A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges

Prof. Shikhar Kr. Sarma[1] *Himadri Bharali*[1] *Ambeswar Gogoi*[1] *Ratul Ch. Deka*[1] *Anup Kr. Barman*[1]
(1) DEPT. OF IT, GAUHATI UNIVERSITY, Guwahati – 781014, India
sks001@gmail.com, himadri0001@gmail.com, ambeswar@gmail.com, rdeka8258@gmail.com, anupbarman.gu@gmail.com

ABSTRACT

To study about various naturally occurring phenomenons on natural language text, a well structured text corpus is very much essential. The quality and structure of a corpus can directly influence on performance of various Natural Language Processing applications. Assamese is one of the major Indian languages used by the people of north east India. Language technology development works in Assamese language have been started at various levels, and research and development works started demanding a structured and well covered Assamese Corpus in UNICODE format. Here we present various issues and problems related to building an Assamese text corpus. We review our experience with constructing one such corpus including about 1.5 million words of Assamese language. It will provide a significant effort by serving as an important research tool for language and NLP researchers.

KEYWORDS: Assamese, Corpus, linguistics, Natural Language Processing.

Proceedings of the 10th Workshop on Asian Language Resources, pages 21–28,
COLING 2012, Mumbai, December 2012.

1 Introduction

Language corpora are extensively used in language technology and linguistic researches. There arose a tremendous interest in building and developing computerized language corpora in recent few years. The study of digital corpora of various languages offers the students and the researchers an opportunity to work with language data with variety of tools and techniques in terms of computational procedures and programs.

Assamese is one of India's national languages and belongs to the Indo-Aryan language Family. It is spoken by about 15 million people. The matter of fact is that Assamese lacks computational linguistic resources. There are no prior computational works on this language, spoken widely in north-east India. Recently, researchers have begun to involve in the development and enrichment of the language of Assamese in the field of Natural Language Processing (NLP). Such NLP activities demanded the need of building up a large amount of corpora in the languages.

The term 'corpus' is used to refer to a collection of linguistic data (covering spoken and written) in a language for some specific purposes and these data are to be stored, managed and analysed in digital format. A corpus may be quite small, for instance, consisting of 50,000 words or texts, or very large, consisting of millions of words. The Cambridge International Corpus collected by Cambridge University Press contains 700 million words or text and it has being increased all the time. The Brown Corpus, the first computer based corpus, comprising one million words of edited written American English, was created at Brown University in early sixties. Corpus is assumed to be a representative of a given language so as to make it useful for linguistic analysis. The word 'corpus' is derived from Latin meaning 'body'. Theoretically, corpus is (C)apable (O)f (R)epresenting (P)otentially (U)nlimited (S)elections of texts.[1]

Corpus is the basis of all kinds of linguistic researches. The scope of corpus is a vast one. The areas of corpus-based researches are – grammatical studies of specific linguistic construction, building of reference grammar, lexicography, language variation and dialectology, historical linguistics, translation studies, language acquisition, language pedagogy, and natural language processing and so on.

The need of language corpora has given rise to the study of corpus linguistics. It is not a branch in linguistics, but a methodology which helps in pursuing linguistic research. From the very beginning, modern corpus linguistics has been closely associated with the development of computer software for corpus analysis. In modern corpus linguistics, the linguists and the computer scientists share a common goal that it is important to depend on the real or actual language data (speech or written) for carrying out any kind of linguistic analysis. Moreover, it is an approach which satisfies two main purposes: how people use language in day-to-day communication and to build up intelligent system to interact with human beings.

2 Related Study

Modern day corpora are of various types. In fact, it is a very crucial task of classifying language corpora into different types. However, written corpus, spoken corpus, general corpus, monolingual corpus, bilingual corpus, un-annotated corpus, annotated corpus, parallel and learner corpus are worth mentioning.

India is a land of diverse linguistic groups. But in comparison to other advanced countries, it possesses no language corpora due to the lack of language technology development. All the linguistic researches are done in traditional mode. But recently it has made a deliberate attempt to build digital language corpus. Generation of corpora could enhance various linguistic and NLP developments and thus protect languages from extinction.

