioin:

The second part in creating the corpus was to convert all the files to unicode format. This was needed because Unicode support for Bangla is much more rich than any other format. Prothom-alo uses two types of fonts, namely “Bansi Alpona” and “Prothoma”. The previous was in use up to 2005 and currently they are using “Prothoma” for the on lone version of the newspaper. So a Java application was written which recursively searched the folders and sub folders and convert all the text files to Unicode.

#### 4. Processing the text: Categorizing the news:

After converting to Unicode the corpus is now ready for any further processing required. An important and useful processing could be categorizing the news. Prothom-alo presents news in 27 different categories. Each category has a category id; i.e. category 1 is for “prothom pata”, category 2 is for “sesh pata”, etc. So if all the news that belong to the same category can be merged together it will enable us to analyze and carry out some research on like text categorization etc. A Java application is used which surfs through the news of all the days and collects news of the same category in one file.

The corpus is also available as a single text file.

#### 5. Analysis :

We now have a corpus which is:

- 318 Mb in size
- 12 million word/token count

which is a big one. Some basic statistical analysis is now on offer which can help others to carry out research. The analysis that i have done are vocabulary creation, word frequency analysis, N-gram analysis etc. The results of the analysis have been attached at the end of this document. The “CMU-Cambridge Statistical Language Modeling Toolkit” version 2 was used for the analysis.

#### 6. Methodologies to create a Bangla corpus:

Creating a corpus is a time consuming task. It involves serious planning, hard labor and one needs to be very patient in the process. One must consider what types of researches will be carried out using the corpus, what the size will be(in word count), what genres are to be included in the corpus etc. no matter what it is, collecting the text is the major problem. There are two types of resources available:

- Electronic media or the World Wide Web

- Resources such as printed media(books, magazines), magnetic media etc

## 6.1 Electronic media: The Web-

For sophisticated applications like transliteration, historical analysis of language or text categorization we need very large corpus. Although such large corpora exist for most of the “large”(widely spoken) languages of the world (Chinese, Hindi, English, French, etc.), it is difficult to collect enough text for the vast majority of languages, such as Bangla. It is because working or developing applications for Bangla dates very recent past. Although copious text is available in book format, we have to face difficulties such as copyright issues . Again the manual labor required to scan these books is so huge that no corpora are generally available for most indigenous languages in the world.

Against all this, the Internet is an extremely valuable resource. Enormous amount of text are available on the Internet in electronic format and copyright is not a problem for much of it. We can collect these articles to build up a corpus. But there is a problematic issue - the articles are very much scattered. I have searched for a two day period and the number of texts that i found written in Bangla is not worth mentioning. To address this problem two automated approaches can be taken:

- 1) Use a web crawler.
- 2) Take the help of a commercial search engine.

### 6.1.1 Use a web crawler:

In this approach a program that search for pages containing words of bangla. Initially a list of words from the dictionary(300 words, or any magical number) will be used as search terms. The crawler will start from a root page. We can save the returned pages and process for further links which will provide pages and the process continues recursively. It can run for 2 days, for example.

The problem with this approach is that in the web English texts predominates other language and the percentage of the Bangla text found can be very low.

### 6.1.2 Use a search Engine:

We can use a search engine to overcome this problem. An Internet search engine limited to a region is more likely to contain texts of that region. for Bangle, again we can use a list of 300 words and limit the search to Asian or south Asian region. we can then try to figure out what language the pages contain. This approach can give us better results.

### 6.1.3 Processing the text :

#### 6.1.3.1 Finding out the language of the text:

To find out whether a text contains Bangla is a challenging task. The html source of the page can be analyzed to find out the font used. Each html source has a FONT tag where the font used in the document is mentioned. We can extract it from the source, try to match it against a database containing all the names of Bangla font currently present.

The other way to do this is to use an N-gram model. From the collected text we can randomly choose 100 words (again it can be any number, but the higher the better) and see how many of those are found in the Bangle dictionary. If the percentage is above a threshold value, the text can be considered as containing Bangla. However there can be other methods of doing this and this is a topic worth further research.

#### 6.1.3.2 Converting to Unicode:

As mentioned earlier, converting the text to Unicode is as important as anything. If web pages contain Unicode text, then no conversion is required. However, if True Type Font (TTF) is used, then we have to convert the text to Unicode. True Type Fonts has own encoding format. By scanning the font tag and running font specific application, the conversion can be achieved.

### 6.2 Non-electronic media:

Other than the web there are resources such as printed media, which includes magazines, books, articles, government announcements etc. But since bangla do not have a very good OCR system yet, the scanning process will produce erroneous text. optical media can be another source for Bangle

text. If books are available in optical form such as CD or DVD then we can collect that to add to our corpus. Bangla academy can be a good place to start with, Both for printed and optical media.

#### 7. Future development on “Prothom-alo Corpus”:

Although the “Prothom-alo Corpus” is a big corpus(12 million words), it is not balanced. It contains only news, which will be mostly useful for researching on different constructions of Bangla grammar. For wider research needs more genres need to be included.

#### 8. References:

- [1] English Corpus linguistics”- by Charles F. Mayer, Cambridge University press, 2002.
- [2] Dewan Shahriar Hossain Pavel, Asif Iqbal Sarkar and Mumit Khan, A Proposed Automated Extraction Procedure of Bangla Text for Corpus Creation in Unicode, Proc. International Conference on Computer Processing of Bangla (ICCPB-2006), Dhaka, Bangladesh, 17 February, 2006.
- [3] Corpus Linguistics” - by Tony McEnery; Andrew Wilson, Edinburgh University Press, 1996.
- [4] Gerrit Botha and Etienne Barnard, Two approaches to gathering text corpora from the World Wide Web
- [5] [www.google.com](http://www.google.com)