A survey of developing Bishnupriya Manipuri Corpus for Multilingual Dictionary

Dr Prajadhip Sinha [1]
[1] Assistant Professor, Dept of Computer Science
Jotsoma, Kohima, Nagaland
(E-mail: prajadhip@rediffmail.com)

***Abstract*—**This paper discuss about the corpus and its requirements during building a multilingual dictionary of Bishnupriya Manipuri language. Due to the lack of awareness like other Indian languages, this language is studied less frequently. As a result the language still lacks a good corpus and basic language processing tools. This paper discusses about the different types of corpus, methods and methodology of being develop the corpus for Multilingual Dictionary. The paper also analyzes the applications areas of corpus and in brief the Tagging process of corpus in NLP and computational linguistics areas.

Keywords*—Bishnupriya Manipuri, Corpus, NLP, Multilingual Introduction*

I. INTRODUCTION

The term corpus, derived from Latin, usually refers to a body of texts (collection of linguistics data) either in written or spoken form (transcribed recorded speech). It is a representative sample of different varieties of language preserved in machine readable form which can be used as a starting point of linguistic description or a means for verifying hypotheses about a language (Crystal 1980). According to Sinclair (1996) corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of language. A corpus, designed methodically, should have the following characteristics features.

- It should be large in size containing a healthy amount of language data.
- It should be authentic and reliable in representation of language.
- It should consist of structured collection of text specifically compiled.
- It can be either a simple plain text or a text with annotation
- It should be user-friendly for easy handling.
- It should be properly and systematically documented
- It should have authentic referential value

II. BISHNUPRIYA MANIPURI LANGUAGE

The Bishnupriya Manipuri Language comes under the group of Indo-Aryan languages. The structure of the language is undoubtedly of Indo-Aryan origin, but it also retains some older sounds of medieval Meitei. The vocabulary is influenced by many Indo-Aryan and Tibetan-Burmese terms. There are two dialects in the Bishnupriya Manipuri language, namely, the Madai Gang dialect or the dialect of the village of the queen and the Rajar Gang dialect or the dialect of the village of the king. The Madai Gang dialect is also known as Leimanai and the Rajar Gang dialect as Ningthaunai. The term Leimanai is derived from Leima (queen)+(ma) nai (attendant), meaning the attendants of the queen, and the word Ningthaunai, from ningthau (king)+(ma) nai (attendant) meaning the attendants of the king. Unlike the dialects of other tribes, these dialects of Bishnupriya are not confined to distinct geographical areas; they rather exist side by side in the same localities

Thirty-five principal phonemes present in Bishnupriya Manipuri of which eight vowel sounds, such as i, e, ε, a, α, ∂, ò and u; twenty-five consonant sounds such as h, p, b, t, d, ţ, ď, ?, ph, th, ţh, kh, c∫, ∫δ, m, n, η, l, r, φ, s, ∫, ĥ and ђ and two semi vowels ŏ and ĕ. The vowel sounds can be represented in a tabular form as follows:

	Front	**Back**
Close (High)	i	u
Half-Close (High-mid)	e	ò
Half-Open (High-mid)	ε	∂
Open (Low)	a	α

The consonant sounds can be represented in a tabular form as follows:

	Bilabial	Dental	Alveolar	Palato-Alveolar	Palatal	Retroflex	Velar	Glottal
Plosive	p,b	t,d				ţ, ď	k,g	?
Aspirate	ph	th				ţh	kh	
Plosive with glottal	b'	d'				đ	ġ	

Affricate				cʃ, ʃδ				
Affricate with glottal				ʃδ'				
Nasal	m		n				η	
Lateral			l					
Flapped			r					
Fricative	φ	s		ʃ				h,ђ
Semi-vowel	ŏ(w)				ĕ (y)			

The voice aspirates, such as as, bh, dh, gh and jh never occur in this language. They are replaced by four stops and an affricate with glottal closure, such as h, b',d',g', z' etc. The -ch- sound is also not found and it is pronounced as -s-.We have analysed the word structure of the Bishnupriya Manipuri language from the data of the corpus. Some portions of our result are shown below

Nouns

Bishnupriya Manipuri nouns that denote male and female beings are sometimes distinguished by suffixation or through pairs of lexically differing terms. the word মুনি (muni) and জেলা (ʤela) are used before the word to indicate masculine and feminine genders respectively .

E.g. – মুনি মানু (munimanu : man),
জেলামানু (ʤelamanu : woman)

The feminine gender is generally indicated by the use of the word জেলা (ʤela) after the words indicating common gender. Feminine gender is formed by adding the following suffixes to the masculine forms of words:

i) ◌ী (i): খুড়া (k^{h}uɹa : father’s younger brother) -> খুড়ি (k^{h}uɹi : the wife of father’s younger brother), জেঠাবা (ʤet^{h}aba : father’s elder brother) -> জেঠীমা (ʤet^{h}ima : the wife of father’s elder brother)

ii) ◌ী◌া◌ি (ani): চাকর (sakɔɹ : servant) -> চাকরানী (sakɔɹani : maid servant) iii) ◌ি (ni): চামার (samaɹ : cobbler) -> চামারি (samaɹani : female cobbler)

Being an agglutinative language, Bishnupriya Manipuri has the capability of generating hundreds of words from a single noun and verb root.

