Is Sanskrit the Most Suitable Language for Natural Language Processing?

Rashmi Jha
Associate Professor (CS), GIBS
New Delhi, GGSIPU, INDIA
Email ID: rashmijha1909@gmail.com

Aman Jha
Reporting & Analysis, Business Analyst
Credit Suisse, Polaris, Pune, INDIA
Email ID: amanjha2011@gmail.com

Deeptanshu Jha
CS Deptt., San Francisco State University
California, UNITED STATES
Email ID: jhadeeptanshu@gmail.com

Sonika Jha
IT Deptt., Amity University
Noida, UP, INDIA
Email ID: sonika.jha00911@gmail.com

Abstract – **Natural Language Processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages. The idea of using a natural language for computer programming is to make it easier for people to talk to computers in their native tongue and spare them the pain of learning a computer friendly language like assembly, C, C++, Java, LISP etc. Among all the Natural Languages, Sanskrit in its style is identified to be the best language which has minimum deviation. Panini, the creator of Sanskrit formulated 3,949 rules. This research paper explores varied distinctive features of AI like NLP, Semantic Net, Vibhakti, Dual Case, Inflection based Syntax etc. and how Sanskrit effectively triumphs over these limitations and fulfills the prerequisites of a Natural Language Processor.**

Keywords – Artificial Intelligence (AI); Dual Case; Equivalence; Inflection; Natural Language Processing (NLP); Panini ; Sanskrit; Semantic Net; Shastric Sanskrit; Vibhakti.

I. INTRODUCTION

"The structure of Paninian Grammar is nothing but a computer program" – Charles Babbage. It has captured the base of universal principles of all languages. Computational Linguistics requires formal rules for analysis and generation of language. Slowly Chomsky and others are turning towards Panini and in the past 60-70 years; much time, effort and money has been expended on designing computers that can work with phrases, expressions, idioms, oratory, rhetoric, ambiguity, vagueness, idiosyncrasy and peculiarity of human language for NLP programming. Understandably, there is a widespread belief that natural languages are unsuitable for the transmission of many ideas that artificial languages can render with great precision and mathematical rigor. These efforts have centered on creating schema designed to parallel logical relations with relations expressed by the syntax and semantics of natural languages which are clearly cumbersome and ambiguous in their function as vehicles for the transmission of logical data[1][2].

At the same time, our ancient Sanskrit linguists, philosophers and grammarians have considered and worked on many complex problems and issues pertaining to natural language understanding, creation and generation.

There is at least one language, Sanskrit which for the duration of almost more than 10,000 years was a living spoken language with a considerable literature of its own. The period of Classical Sanskrit is designated as c. 500 BCE to 1000 CE.

All great scholars of the world like Shakespeare, Voltaire, Briggs, Kant etc consigned to the fact; "Sanskrit is the mother of all Languages". Later many proofs supported the fact that Sanskrit is root of all Indo-European languages.

Besides works of literary value, there was a long philosophical and grammatical tradition that has continued to exist with undiminished vigor until the present century. Among the accomplishments of the grammarians can be reckoned as a method for paraphrasing Sanskrit in a manner that is identical not only in essence but in form with current work in Artificial Intelligence. To examine the possibility of seeking solutions for these problems, this research article demonstrates that a natural language can serve as an artificial language with the help of NLP, Semantic Net, Vibhakti, Dual Case; Inflection based Syntax, Shastric Sanskrit and Equivalence etc.

II. NLP & SANSKRIT

In a marvelous paper "*Knowledge Representation in Sanskrit and Artificial Intelligence"* AI researcher Rick Briggs (1985) observed; "Certain problems of concern in NLP have either been thought about or actually solved in the old tradition of Sanskrit grammar and philosophy. If the grammatical view is not examined, current AI work will go on unnecessarily re - inventing a wheel millennia old"[1]. This development in computer world triggered the Sanskrit Academy to initiate the "Kriya Research Project" under the aegis and protection of Indian Government Department of Electronics. Various Sanskrit Vyakarana Pundits and Professor became involve in

978-9-3805-4421-2/16/$31.00

simplification of basic features of Sanskrit language grammatical system. While designing the software for the system, ancient shastras provide the basic tenets for NLP creation and generation. In 1995, the second phase of this important project "Sabdabodha" came into existence. Parsing of the sentences involves three processes. First it identifies the words, then classify and interconnecting the words in meaningful way to create a relevant sentence. This process is applicable in machine translation and in other methods of handling natural language. Therefore Sanskrit
Academy built up modules or units for further recognition/ identification, generation/creation of diverse forms and analysis of nouns, verbs and participles in Sanskrit.
NLP is correlated with the field of interaction between human being and computers. It mainly entails the challenges like natural or human language understanding (spoken as well as written), so that computers can understand it. Further it can enable computers to obtain exact meaning from input of natural language and others, thereby helping in natural language creation and generation. In general as part of NLU, any language for human communication has three basic tenets [6]:

A. Syntax:

It depicts the basic form of any natural language, typically grammar specified. Here only that language can be best fit for the computer programming or NLP or Robot programming which has sound grammar base [6].

