A Quantitative Analysis of Synset of Assamese Wordnet: Its Position and Timeline

Shikhar Kr. Sarma
Gauhati University
Guwahati,Assam,India.
sks001@gmail.com

Dibyajyoti Sarmah
Gauhati University
Guwahati,Assam,India.
dibyasarmah@gmail.com

Ratul Deka
Gauhati University
Guwahati,Assam,India.
rdeka8258@gmail.com

Anup Kr. Barman
Gauhati University
Guwahati,Assam,India.
anupbarman.gu@gmail.com

Jumi Sarmah
Gauhati University
Guwahati,Assam,India.
jumis884@gmail.com

Himadri Bharali
Gauhati University
Guwahati,Assam,India.
himadri0001@gmail.com

Mayashree Mahanta
Gauhati University
Guwahati,Assam,India.
mayashreemahanta@gmail.com

Umesh Deka
Gauhati University
Guwahati,Assam,India.
deka_umesh@rediffmail.com

Abstract

The synsets in Assamese Wordnet play a significant role in the enrichment of Assamese language. These synsets are built depending on the intuition the native speakers of the language. There is no fixed rule in the arranging the positions of each synset. The present paper mainly aims to make a quantitative comparison of every synset position of Wordnet seeing the occurrences of these synsets in corpus of Assamese (approximately 1.5 million words). The experimental result of this comparison is represented with the help of diagrams. Again, it is an attempt to highlight the timeline of each synsets of Wordnet based on the corpus. It is dealt with the change of the synonymous word forms in course of times.

1 Introduction

Language is a central feature of human identity. Language is the identity of that particular community. No community can survive without a language. The language of the communities live in India is very ancient and rich. Similarly, Assamese language is also one of the ancient and rich languages of the north-eastern languages. Assamese has been regarded as a rich language with its own script and written literary texts since the ancient times. Assamese language belongs to the Satam group of the Indo-European language family. The main root of this language lies to the Indo-Aryan languages.

Dr. Banikanta Kakati has classified the development of Assamese language into three stages (Kakati, 2008):

A. Early Assamese (14th to 16th century A.D.)

This period again may be divided into a) Pre-Vaishnavite and b) Vaishnavite sub-periods. The earliest known Assamese writer is Hema Saraswati, who wrote a small poem ‘Prahlad Charit’. Sankardeva, the great Vaishnavite reformer in Assam, born in 1449 A.D. composed religious songs and drama. In his popularly known as Braja-Bali idioms (Goswami, 1983).

B. Middle Assamese (17th to 19th century A.D.)

The main characteristic of this period is the historical writings initiated under the inspiration of the Ahom court. These historical writings in prose are known as Buranjis. In the Ahom court, historical Chronicles were at first composed in their original Tibeto-Chinese languages, but when the Ahom rulers adopted Assamese as the court language, historical chronicles began to be written in Assamese. The language is essentially modern except for slight alterations in grammar and spelling.

C. Modern Assamese

The modern Assamese period begins with the publication of the Bible in Assamese by American Baptist Missionaries in the first quarter of the 19th century. The currently prevalent standard Asamiya has its roots in the Sibasagar dialect of Eastern Assam. The American Baptist Missionaries were the first to use this dialect in translating the Bible in 1813 A.D. In 1836 A.D., they started a monthly periodical called Arunodoy and in 1848 A.D., Nathan Brown published the first book on Assamese grammar. The Missionaries published the first Assamese-English Dictionary compiled by Miles Bronson in 1867 A.D. The Sibasagar Asamiya dialect came to be formally recognized as the Standard Asamiya dialect when it was made the official language of the state by the schools, courts, and Govt. officers in 1872. This Standard language is accepted by all other Asamiya dialect as the standard norm and was used for all formal occasions – in writing, in the classroom, in meetings, in the courts and offices and for inter-dialect communication also.

2 Assamese Corpus

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 stored, managed and analyzed in digital format. There is a huge amount of corpus in Assamese language consisting of approximately 15 or 20 lakh words based on the various Assamese literary or non-literary texts (such as magazines, newspaper, dramas, novels, stories, articles etc.). Words are collected from various texts ranging from 19th to 21st centuries (Sarma et al., 2012).