For example,
The root word মানু (manu: man) may form different inflected words.

1. মানুে◌য় (manuje) → মানু (manu) + ে◌য় (je) → Noun + NCM
2. মানুহাি◌ব (manuhabi) → মানু (manu) + হাি◌ব (habi) → Noun + Pl
3. মানুে◌র (manuɹe) → মানু (manu) + ে◌র (ɹe) → Noun + ACM

4. মানুগেয় (manugɔje) → মানু (manu) + গ (gɔ) + ে◌য় (je) → Noun + DM + NCM
5.মানুহাি◌বেয়েহ(manuhabijehe)→মানু(manu)+হাি◌ব(habi)+ে◌য়(je)+ে◌হ(he) → Noun + Pl + NCM + EM A

III. BASIC PRINCIPLES OF CORPUS BUILDING

A Corpus is considered as a building block for any language processing tasks and few principles for building corpus are as follows:

- The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise
- Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.
- Only those components of corpora which have been designed to be independently contrastive should be contrasted.
- Criteria for determining the structure of a corpus should be small in number, clearly separate from each other and efficient as a group in delineating a corpus that is representative of the language or variety under examination.
- Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in application.
- Samples of language for a corpus should wherever possible of entire document or transcription of complete speech events, or should get as close to this target as possible. This means that sample will differ substantially in size.

IV. TYPES OF CORPUS

- **Sample corpus:** A fixed sample of text, often used as a reference corpus for comparing.
- **Monitor Corpus:** A corpus which develops and is added to or filtered depending on the researchers needs**.**
- **Mini-Corpus:** a small corpus (e.g. to be compared with a reference corpus)
- **Multilingual Corpora:** Corpus in a variety of language.
- **Comparable Corpus:** Text in two language or two language varieties but not matched up.
- **Parallel Corpus:** Text are translation of each other, e.g. Canadian Hansard, Corpus of Version of Plato, Bible.
  - **Translation Corpus:** two or more set of text classified as either originals or transition, the purpose being to identify features of translation (Manchester: Baker)
  - **Diachronic Corpus:** Helsinki, LOBA V. FLOB

- **Learner Corpus:** Texts are written by language learner.

V. METHODS OF DATA INPUT

Data from electronic sources: In these process texts from newspapers, journals, magazines, books etc. are included if these are found in electronic form But unfortunately BISHNUPRIYA Manipuri are not available in UNICODE standard.

Data from the websites: This includes texts from web pages, web sites, and home pages, same problem as we have no web pages available in BISHNUPRIYA Manipuri in proper UNICODE format.

Data from e-mails: Electronic typewriting, e-mails, etc. are also used as source of data. People of Manipur used the current trends of technology but still we cannot used our script for such works as due to the above said problems.

Machine reading of text: It converts printed texts into machine-readable form by way of optical character recognition (OCR) system. Using this method, printed materials are quickly entered into a corpus. Here printed or written materials are available but it will be a hard work to convert the scripts in electronic text, mainly in financial.

Manual data input: It is done through typing texts in computer. This is the best means for data collection from hand-written materials, transcriptions of spoken texts, and old manuscripts. The process of data input is indirectly based on the method of text sampling.

. VI: ALGORITHM.

STEP 1 Find each BISHNUPRIYA Manipuri string which occurs more than once in corpus *C*. Record and its frequency of occurrence *F* in an entry in MayBe database.

STEP 2 For each in the entries in MayBe. Find , the strings in MayBe with one more character than , where is a sub-stringof .Compute F = $\sum^{F}$. 1/λ Compute F2 = $\sum^{F}$, where is the first two entries and in , where is , the first *L* characters of , and where is the last *L* characters of .Compute N =F -F1 +F2. We extract each entry whose net frequency of occurrence is more than one as a CFS.

VII. CORPUS CLEANING

The corpora have to be cleaned from such unintended error as –typos, wrong splits, foreign characters, which may have been introduced while keying the text i.e in the process of digitization. For example, some of these corrections include, removal of '-', '~', '_' etc. which are introduce to break words at the end of lines while keying in the text. Sometimes the conversion of corpus text from one standard format to another may have introduced viz. alt, control characters (^C, ^M, ^Z etc.,.) are also removed. The resulting text is free from all such errors. Finally, the entire Language corpora shall be converted to case sensitive roman notations in wx- scheme.

Generally, five types of error may occur at the time of manual data

- Omission of character,
- Addition of Character,
- Repetition of Character,
- Substitution of character, and
- Transposition of character

To remove spelling errors, we need to thoroughly check the corpus and compare it with the actual physical data source, and do manual correction. Care has to be taken to ensure that the spelling of words used in the corpus must resemble with the spelling of words used in the source texts. Also, it has to be checked if words are changed, repeated or omitted, punctuation marks are properly used. Lines are properly maintained, and separate paragraphs are made for each other.

