



## Sanskrit in Natural Language Processing

Divya Teja, Sasidhar Kothuru

Department of Electronics and Computer Engineering  
Sreenidhi Institute of Science and Technology, Hyderabad, India

---

**Abstract:** *This paper explains the importance of Sanskrit in the Natural Language Processing of computer programming. Sanskrit is considered because of its well defined structure and unambiguous rules. This paper also proposes an algorithm for machine understandable sentences through morphological analysis of Sanskrit words.*

**Keywords:** *Natural Language Processing, Artificial Intelligence, Semantic Net, Vibhakti, Dhatu and Sandhi.*

---

### I. INTRODUCTION

Sanskrit is one of the most Indo-European languages, was born in India and is considered as greatest treasure. This is because of its unambiguous representation. This led to consider it as a Natural Language for Artificial Intelligence. The essence of the language is from its grammar rules which were penned by Maharishi Panini in his book "Astadhyayi". Decades have passed, enormous efforts and money has been spent on designing of an unambiguous representation for Natural Languages<sup>[4]</sup>. Among all the Natural Languages, Sanskrit in its style is identified to be the best language which has minimum deviation. Panini discovered nearly 4000 rules for Sanskrit grammar in Astadhyayi<sup>[1]</sup>. Now this language with its rules is considered to be the greatest treasure in the field of Artificial Intelligence. The first part of this paper discusses the problems that are encountered while English is used in Natural language processing. Then the solutions for those problems are discussed using the Sanskrit Grammar and its importance in accomplishing the requirements of a Natural Language Processing. In the second part, algorithms, Morphological Analysis mechanism for the Sanskrit words are given.

### II. NATURAL LANGUAGE PROCESSING

Natural Language processing, a component of Artificial Intelligence is the ability of a computer program to understand human speaking languages and for the interaction with humans<sup>[2]</sup>.

In Artificial Intelligence, a mechanism is required for the machine understandable natural language. Such a human speaking Language like English consists of several ambiguities that make it unsuitable for Natural Language Processing. Few ambiguities encountered in English are:

1. Same Expression gives different meaning in different context

For Example: Where's the water? [In a chemistry lab, it must be pure water]

Where's the water? [When you are thirsty, it must be potable]

Where's the water? [Dealing with a leaky roof, it can be filthy]

2. There are many forms of sentences to express the same meaning

For Example:

Garry was born on November 14<sup>th</sup>.

Garry's Birthday is on November 14<sup>th</sup>.

3. In Natural Language Processing, Knowledge Representation is the primary step to make machine understandable Natural Language. This is possible with Semantic Nets.

### III. SEMANTIC NET

The main idea behind Semantic Net is, the meaning of a concept comes from the way in which it is connected with other concepts. In Semantic Nets, information is represented as a set of nodes connected to each other by a set of labeled arcs which represents relation among the nodes.[4]

Example: Sita gave a pen to Sriram.

In this sentence, as there is only one verb, it is considered as primary element and set of triples for that sentence are formed.

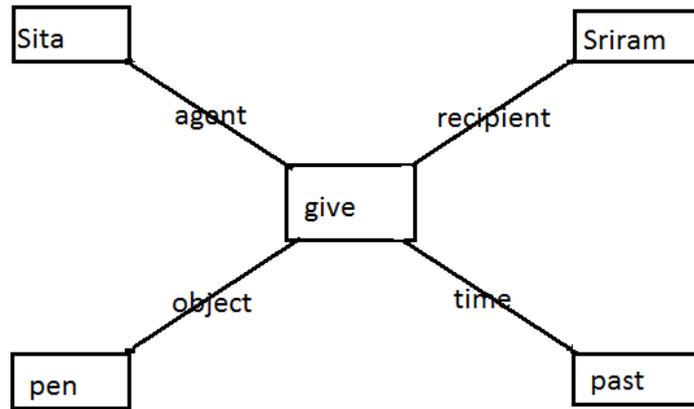
Give, agent, Sita.

Give, recipient, Sriram.

Give, object, Pen.

Give, time, past.

If there are more number of verbs, forming triples and mapping them becomes difficult and complex. This is another problem what English language faces when it is used in Knowledge Representation.



4. Words of English when stored in the database just serve as labels. Things like Data, Information and Meaning cannot be retrieved from that word.<sup>[8]</sup>  
For example: “Apple Marketing”  
When System encounters this word, it does not understand whether apple means fruit or a Brand.  
These problems are not only seen in English but also in many other Natural Languages like French, Spanish etc. Solution for ambiguous free context is found in Sanskrit.

#### IV. KNOWLEDGE IN SANSKRIT WORDS

There are about 1,025,109<sup>[3]</sup> words in English Language and with the invention of new device and other new word gets added into English language. Unlike English, every device has a word in Sanskrit. No new words are invented in this Language. This is because every letter in Sanskrit has a meaning and every word in Sanskrit describes only the property of an object but not the object itself. A person with knowledge in Sanskrit can easily understand the word and also the object it referred to.