3 Assamese Wordnet

Wordnet is a repository of words of a language. Wordnet is basically a synonymous lexical database. Vocabulary plays a main role in building Wordnet. Assamese language possesses a huge amount of vocabulary; it becomes easy to build Wordnet in the language. The task of Assamese Wordnet building is almost ready to provide us with all the lexical words. Yet there are still many words in the language those need to be entered (Sarma et al., 2010).

Assamese Wordnet is built on the basis of Hindi Wordnet (Sarma et al., 2012). Here, words are shown according to the sense of the given context or sentence and accordingly, we can derive different meanings from them. For example, the Assamese word ‘*paani, and ‘farkaal’*’ has different meaning according to its sense in the context.

Paani (noun) –

Paanir para bemar hoi
Kaamtu paani hoi gol

Farkaal (Adjective) – Bataratu bar farkaal (not rainy)

- Raam farkaal monar maanuh
- Khuala manar manuh (Free minded)
- Path farkaal hoise (not muddy, dry)

4 Quantitative Analysis

The main aim of quantitative analysis is a complete description. Quantitative analysis allows for fine distinctions to be drawn because it is not necessary to the data into a finite number of classifications.

The resulting corpus contains over 1.5 billion words and 14958 Assamese Wordnet synset data. Initially, we have tried to find out the position of corpus and synset data. The synset category is classified as noun, verb, adjective and adverb for Assamese Wordnet. Here, we compare the frequency of Wordnet synset to the frequencies of Corpus data.

Some results of words position analysis in Assamese Wordnet with Corpus are mentioned below:

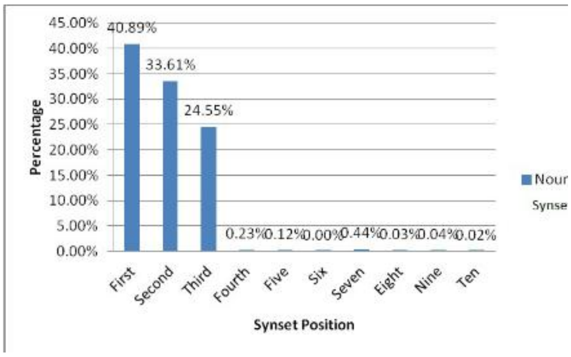

Figure 1: Position analysis of Noun Synset

In Figure 1, we have shown Synset Positions of Noun in Assamese Corpus. For the First position we have found 40.89%, for the second and third 33.61% and 24.55% respectively and so on. Similarly in Figure 2, Figure 3, Figure 4 we have shown Synset Positions and analysis of verb, adverb and adjective in Assamese Corpus.

Finally in Figure 5, it is clear that the finding of first position is always higher than the remaining synset position.

