Corpus Building of Literary Lesser Rich Language- Bodo: Insights and Challenges

Biswajit Brahma[1] *Anup Kr. Barman*[1] *Prof. Shikhar Kr. Sarma*[1] *Bhatima Boro*[1]
(1) DEPT. OF IT, GAUHATI UNIVERSITY, Guwahati – 781014, India
bswjtbrahma@gmail.com, anupbarman.gu@gmail.com,
sks001@gmail.com, borobhatima@gmail.com

ABSTRACT

Collection of natural language texts in to a machine readable format for investigating various linguistic phenomenons is call a corpus. A well structured corpus can help to know how people used that language in day-to-day life and to build an intelligent system that can understand natural language texts. Here we review our experience with building a corpus containing 1.5 million words of Bodo language. Bodo is a Sino Tibetan family language mainly spoken in Northern parts of Assam, the North Eastern state of India. We try to improve the quality of Bodo corpora considering various characteristics like representativeness, machine readability, finite size etc. Since Bodo is one of the Indian language which is lesser reach on literary and computationally we face big problem on collecting data and our generated corpus will help the researchers in both field.

KEYWORD : Bodo language, Corpus, Linguistics, Natural Language Processing.

1 Introduction

The term corpus is derived from Latin corpus "body" which it means as a representative collection of texts of a given language, dialect or other subset of a language to be used for linguistic analysis. Precisely, it refers to (a) (loosely) anybody of text; (b) (most commonly) a body of machine readable text; and (c) (more strictly) a finite collection of machine-readable texts sampled to be representative of a language or variety (Mc Enery and Wilson 1996: 218). Again, Corpus is a machine readable texts (both spoken and written) document stored in machine systematically collected from different sources. It is an important text in digital media world. It is defined as corpus and in plural corpora a collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The main purpose of a corpus is to verify a hypothesis about language - for example, to determine how the usage of a particular sound, word, or syntactic construction varies. Corpus linguistics deals with the principles and practice of using corpora in language study. A computer corpus is a large body of machine-readable texts[1]. So it is the computerization of varieties text (various domains of texts such as literature, science, sports etc.) of a given language. Corpus may be of monolingual, bilingual and multi lingual format of machine readable data etc. It is an annotated and tagged component of parts of speech. It is most important for computing to make it accessible worldwide via internet. Moreover it is a valid machine readable data of a given language which gives us proper information of a language where it follows linguistics principles.

The need of language corpora has given rise to the study of corpus linguistics. It is not a branch of linguistics but the methodology that helps in analysis and research of naturally occurring language through the help of computerized corpora, i.e. with the specialized software. 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.

It is not easy to classify corpora into various types. 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, unannotated corpus, annotated corpus, parallel and learner corpus are worth mentioning.

2 Related Studies

The first computer corpus, "Brown Corpus" was created early in the 1960s by Nelson Francis and Henry Kuccera. But it was not warmly accepted by the linguistics community, yet they are regarded as the pioneer of the Corpus linguistics. Creation of corpus is the most important to keep alive from the extinction of languages from this world. Keeping in the notice for the development of the Indian scheduled languages the government of India also started corpus generation revolution in India. As a consequence of its view the government of India emphasized for the development of Indian scheduled languages in technological media world and initiated the technological development works on scheduled languages in 1991. Accordingly machine readable texts have been developed in some major languages in India viz. Hindi, Indian English, Punjabi, Telugu, Kannada, Malayalam, Marathi, Gujarati, Oriya, Bengali, Assamese, Sanskrit, Urdu, Sindhi and Kashmiri in many universities and technology Institutes of India. Later development of corpora for the remaining languages had been done as to run parallel with the other languages for the better gaining to all.

Bodo language belongs to the Sino Tibetan language family under the sub branch of Assam-Burmese group. This language speakers have spread highly in the northern part of the Brahmaputra valley. They are also scattered in all the districts of Assam state more or less. Apart from they can be found in the North-Eastern states like Arunachal, Nagaland, Mizoram, Manipur, Tripura, Northern parts of West Bengal, Bihar and adjoining part of the Bangladesh, Nepal and Bhutan in small concentration. This language has the three distinct dialects according to some researchers. But Promod Chandra Bhattachrya in his doctoral thesis book "A descriptive analysis of the Boro Language" stated four dialects of Bodo language. These are i)

1. Crystal, David. 1992. *An Encyclopedic Dictionary of Language and Languages*. Oxford,85 (cf.)

North-west dialects areas having sub dialects of North Kamrupand North Goalpara district ii) South-West dialect area comprising South Goalpara and garo hills district iii) North-Central Assam dialect area comprising Darrang lakhimpur districts and a few places of Arunachal Pradesh iv) Southern dialect area comprising Nowgong North Cachar , Mikir Hills, Cachar and adjacent districts. It has two types of tone high and low tone. Intonation, juncture, agglutinating features is there in this language. Use of high back unrounded /w/ vowel is more frequent in this language. There are 22 phonemes 16 consonant and 6 vowel phonemes. Highly use of monosyllabic word can be found in this language. Devnagiri script is the main script of this language.