Example: Moorkh (Sanskrit word): it means property of being Stupid.<sup>[5]</sup>

So to refer a stupid person, Vibhakti concept must be appended to the word. If first Vibhakti is used, word becomes Moorkaha which now means a stupid person.

Every Sanskrit word gives the Data, Information and Message because it is formed from primary elements called Dhatus which has a meaning. There are nearly 2200 dhatus in Sanskrit from which words are constructed. They are constructed by adding prefixes and suffixes to the dhatus. Through this method infinite number of words can be formed.

For example: Consider a word “Kshanaha”

It is formed from the dhatus: ksha + ana

“Ksha” means “exist invisibly”.

“ana” means “atom, very small quantity”.

Therefore “Kshanaha” means invisibly atom, small unit.

Sanskrit words also consist of information like Gender, Quantity, and tense used.

For Example: another word “Gachaami” which means going. It is derived from the dhatu, “Gachaa”. It also indicates its gender, quantity and tense.

#### V. WORD STRUCTURE AND INFLECTION CONCEPT

Word Structures in Sanskrit is in the below format<sup>[8]</sup>.

<prefix> <dhatu> <suffix>

In English language, in order to change the meaning of a sentence, introduction of new words into the sentence has to be done. Unlike English, most of the sentences in Sanskrit only require addition of prefix or suffix to a word.

For Example: “Gachaami” means “going”.

“aagachaami” means “coming back”.

In Sanskrit, words like is, an, the etc... doesn't have a separate word. In order to use it an addition of suffix or prefix to a word is done. But due to the concept called inflection words give the accurate meaning.

For example, consider a sentence: This is an Elephant.

In Sanskrit: Eshaha Gajaha

A sentence with four words in English is described with only two words in Sanskrit. Thus it decreases the storage space. All these features make Sanskrit unambiguous. This language does not use concepts like mapping or diagrammatic representation. It only follows the grammar rules penned by Panini. These rules make it ambiguous free and also help it to consider as a treasure in the field of Artificial Intelligence.

#### VI. ANALYSIS OF SANSKRIT WORDS

Understanding the actual meaning of Sanskrit sentences is very tricky. A step by step procedure with different techniques is given to find morphological analysis of Sanskrit words. An algorithm called Divide and Conquer is used. First

sentence is divided in to words. Later, after finding the meaning of each word they are combined. This technique is theoretically studied; practical studies are yet to be checked.

**Algorithm:**

Algorithm Divide

```
{
If small (p) [if there is only one word] then perform step 2, 3, 4;
return s(p)
else
{
divide p into smaller instances p1,p2,p3.....pn
[Sentence is divided whenever space is encountered]
Perform step 2, 3, 4
return
combine (s(p1), s(p2), s(p3),.....s(pn))
}
```

‘p’ refers to a sentence

p<sub>1</sub>, p<sub>2</sub>, p<sub>3</sub>...p<sub>n</sub> refers to ‘n’ words in the sentence

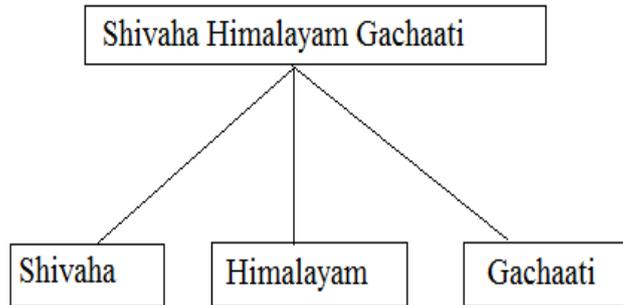
Steps to understand a sentence and the words.

**Step1:** Divide the sentence into separate words. A sentence is a combination of words. By considering Divide and Conquer technique, a sentence is considered as a larger problem and words in it as sub problems. In order to find solution for larger problem, smaller ones are solved first. Dividing the sentence in to words is done whenever space is encountered. Full stop defines end of the sentence. This division is represented in the form of a tree. Parent node consists of a sentence and their child nodes consist of the words in that sentence.

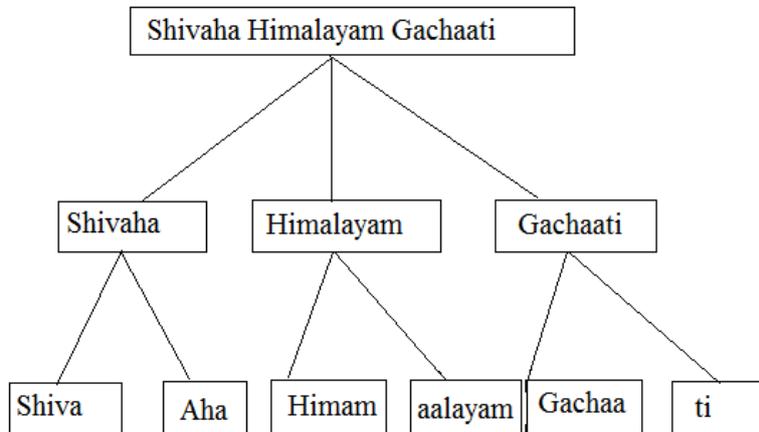
For example: Consider a sentence:

“Shivaha Himalayam Gachaati.”