B. Semantics

Here semantics denotes the connotation, significance or real meaning of the words or sentences of the language. Although general semantic theories exist, when we build a natural language understanding system for a particular application, we try to use the simplest representation we can. Usually knowledge base has fixed correlation, association or mapping of word and concepts in the real world [6].

C. Pragmatics

The pragmatic component explains how the expressions or speech or words relate to the world. To understand language, an agent should consider more than the sentence; it has to take into account the context of the sentence, the state of the world, the goals of the speaker and the listener, special conventions, and the like[6].
As per any computational linguistics, any sentence must be syntactically correct, semantically correct and pragmatically right, otherwise NLP programmer will face lots of difficulties. This concept will be more clearly depicted by below given examples;

a) That seminar is about soft computing.
b) The blue birds sing sweet.
c) Drab red thoughts slumber frantically.
d) Frantically slumber thoughts red drab.

Here the first sentence is well formed. While the second sentence is syntax and semantics wise correct, pragmatically it does not sound well. The third sentence semantically does not make sense but it is syntactically right. The last sentence does not make any sense. This is incorrect syntactically, as well as semantically and pragmatically [6]. If we want binary/twofold code/ machine language as the ground of all computer activities, a well controlled, structured and unambiguous method followed in Sanskrit is essential to fulfill the required objectives defined earlier. Although many languages other than Sanskrit have a well defined grammar, still the same word and sentence carry different meanings in different perspectives.
For example, a phrase in English language "I like apple" can suggest a brand of Computer or a kind of fruit apple. In other example, "Do you see the man with the glasses?" suggest from linguistic analysis point of view varied meanings for different things or situations. At one place, it suggests "A man uses the glasses or eyeglasses" ("A device to compensate for defective vision or to protect the eyes from light and dust") or things made of glass ("a hard, brittle substance, typically transparent or translucent, made by fusing sand with soda and lime and cooling rapidly which is used to make windows, drinking containers and many other things") or at some other place, it depicts a drinking pot or container made from glass, e.g. a glass of milk.
In above examples, both sentences are okay grammatically. Their correct meaning varies with respect to their existing contexts or background. These vagueness or ambiguity must be properly dealt by NLP Programmers so that Sanskrit can become the most excellent language suited for NLP.

III. SEMANTIC NET

In Natural Language Processing, Knowledge Representation is the primary step to make machine understandable Natural Language with the use of Semantic Nets. Throgh Semantic Nets concepts in AI, information are symbolized as a group of nodes connected to each other by a set of labeled arcs which depicts associations or relation among the nodes.
The concept of semantic nets is mainly employed to symbolize semantic associations / relations between or among the concepts by using undirected or directed graphs. One needs to reveal the real meaning of the sentence after removing ambiguities in the sentence. People think that it is simple to map one lexical item to other one, for sentence translation so that computer can understand its form. To some extent, they are right. Here in this example; "Rashi gave a pen to Ankit"; the information can be arranged and stored as a set of triple:

"give", "agent", "Rashi"

"give", "object", "Pen"

"give", "recipient", "Ankit"

"give", "time", "Past"

In figure 1, one can observed that vital importance is put on verb (Give) which is treated as distinctive feature of the sentence "Rashi gave a pen to Ankit". 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. In this semantic net diagram, past tense of the defined node and arc has been grammatically changed accordingly.

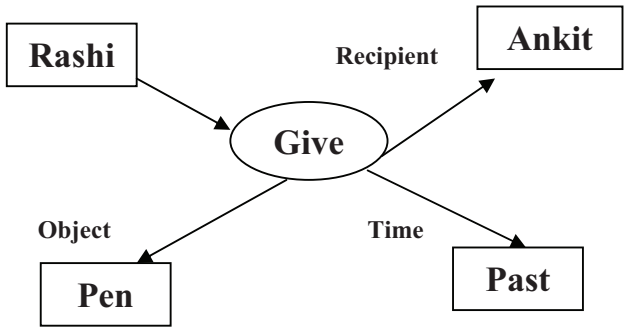

Fig. 1. Basic Semantic Net for "Rashi gave a Pen to Ankit"

English language is natural and unmanageable sometimes as English literature came first, Grammar defined later. While in case of Sanskrit, whole edifice or structure of its literature based on flawless constructs of Sanskrit Grammar since thousands of years. Sanskrit Pundits defined all theoretical, conceptual notions in grammatical sentences. We can take help of Semantic nets to make easy transformation of English sentences.