Recently the language has recognized as the scheduled language by the government of India in 2003. The language is the medium of instruction up to the 10th standard in school from 1963. In 1984 the language is recognized as the state associate official language in the districts of Kokrajhar and Udalguri. This language is introduced as major subject in the colleges under Gauhati University affiliation in the very recent.

3 Bodo Text Corpus

Consideration of size or length of corpus is an important factor. Overall size of Bodo corpus is determined as 1.5 million words. It is also determined of the availability of data, time for computerizing them in the format. The determined size of the corpus is collected from the expected three main category- Media, Learned and Literature. These categories are again classified into sub categories during the creation of Bodo corpus as given against in the following table. Thus the corpus generation is done keeping in mind of determined target from the different domain collection resources in Bodo. In Bodo media house collection news paper like dailies, weeklies; bi-weeklies and magazines monthlies, bi-monthlies etc are very less. And medical science, engineering, technological word terms very rare, those terms words are taken from the "Glossary of Administrative Terms" published by the Ministry of Human Resource Development (Department of Higher Education), government of India. Entire collection of the data was taken from the written texts document from the various resources as given in the following tree diagram. In Media category total 637000 roots words have been entered comprising category and subcategories. 229250 root words from the learned material category including category and sub categories and a total count of 711500 root words from literature category have been computerised in the text format as shown in the following tree diagram. Having all these three category the Bodo corpus has been created and shaped a total word counting of 1.5 million words (total 1,577,750 words).

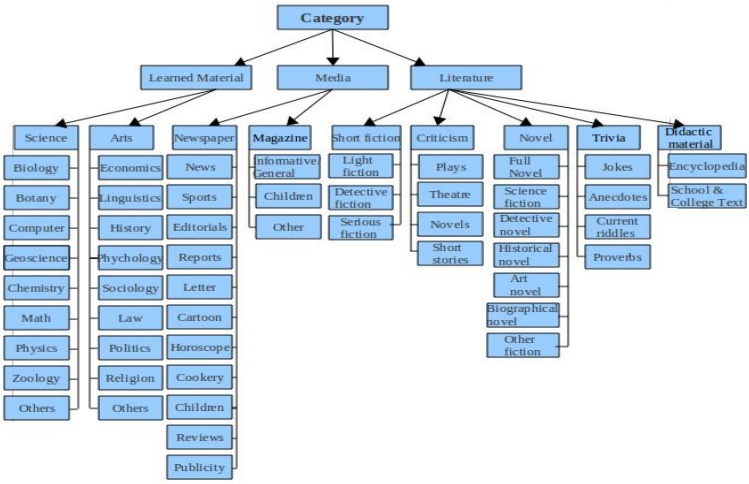

FIGURE 1 – A tree diagram showing categories of corpus contents

3.1 Content Selection

A large number of written genres are selected keeping in the mind of its purpose and utility of a corpus. But poetry genre is not included in our selection. Some genres are not in Bodo like Obituaries, Classified advertisements in the news paper. So these are cannot be found in the format data. There is no film's and women's magazine in Bodo but getting a few representations in the magazines it was included in the corpus. All these genres represent the actual sense of the language and they are listed in the above given diagram.

It is the second task after selecting the genres to determine how many the numbers of texts and the range of writers to be included in the Bodo corpus. 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. But in case of news paper and magazine we select all the news papers and magazines published in Bodo as news paper items are not available in the language. In case of learned material also we try to cover up all necessary domains. And in literature the science fiction and sentimental fiction are also not available, so they are avoided in the corpus while generating the corpora.

3.2 Data Collection

For building a corpus in Bodo, data are collected from the written texts of the language. In order to collect data, we mainly go for buying books, use of library materials, some texts are also photocopied and scanned etc. The issue of copyright is always kept in mind.

3.3 Computerizing data

The collected data are now ready for entering on to the computer. The task of computerising the text materials is a very crucial. These data are compiled by the native speakers only. Trio-lingual a (Bodo-English-Hindi) dictionary of Bodo Sahitya Sabha published by Onsumwi Library, Kokrajhar, Assam is followed while entering the texts in the format for standardization of the language and in some cases linguistics standardization is also followed.

3.4 Validation

The next process is the validation of those typed data. Validation must be done by the expert. He should be a native speaker of Bodo langauge who has the linguistic command over the Bodo langauge. Generation of Bodo corpus is based the standardization of "Boro-Ingraji-Hindi Swdwbbigung" a trio-lingual (Bodo-English-Hindi) dictionary of Bodo Sahitya Sabha published by Onsumwi Library, Kokrajhar, Assam and in some cases linguistics standardization is also followed. Present discussion is done generated raw corpus in Bodo of few years back. Validation is done manually because this language does not have still tagged corpus and annotated texts. It has a long way to reach its fruitful goal.