The Kolhapur corpus of Indian Languages, created at the Shivaji University, Kolhapur in 1988. It consists of approximately one million words of Indian English data. But it fails to represent Indian national language used in the country.

The urge to build corpora in Indian languages is fueled by the all round growth of language technology in India. Consequently, the Department of Electronics (DOE), Govt. of India begun corpus development from 1991. The technology development for Indian languages (TDIL) programme had taken initiation in building machine-readable corpora of nearly 10 million words within three years for all Indian national languages. Indian Institute of Technology (IIT), Kanpur was entrusted to develop tool for language processing and machine-aided translation system from English to Indian languages. However, the Department of Electronics (DOE) could develop corpora of 3 million words for each Indian language and had to suspend further continuation by the end of 1994.

Later on some Indian experts had decided to start more corpus generation and processing. The Ministry of Information Technology (MIT), Govt. of India, Department of Information Technology (DIT), the Central Institute of Indian Languages (CIIL), Mysore had taken steps to create corpora in major Indian languages (Hindi, Nepali, Marathi, Konkini, Assamese, Manipuri, Kannada, Sanskrit, Bangla, Telegu, Tamil, Gujrati, Oriya, Punjabi, Malayalam, Urdu, Kashmiri). These corpora are in UNICODE and annotated according to the Corpus Encoding Standard (CES) guidelines.

[1] Dash, N.S. (2005) Corpus Linguistics and Language Technology: With Reference to Indian Languages, New Delhi, Mittal Publication

3 Text Corpus Generation In Assamese

In the present study, we mainly deal with the building the structure of Assamese un-annotated raw corpus comprising approximately 1.5 million words (total 1,577,750 words) and also try to highlight the problems faced during the process of building it. This huge collection of texts would be helpful in the linguistic and non-linguistic studies, cross-linguistic comparisons and, all other areas of language technologies.

There are various issues that are associated with the design, development and management of corpus. Such issues vary according to the type and utility of the corpus. In fact, speech corpus development is different from text corpus. Developing a text corpus in Assamese is concerned with the issues like the overall size or length of corpus, selection of the type of genres, the number of text and range of writers, data collection, computerizing the data and validation of raw corpus. These are discussed below:

1. The overall size or length of the corpus

Size or length of corpus is an important factor of consideration. The overall size of Assamese corpus is determined as 1.5 million words (total 1,577,750 words). But before determining the length of the corpus, certain decisions are taken such as – availability of resources, time for data collection and computerizing them. So far as time factor is concerned, the present corpus is expected to be completed within approximately two years. The matter of fact is that the length of a corpus is determined not by focusing on the overall length of the corpus, but focusing more on the internal structure of the corpus: the number of genres is to be included in the corpus, the length and number of individual text samples. The expected words would be collected from three main categories: Media, learned material and literature. These main categories are again divided into some sub-categories. And accordingly, the collection of the total 1.5 million words is shown in below table:

Main Category	Category	Sub-category	Expected words per category	Root category count
Media				637000
	Newspaper		337500	
		News		
		Sports		
		Editorials		
		Reports		
		Letter		
		Cartoon		
		Horoscope		
		Arts related news		
		Science related news		
		Cookery		
		Reviews		
		Obituaries		
		Classified ads		
		Publicity		
		Trivia		
	Magazine		299500	
		Film		
		Women's		
		Informative/General		
		Children		
		Others		
Learned Material				229250
	Science		116250	
		Biology		
		Botany		
		Computer		
		Geoscience		
		Chemistry		