Words of English when stored in the database just serve as labels. Things like Data, Information and Meaning cannot be retrieved from that word. 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. WORD STRUCTURE & INFLECTION CONCEPT

Word Structures in Sanskrit is in the below format;
<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 [5].

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. Sanskrit is such a highly inflected language that word order almost does not matter. For prose Sanskrit had the preferred word order of Subject-Object-Verb (SOV). For poetry and the like other word orders were used frequently for their effect.

V. VIBHAKTI

The computational grammar described here takes the concept of vibhakti and karaka relations from Panini framework and uses them to get an efficient parse for Sanskrit Text. Vibhakti guides for making sentence in Sanskrit and there are seven kinds of vibhakti [3]. Vibhakti also provides information on respective karaka. These seven vibhktis are:

- "Prathama – Nominative"
- "Dvitiya – Accusative"
- "Tritiya – Instrumental"
- "Chaturthi – Dative"
- "Panchami – Ablative"
- "Shhashhthi – Possessive"
- "Saptami – Locative"
- "Sambodhana – Denominative"

Karaka approach helps in generating grammatical relationship of nouns and pronouns to other words in a sentence. The grammar is written in 'utsarga apavaada' approach i.e. rules are arranged in several layers each layer forming the exception of previous one. Now with the competent use of classes, objects and pointers, we can see that how Sanskrit language makes the use of vibhakti in computer programming language, to make it short, concise and succinct.

For example;

मूर्खः परिहर्तव्यः प्रत्यक्षः द्विपदः पशुः ।

This Sanskrit shlokas suggest; "A stupid person is like a two legged animal in front of the eyes so he must be avoided". This sentence is like "Gagar main Sagar", very economic in selecting and using words. Only few languages of the world can match this effectiveness [4]. It has a distinct property to maintain one-to-one relation between the word/ objects represented and its associated properties.

मूर्ख = (the property of being) stupid
परिहर्तव्य = (the property that makes one) avoidable (by others)
प्रत्यक्ष = (the property of being) in front of the eyes
द्विपद = (the property of) having two legs
पशु = (the property of usually being) tethered

Generally in day to day language, the words/ objects do not reflect attribute/property. We should use our word effectively in vibhaktified form [4].

So, मूर्ख represents the property of being stupid, but मूर्खः (which is a vibhakti of the word मूर्ख) represents an object/person who is stupid.

In the same way, below mentioned word can be understood;

परिहर्तव्यः = an object/person who must be avoided
प्रत्यक्षः = an object/person located in front of the eyes
द्विपदः = a object/creature having two legs
पशुः = an object/creature who is tethered = a beast or cattle

Here all different words signify distinct attributes/properties. Now in figure 2, all five words have been converted into five objects with different associated properties. Here we use the concept of Vibhakti. The guiding principle of Sanskrit grammar states; "Words having the same vibhakti represents the same object i.e. the five different words are like pointers that point to the same object because they all have the same vibhakti" [4].

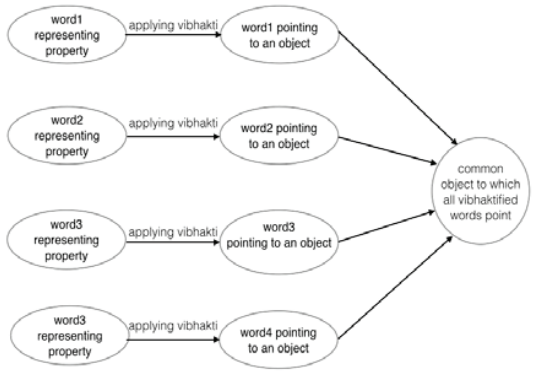

Fig. 2. Vibhaktis Used as Object Pointers

Here the concept of class in OOPS has been used (devoid of function). It has been used as a pointer to an object when used with vibhakti.

VI. DUAL CASE

In comparison to others, Sanskrit is a highly inflected language by way of three grammatical genders (masculine, feminine and neuter) and three numbers (singular, plural and dual). Sanskrit supports this exclusive characteristic of using and handing out singular, plural as well as dual [4]. We have made a comparative study among Sanskrit, English and other languages as mentioned here under;

1) Sanskrit:
"Singular Case: baalkey (in the boy)"
"Dual Case: baalkayo: (between the boys)"
"Plural Case: baalkeshu (among the boys)"
2) English:
"Singular Case: in the boy"
"Dual Case: between the boys"
"Plural Case: among the boys"
3) Spanish:
"Singular Case: en el nino - in the boy"
"Dual Case: entre los chicos- between the boys"
"Plural Case: entre los chicos among the boys"
4) French
"Singular Case: chez le garcon- in the boy"
"Dual Case: entre les garcons- between the boys"
"Plural Case: parmi les garcons among the boys"

In above examples, it has been observed that in comparison of other languages, only Sanskrit makes a comprehensible, plain difference among singular, plural and dual case. Therefore in future, we have an error free Natural Language Processing environment for AI programmers.