4 Issues related to Bodo corpus generation

The size and quality of the corpus depends on the data of a respective language on its resources. Bodo does not have such a rich resources in various fields of its language and the literature and in the science (Chemistry, physics etc.) and in the media house whatever it is electronics or print media. Child literature is very less as compared to other literature and medical science and engineering and the terms of respective subject's words are very rare. Medical science, administrative engineering terms words are entered in the corpus from the glossary book published by MHRD, government of India. Provisions like obituary, classified advertisements etc. are not there in the news paper. In these entire field the resources is increasing day to day. Here we mention some challenges faced during building period of Bodo corpus:

Spelling variation

It is a major problem in Bodo literature as well as in other writing fields also. No standard or uniform spelling system is followed by the authors or writers in this language for their writings though standardized language is followed. Many authors and writers go their own wishes. So it is found very difficult while entering texts documents in the format. As for example: [थाखाय, थाखाइ (*thakhai*): for], [बायदि, बाइदि

(*baidi*): etc.] here whereas both the word [थाखाय, थाखाइ (*thakhai*): for] is used to mean the same meaning but spelling is changed in the last letter of the word i.e. य letter is changing to इ in the second word and also in second example [बायदि, बाइदि (*baidi*): alike] it also refers same meaning though the word spelling in the middle is changed from य to इ. Both in the above example there is no change in their word meaning but its spelling is varying in both the words. So it is one of the major problems which one has to be follow while entering the text for corpus.

Word Split

Splitting of words is found frequently in Bodo while entering the texts into format. These words are edited and correctly entered by the compiler. For example:

BS: बुंदोंमोन दि TF:bungdwngmw di

Correct:BS: बुंदोंमोनदि TF:bungdwngmwdi

Joined Sentence/Word

Many times joined sentence is found in the texts while entering the texts. The compiler itself corrected the sentence and entered in the format.

BS: गस्लाखौ गानहां जाबायमोन।रामोना आंखौ लिंहरो।

TF:goslakhou ganhan jabaimwn.Ramwna angkhou linghorw

Correct:BS: गस्लाखौ गानहां जाबायमोन। रामोना आंखौ लिंहरो।

TF:goslakhou ganhan jabaimwn. Ramwna angkhou linghorw.

Punctuation Error

A large number of punctuation incorrect marks are found in the texts materials. These are removed and corrected by the compiler. As for example

BS: खोलाहा थिङै। सानैजों गोबालायासै । TF: khwlaha thingwi. Sanwijwng gwbalayaswi.

Correct:BS: खोलाहा थिङै सानैजों गोबालायासै। TF: khwlaha thingwi sanwijwng gwbalayaswi.

Dialect Words

Sometime many dialect words are found in the texts. These words are corrected by the compiler and entered in the data format for the corpus. For instance

BS: कोरर्टारखो TF: quarterkhw

Correct:BS: कोरर्टारखौ TF:quarterkhou

Grammatical error

There are lots of sentences which are found grammatically incorrect in the texts. Those sentences are edited and entry is done correctly by the compiler as given in the following example.

BS: जेब्ला रांरासिया गहेल थानाय कोरर्टारखो मोनहैयो अब्ला हर 10 टासो जाबायमोन ।

TF: jebla rangrasiya gohel thanai quarterkhou mwnhwiyw obla hor 10 tasw jabaimwn

Correct:BS: जेब्ला रांरासिया गहेल थानाय कोरर्टारखौ मोनहैयो अब्ला हरनि 10 टासो जाबायमोन ।

TF: jebla rangrasiya gohel thanai quarterkhou mwnhwiyw obla horni 10 tasw jabaimwn.

Hyphenated words

Bodo also have hyphenated words, those are in case of multiword expression words. But surprisingly, there are a few hyphenated words in Bodo within a word which are found in the texts. Those words are compiled and entered by the compiler in the format. For example

BS: गामि-आरिफ्रा TF: gami-arifra

Correct:BS: गामिआरिफ्रा TF: gamiarifra

Incomplete sentence

Incomplete sentences in the texts are very frequent in the Bodo texts. Complier has to face problem . For instance

BS: बियो गाज्लं Ø थाङो । TF: biyw gajlaong Ø thangw.

Correct:BS: बियो गाज्लं गाज्लं थाङो। TF: biyw gajlong gajlong thangw.

Conclusion

It is seen from the above discussion that there is no developed fonts in Bodo. Due to in-uniformity of spelling the compiler of the corpus has to face several problems while entering the text into the format. In such cases they have to correct themselves. There is no science and sentimental fictions in Bodo and in some fields like journals like women's, children's, whether it is monthlies, bi-monthlies and news papers whether it is dailies, weeklies etc are very rare. The entire generation of Bodo corpus is based the standardization of trio-lingual (a Bodo-English-Hindi) dictionary of Bodo Sahitya Sabha published by Onsumwi Library, Kokrajhar, Assam in some cases and linguistics standardization is also followed. Present discussion is done generated raw corpus in Bodo of few years back. Validation of this generation corpus is done manually as this language does not have still tagged corpus and annotated texts. It has a long way to reach its fruitful goal.

References

Brahma, Promod Chandra (Compiler): *Boro-Ingraji-Hindi Swdwbbigung*, Onsumwi Library 2003, Kokrajhar Assam.

Ministry of Human Resource Department. Government of India 2007, *Glossary of Administrative Terms*

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).