		Mathematics		
		Physics		
		Medicine		
		Zoology		
		Others		
	Arts		113000	
		Economics		
		Linguistics		
		History		
		Psychology		
		Sociology		
		Law		
		Politics		
		Philosophy		
		Religion		
		Other		
Literature				711500
	Short fiction		120000	
		Light fiction		
		Sentimental fiction		
		Science fiction		
		Detective fiction		
		Serious fiction		
	Criticism		52500	
		Plays		
		Theatre		
		Novels		
		Short stories		
	Theatre		75000	
		Full length plays		
		Comedy		
		Tragedy		
		Art plays		
		Light theatre		
	Novel		300000	
		Full length novel		
		Sentimental novel		
		Science fiction		
		Detective novel		
		Historical novel		
		Art novel		
		Auto-biographical novel		
		Other light fiction		
	Trivia		15000	
		Jokes		
		Anecdotes		
		Fables		
		Current riddles		
		Proverbs		
	Art and craft		37000	
	Letter		18500	
		Administrative		

		Personal		
	Didactic material		75000	
		Encyclopaedia		
		School and college texts		
			Total	1577750

TABLE 1- Representation of the collection of 1.5 million words from various genres.

2. Selection of genres included in the corpus

Genres are selected keeping in mind the purpose and utility of a corpus. A large number of written genres are included in Assamese corpus. These genres are listed in the Table 1 and they represent the language in true sense. Importantly, in selecting the genres we do not consider the poetry since the language structure is very much flexible depending on the writer's views.

3. Determining the number of text and range of writers included in the corpus

After selecting the genres, next task is to determine how many the numbers of texts and the range of writers to be included in the corpus of Assamese. There are a huge number of texts available in the languages, but we are very selective in determining the number of texts. Similarly, in the selection of the range of authors, we give importance to both eminent authors and little-known authors. In the selection of newspaper and magazines, we are very much selective. In case learned material, we try to cover up all necessary domains (as shown in Table 1). Regarding the text selection we also consider the time factors so that we can include texts from various time periods.

4. Collecting data

Data collection is a crucial task of building a corpus. There are various ways to collect written texts for Assamese corpus such as buying printed texts, use of library (with necessary permission), photocopying and scanning the texts etc. In this context, the issue of copyright is well maintained.

5. Computerizing data

After data collection, we prepare for entering those data in an electronic format. It is a very laborious process. And most importantly these data are only typed by the native speaker of Assamese language because a non-native speaker is not familiar with the writing style of a given language.

The composer has the most important task of entering the metadata also. He has to give certain information about the text, for example, genre of the text, type of the text (report, fiction, drama, article etc.), the name of the text, the name of author and editor, name of publisher, date of publication, place of publication, the page numbers of the texts etc.

6. Validating the raw corpus

The process of validating the whole raw corpus starts just after the completion of entering the computerized data. It is done by the experts (must be a native speaker) who possesses linguistic knowledge of Assamese. Sometimes, the data compiler validates the data himself. But the cross-validation of the data is best deserved.

4 Problems Faced During the Overall Process of Building of the Corpus:

1. Problem of availability of data

In corpus building of Assamese corpus, if the composer sometimes fails to find out certain selected text material, then he can replace that selected text by another text to that corresponding author. Besides certain academic materials like engineering and medical books are not generally found written in Assamese language. In such cases, we need to replace those materials with some other related materials available in the language.

2. Linguistic problem

In computerizing the data, it is seem to face certain linguistic problem such as

- Spelling error

The compiler faces certain spelling errors in the text materials. And it is the task of the compiler to enter the correct forms of the word in computerizing the data. Some of the common spelling errors are dealt with in building Assamese corpus mentioned below: (AS: Assamese Sentence; TF: Transliterated Assamese Form; ET: English Translation)

Error: AS: সমিচীন, দূর্গা, পুৰস্কাৰ

TF: *samiciin, duurgaa, puraskaar*

Correct: AS: সমীচীন, দুর্গা, পুৰষ্কাৰ

TF: *samiiciin, durgaa, purashkaraar*

ET: suitable, goddess Durga, award

- Spelling variation

In Assamese language, there are certain words which have more than one accepted spellings. These spelling varies from text to text depending on the writer's acceptance. Sometimes the composer seems to become confused seeing spelling variation for the same word in the text materials. He has to take crucial decision in this regard of selecting different spelling forms of the same word. In Assamese texts also such kinds of spelling variations are very frequent. Depending on the frequency of the different word forms, the composer has to keep all of them in the digital files. For example:

To represent river Ganga, two accepted spellings are গংগা (gangaa: river Ganga), গঙ্গা (gangaa: river Ganga)

To represent office, two accepted spellings are কার্যালয় (kaarjyaalay: office), কার্য্যালয় (kaarjyaalay: office)

- Syntactic error

Syntactic errors are commonly found in Assamese texts and it is the responsibility of the compiler to write the correct forms. For example

Error: AS: মানুহজন ঘৰলৈ যাম।

TF: *maanuhjan gharaloi zaam.*

Correct: AS: মানুহজন ঘৰলৈ যাব।

TF: *maanuhjan gharaloi zaaba.*
ET: The man will go to home.

- Dialectical variation

Assamese corpus texts contain a large amount of dialect words. These words are retained as it is. For example

AS:মাকজনীয়ে কেঁচুৱাটোৰ *হেনাহতে* মৰেমৰে।

TF: *maakjaniiye kecuwator henaahate mare*
ET: The mother has deep love for her baby.

AS: '*ঐ আপা,* এইফালে আহ', মানুহজনে মাত লগালে।

ET: Hello boy, come here, the man called.
TF: 'oi aapaa, eiphaale aah,' maanuhjane maat lagaale.

- Junk characters

Junk characters are occurred profusely in the texts due to typing error. In Assamese texts too, junk characters are dealt with care. For instance

Error: AS: পূ্জা, মৌ্, আৰু্ etc.
TF: puuuujaa, mouou, aaruu
Correct: AS: পূজা, মৌ, আৰু
TF: puujaa, mou, aaru
ET: worship, bee, and

- Incomplete sentence

Incomplete sentences in the texts create problem for the compiler. It is important to avoid incomplete sentences while entering the data by the compiler. Incomplete sentence found in Assamese text materials is given below:

Error: AS: ডকাইতিৰ কথা শুনি মানুহজনে …
TF: dakaatir kathaa shuni maanuhjane …
ET: hearing about the robbery the man …
Correct: AS: ডকাইতিৰ কথা শুনি মানুহজনে *উচপ খাই উঠিল*।
TF: dakaatir kathaa shuni maanuhjane *ucap khai uthil*.
ET: hearing about the robbery the man became shocked.

- Hyphenated words

Assamese possesses hyphenated word forms. But these are not uniform in all the texts. Therefore, hyphens between words are removed in Assamese texts, except reduplicated forms.

Error: AS: লাহে-লাহে, লগে-লগে
TF: laahe-laahe, lage-lage
ET: slowly, instantly
Correct: AS: লাহে লাহে, লগে লগে
TF: laahe-laahe, lage-lage
ET: slowly, instantly

- Punctuation markers

In some texts, punctuation markers like full stop, comma, dash etc. are not marked uniformly. Two or more sentences are joined together without any overt connectors. In that case, the compiler puts appropriate punctuation markers reading out the data in the texts. Some of these errors are commonly found while building Assamese corpus, such as

Error1: AS: কৌটিল্যই লিপিকাৰৰ গুণাগুণ বিচাৰ কৰি কৈছে যে লিপিকাৰে কেৱল আখৰকেইটা লিখিব জানিলেই নহ'ব।

TF: koutilyai lipikaarar gunagun bicar kari koise ze lipikaare kewal aakharkeitaa likhib janilei nahaba
ET: after examining the writer's creations Kautilya commented that it is not sufficient for the writer only to know how to write

Error2: AS: মূর্ত্তিৰ কথা শেষ নহ'ল খঙতে লালজীয়ে ৰুদ্ৰ মূর্ত্তি ধৰিলে
TF: muurtir kathaa shekh nahal khangate laalajiiye rudra muurti dharile
ET: Murtty did not completed his speech Lalaji became raged in anger

Correct1: AS: কৌটিল্যই লিপিকাৰৰ গুণাগুণ বিচাৰ কৰি কৈছে যে, 'লিপিকাৰে কেৱল আখৰকেইটা লিখিব জানিলেই নহ'ব।'

TF: koutilyai lipikaarar gunagun bicar kari koise ze, 'lipikaare kewal aakharkeitaa likhib janilei nahaba।'

ET: After examining the writer's creations Kautilya commented that, 'it is not sufficient for the writer only to know how to write.'