[In English: Shiva is going to a place called Himalaya]



**Step 2:** Dividing the words into sub words is done using a process called sandhi division. This is done in order to understand the meaning of the word. The process of analysing the given word to extract the information encoded in the word is called as Morphological Analysis. This Analysis help us in finding the root word, prefixes, suffixes and features of grammar like number, gender etc.



**Step 3:** In sanskrit each dhatu has a meaning and every word is formed from the combination of dhatus. Last letter of the word before dividing, it checked in the database for the grammar meaning. If it does not have a grammar meaning then it is divided. Ending of the word in Sanskrit always has a grammar meaning. Division is avoided if word encounters a dhatu.

A database would be designed which consists of grammar rules and also the meaning of each dhatu

**Step 4:** Now meaning for each leaf node is checked . If a meaning for sub word is not found it is combined with the nearest node and again searched for the meaning. If meaning for a letter is found it is again combined with nearest node and checked for the meaning. The meaning which is most relevant to the sentence is retrieved.

In the above example, these are the meanings of each letters

Shiva-the destroyer

aha-[it is the end of the word, meaning of this is checked in grammar rules] Nominative which is singular and it is masculine.

Himam-snow

Aalayam-the abode

gachaa- [it is a dhatu formed from the dhatu 'ga'. It has to be predefined in the database] to move

ti-Nominative which is singular.

**Step 5:** Now each words end letter is analysed and information about that word and sentence is tabulated.

Word	Vibhakti	Gender	Person
Shivaha	1	Male	Singular
Himalayam	1	-	-
Gachaati	1	Male	Singular

**Step 6:** Now merging of nodes takes place. While merging meaning of that word is analysed. After analysing the meaning of each word, the meaning of all the words in the sentence are combined and meaning of the sentence is analysed by the system. This may not give the accurate meaning of the sentence but strong database development and further improvement in sandhi rules might give the accurate results.

From the example, we can derive that

Shivaha- a single person who is male and a destroyer.

Himalayam- an abode of snow.

Gachaati-going

Answer: A person who is a destroyer is going to place which is the abode of snow.

## VII. CONCLUSION

First part of this paper was dealt with reasons to support sanskrit as a best language for Natural Language Processing which is because of its grammar rules and word structure. Various examples were given to support it as the best language in Knowledge representation.

Second part of this paper presents a technique and also an algorithm to find meaning and to understand the sanskrit words and sentences through basic grammar rules. This is done in order to show how a machine understands sanskrit sentences. And also included algorithms similar to divide and conquer. The procedure is diagrammatically explained using a tree structure. Ambiguity aroused while combining or dividing words could be eliminated by Fuzzy logic or Fuzzy reasoning techniques.

## VIII. FURTHER STUDIES

Further research in this area has to be done in order to give appropriate meaning for all types of sanskrit sentences. Concentrating on parsers for Sandhi construction might lead us one step forward. A well structured Database and predefined dhatus have to be designed. Feasibility and desirability of Sandhi construction in computer system has to be checked. If a parser which divides a combinational word at a right place using Sandhi principles is designed, then understanding the word becomes more easy.

## REFERENCES

- [1] Panini's Grammar in Computer Science by Parul Saxena<sup>1</sup>, Kuldeep Pandey<sup>2</sup>, Vinay Saxena<sup>2</sup>.
- [2] Artificial Intelligence (Second Edition) by Elaine Rich and Kevin Knight.
- [3] [www.languagemonitor.com](http://www.languagemonitor.com)
- [4] Rick Briggs, Knowledge Representation in Sanskrit and Artificial Intelligence, <http://dx.doi.org/10.1609/aimag.v6i1.466>
- [5] Uttishthabharata.wordpress.com
- [6] Shashank Saxena and Raghav Agarwal, Sanskrit as a Programming Language and Natural Language Processing, Global Journal of Management and Business Studies, ISSN 2248-9878, Volume 3, Number 10 (2013), pp. 1135-1142.

- [7] Kridanta analysis for Sanskrit by N.Murali Dr.R.J.Rama Sree and Dr.K.V.R.K Acharyulu  
[8] Nandish V.Patel: On the relevance of declared shabda knowledge representation scheme in sanskrit to design of data structures for computer systems.

#### **AUTHORS**

- [1] **Ms. Divyateja** is a final year BTech student in the department of Electronics and Computer Engineering (ECM). She is one of the top students and interested in research in the field of Artificial Intelligence. She also has been placed in the campus placements.
- [2] **Dr. K.Sasidhar** is currently Head In-charge for the department of Electronics and Computer Engineering, SNIST, Hyderabad. He has 17 years of teaching experience. Encouraging students through his superior guidance for innovative projects, internships and for publishing papers.