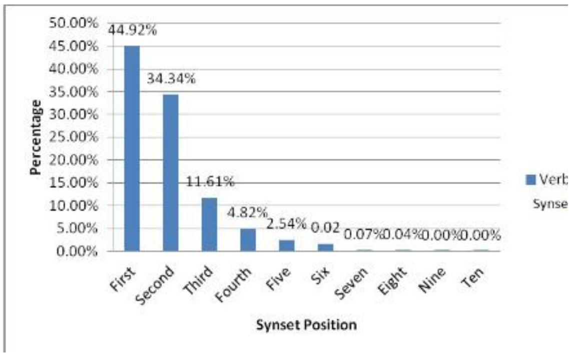

Figure 2: Position analysis of Verb Synset

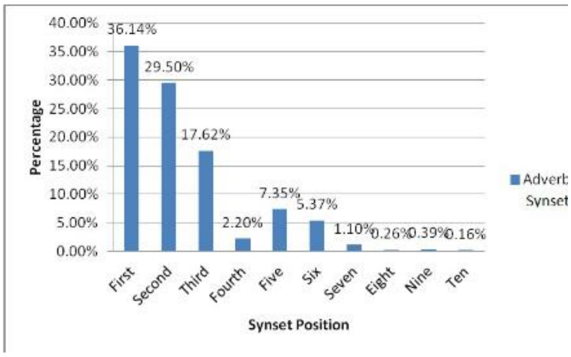

Figure 3: Position analysis of Adverb Synset

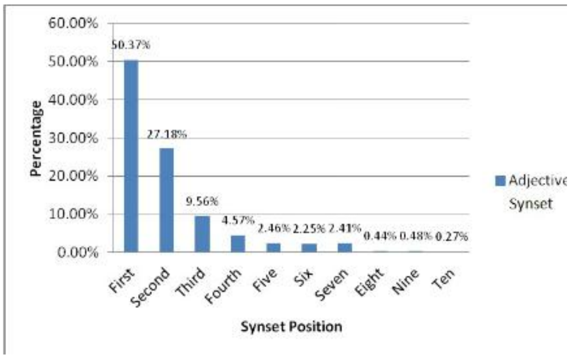

Figure 4: Position analysis of Adjective Synset

Final Result of Analysis

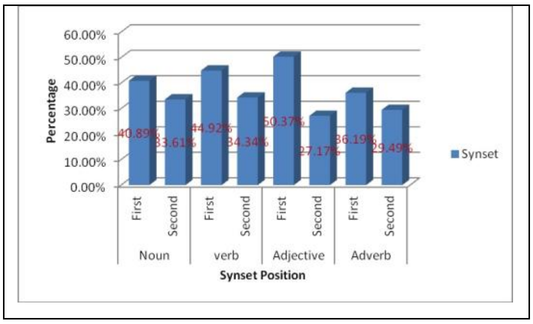

Figure 5: Final result of analysis

5 Timeline of Words

Wordnet has been built taking various words of hundred years. There are 38 synset positions in Wordnet. Especially, words are found to be most frequently used in the synset positions like 1st, 2nd and 3rd which cover a period from 1900 to 1995. It is worth mentioning here that we have not found any synonymous words after the synset position 17. Most of the words starting from synset position no. 1 to 5 we have seen words have became change from the old Assamese to modern forms. Thus, it has enriched the words in Assamese language.

While studying the synset in Assamese language, it is seen that most of the words used by the Christian missionaries have not been used at present times. It does not mean that these words have disappeared completely, but these are used less frequently with change in the forms of those words.

Examples of words change in Assamese language are mentioned below:

Synset Position	Forms of 20th Century	Present Forms
3	সইত Soit (true)(1918)	সত্য Satya(true)
7	আৰাৱ্‘aaraaw’(high sound) (1963)	চিঞৰ ‘chiyar’(high sound)
7	ক্লেশ ‘klesh’ (sorrow)(1900)	বেদনা‘bedanaa’ (sorrow)
3	ব্যাঘ্র ‘byaghra’ (tiger)	বাঘ ‘baagh’ (tiger)
11	কৰাইচ ‘karaaich’ (miser) (1938)	কৃপণ ‘kripan’(miser)

Table 1: Word Change of Assamese Language

In Table 1 we have shown the Synset Position in 1st Column and in the 2nd and 3rd column we have shown the words forms of 20th century and present day respectively.

Conclusion

The present paper makes an examination on the timeline of synset positions of Assamese Wordnet. In order to perform this task, mainly we refer to Assamese corpus covering time period from 1900 to 2008. In this corpus, there are more than 1.5 million texts. We consider all the synsets of Assamese Wordnet entered till date as it is in a developing stage. First we determine the timeline of all the corpus entries and secondly we map up these entries with their corresponding synset entries. While mapping we also consider the respective positions of each synset entries. After analysis the data, we basically found that from first to fifth position of synset entries are occurred frequently in the time period of our given corpus. But the results varied from different word categories those are clearly depicted in the above sections.

References

Golock C. Goswami. 1983. *Structure of Assamese*, Gauhati University , Assam Guwahati,Assam

Banikanta Kakati. 2008. *Assamese: Its Formation and Development*, Lawayers Book Stall, Guwahati , Assam

Shikhar Kr. Sarma, Moromi Gogoi, Rakesh Medhi, Utpal Saikia. 2010. *Foundation and Structure of Developing an AssameseWordnet*, Department of Computer Science Gauhati University, Proceedings of the 5th Global Wordnet Conference,Narosa Publishing House.

Shikhar Kr. Sarma, Utpal Saikia, Mayashree Mahanta, Himadri Bharali. 2012, Assamese Vocabulary and Assamese Wordnet Building: An Analysis. Global Wordnet Conference, Matsue, Japan

Shikhar Kr. Sarma, Himadri Bharali, Ambeswar Gogoi, Anup Barman. *A Structured Approach for Building Assamese Corpus*: Insights, Applications and Challenges, coling 2012, India