Correct2: AS: মূর্ত্তিৰ কথা শেষ নহ'ল। খঙতে লালজীয়ে ৰুদ্ৰ মূর্ত্তি ধৰিলে।
TF: muurtir kathaa shekh nahal. khangate laalajiiye rudra muurti dharile
ET: Murtty did not complete his speech. Lalaji became raged in anger.

Conclusion

In this paper, we have presented a description of processes involved in creating the raw corpus in Assamese and also a discussion on the problems faced during the process. Corpus is being regarded as a multi-dimensional in nature. Corpus in Assamese opens up new avenues in the field of language technology, communication, exchange of information, language education and linguistic activities etc. In the future, it should be our great responsibility to create bigger corpora, consisting of billions of words, in our native language. Besides, steps are to be taken in annotating the raw corpus which would result in building morphological analyzer, spell checking tool, concordancer, machine translation, speech recognition etc. in the language of Assamese.

References

Bora, L.S. (2006). *Asamiya Bhasar Ruptattva*, M/s Banalata, 2006.

Goswami, G. C. (2009). Asamiya Vyakaran Pravesh, 3rd edition. Bina Library, Guwahati. 2009.

Goswami, G. C. (2004). Asamiya Vyakaranar Maulik Vicar Pravesh, 4th edition. Bina Library, Guwahati. 2009.

Aston, G (Ed. 2004) *Learning with Corpora*. Cambridge: Cambridge University press.

Jayaram, B.D and Rajyashree, S.K.: *Corpora in Indian Languages. Central Institute of Languages Manasagangotri,* Mysore 570006, India.

Jayaram, B.D. (1996). *Development of Corpora in Indian Languages: Problems and Suggested Solutions*. Paper presented at workshop of Indian Language Corpus and its applications at CIIL, Mysore.

Ganesan, M: *Tamil Corpus Generation and Text Analysis:* Annamalai University, Annamalainagar, Tamilnadu, India.

Jaimai Purev and Chimeddorj Obdayar. (2008). *Corpus Building for Mongolian Language* in Proceedings The 6th Workshop on Asian Languae Resources, 2008

Steven A. and Steven B. (2010). *The Human Language Project: building a universal corpus of the world's languages*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden.

N.S. Dash (2005). Corpus Languistics and Language Technology with Reference to Indian languages: Mitali Publication, New Delhi.

Charles F. Mayer: English Corpus Linguistics An Introduction. Published by the press Syndicate of the University of Cambridge.

Stella E.O. Tagnin: A Multilingual Learner Corpus in Brazil. Published: Rodopi.

McEnery and Andrew Wilson: Corpus Linguistics. Published by Edinburge University press.

Michael McCarthy: Touchstone From Corpus to Course Book. Published by the syndicate of the University of Cambridge.

Kenji Imamura and Eiichiro Sumita (2002). Bilingual Corpus Cleaning Focusing on Translation Literality. In: 7th International Conference on Spoken Language Processing (ICSLP-2002).

Dash, Niladri Sekhar: Language corpora. A Mittal Publication.

Dash, Niladri Sekhar. (2004). Language corpora: Present Indian Need. In the Proceedings of the SCALLA 2004 Working Conference.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz: Building a Large Annotated Corpus of English: The Penn Treebank. Published in: Journal Computational Linguistics – Special issue on using large corpora: II.

Motaz K. Saad, Wesam Ashour. (2010) OSAC: Open Source Arabic Corpora: Published at the 6th International conference on Electrical and Computer System (EECS,10), Nov.25-26,2010, Lefke, North Cyprus.