Besides error correction, we have to verify the omission of foreign words, quotations, dialectal forms, etc. after generation of a corpus. The naturalized foreign words are, however, allowed to enter into the corpus. Others should be omitted. Dialectal variations are allowed. Punctuation marks and transliterated words are faithfully reproduced. Usually, books on natural and social sciences contain more foreign words, phrases and sentences than books of stories or fiction. Similarly, quotations from other languages, poems, songs, mathematical expressions, chemical formulae, geometric diagrams, images, tables, pictures, figures, flow-charts and similar symbolic representations of the source texts are not entered into corpus. All kinds of processing and reference works become easier and authentic if corpus is properly edited and errors are removed. For cleaning corpus we used Perl program to clean the raw corpus.

VIII. APPLICATIONS

The purpose of corpus is not merely to gather a big file of different texts and store it on the computer, but rather to prepare the texts and put them in a certain format so that they can be used by search tools and the results of the search can be displayed in a way that is meaningful and useful to the linguist, teacher and learner especially at the advanced level. For example, scholars, teachers and learners can explore the use of a word in different types of texts to see how frequently this word is used, how many meanings it has, what syntactic environment it occurs in, whether the word has the same frequency of occurrence in all types of texts. Teachers can identify the most frequent words and select them as a basis for their material. There is also the study of syntactic structures and analysing the distribution of competing structures. For example, the uses of verb–subject vs. subject–verb word order in Bishnupriya Manipuri: which word order is more preferred in children's stories, interviews, and scientific documents? The MMD corpus is to be annotated with XML mark-up which includes information about the text, author, and source; this gives the opportunity to conduct empirical analyses which control extra-linguistic factors (such as age, sex, region, social class, and education level) and examine the accompanying linguistic variations. We hope our corpus would be further enriched with other information such as tagging which signifies information on word classes. This would make retrieval of useful information qualitatively and quantitatively much richer and easier to handle.

Corpus as knowledge resource: corpus is used for

- developing multilingual libraries,
- designing course books for language teaching,
- compiling monolingual dictionaries (printed and electronic),
- developing bilingual dictionaries (printed and electronic),
- multilingual dictionaries (printed and electronic),
- monolingual thesaurus (printed and electronic version),
- various reference materials (printed and electronic version),
- developing machine readable dictionaries (MRDs),
- developing multilingual lexical resources,

Corpus for translation support systems: corpus is used for

- language resource access systems,
- Machine translation systems,
- multilingual information access systems, and
- cross-language information retrieval systems, etc.

Corpus in speech technology: Speech corpus technology is used to develop general framework for

- speech technology, • phonetic,• lexical, • pronunciation variability in dialectal versions, automatic speech recognition, • automatic speech synthesis, • automatic speech processing, • speaker identification, repairing speech disorders, and • forensic linguistics, etc.

CONCLUSION

In this paper we provided an extensive survey how corpus can develop for Multilingual Dictionary, mainly based parallel corpora and mono corpora. We envisage that not only would this corpus fill a gap in the general field of corpus linguistics but it would also have a role in providing authentic material for teaching Bishnupriya Manipuri as a foreign language. In future, we will further increase the size of this corpus and will add more sections to the corpus. We are also planning to develop language processing tools on this language.

References

[1] Berwick, Robert C. 1989. Learning Word Meanings from Examples. In Semantic Structures: Advances in Natural Language Processing. Lawrence Erlbaum Associates, chapter 3, pages 89-124.

[2] Brill, E. 1994. Some Advances in Rule-based Part of Speech Tagging. In Proceedings of the Twelfth National Conference on Artificial Intelligence, pages 722-727. AAAI Press/The MIT Press.

[3] Carbonell, J. G. 1979. Towards a Self-Extending Parser. In Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics, pages 3-7.

[4] Cardie, C. 1993. A Case-Based Approach to Knowledge Acquisition for Domain-Specific Sentence Analysis. In Proceedings of the Eleventh National Conference on Artificial Intelligence, pages 798-803. AAAI Press/The MIT Press.

[5] Church, K. 1989. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In Proceedings off the Second Conference on Applied Natural Language Processing.

[6] Granger, R. H. 1977. FOUL-UP: A Program that Figures Out Meanings of Words from Context. In Proceedings of the Fifth International Joint Conference on Artificial Intelligence, pages 172-178.

[7] http://manipuri.freeservers.com/bpm.html

[8] Nayan Jyoti Kalita, Navanath Saharia and Smriti Kumar Sinha: Towards The Development of a Bishnupriya Manipuri Corpus

About Author:

Dr. Prajadhip Sinha is working as Assistant Professor in Kohima Science College. His research interests include Natural Language Processing (NLP), E-learning, Corpus Based Learning and Computer Applications etc.