VII. DECLENSION FOR FUNDAMENTAL NOUN & ADJECTIVE

As shown in table 1, the fundamental method of suffixation is applicable for approximately every adjectives and nouns. Nevertheless, as per the gender and the last vowel or consonant applicable to the uninflected word stem, we have fixed tenets/rules of required "Sandhi" that finally give the inflected word [9]. For neuter gender, parentheses act as the case terminations, what remains applies to feminine gender and masculine gender.

We have both Devanagari script and International Alphabet of Sanskrit Transliteration (I.A.S.T.) transliterations in the following table for clear understanding. Transliteration allows complete and lossless Romanization of Indic text or scripts including Sanskrit, Pāḷi, Prākṛta and Apabhraṁśa. Here final Visarga '(ḥ)' is alike to's' originated from Sandhi changes [9].

TABLE I. SANSKRITISED SUFFIX FOR ALL NUMBERS

	"SINGULAR"	**"DUAL"**	**"PLURAL"**
"NOMINATIVE" "(*KARTĀ*)"	"-स् –S" "(-म् -M)"	"-औ -AU "(-ई -ī)"	"-अस् – AS" "(-इ -I)"
"ACCUSATIVE" "(*KARMAN*)"	"-अम् –AM" "(-म् -M)"	"-औ –AU" "(-ई -ī)"	"-अस् – AS" "(-इ -I)"
"INSTRUMENTAL" (*KARAṆA*)	"-आ –Ā"	"-भ्याम् – BHYĀM"	"-भिस् – BHIS"
"DATIVE" "(*SAMPRADĀNA*)"	"-ए –E"	"-भ्याम् – BHYĀM"	"-भ्यस् – BHYAS"
"ABLATIVE" "(*APĀDĀNA*)"	"-अस् –AS"	"-भ्याम् – BHYĀM"	"-भ्यस् – BHYAS"
"GENITIVE" "(*SAMBANDHA*)"	"-अस् –AS"	"-ओस् –OS"	"-आम् – ĀM"
"LOCATIVE" "(*AdhikaranA*)"	"-इ –I"	"-ओस् –OS"	"-सु –SU"
"VOCATIVE"	"-स् –S" "(- -)"	"-औ –AU" "(-ई -ī)"	"-अस् – AS" "(-इ -I)"

VIII. SHASTRIC SANSKRIT WORD PRESENTATION THROGH OBJECT's ATTRIBUTES

Sanskrit differs from other language in the sense that it has

one-to-one unique relation or correspondence between the available words and the entities, they stand for. In English, a word tree simply denotes the word tree, not reflecting its associated features, attributes or properties. While in Sanskrit, the word वृक्ष (tree) corresponds to tree's features additionally and not denote the tree itself only. Likewise many other words in Sanskrit that describe the attributes of a tree, signify the same word tree as shown below in the example given [4].

वृक्ष = something that is cut and felled down

तरु = something that floats

पादप = something that drinks using its feet

Here above mentioned words represent the same object (a tree) because a tree has the above properties. This advantage can be fully utilized by NLP programmers.

IX. COGNATE SANSKRIT WORDS AND THE EUROPEAN LANGUAGES

In 1785, Sir William Jones (a Judge) in order to systematize the native law of India (so that Britain could rule India by native law which was logically consistent) started studying Sanskrit and surprisingly, became a linguistic scholar. He found Sanskrit to be a marvelous language as he could guess the meaning of some Sanskrit words from his knowledge of Latin and Greek. After four months of study he wrote and delivered a paper in which Jones (1786) said [7]:

"The Sanskrit language, whatever be its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either, yet bearing to both of them a stronger affinity, both in the roots of verbs and in the forms of grammar, than could possibly have been produced by accident; so strong indeed, that no philologer could examine them all three without believing them to have sprung from some common source, which, perhaps, no longer exists."

Jones became a Sanskrit aficionado and communicated that passion to the intellectual scholar world of Europe of the time through his writings. The instances of cognate words cited are examples like such as *raj* for king in Sanskrit and *rex* in Latin. These are alright but the most wonderful evidence comes from such common words as the names for numbers [7].

TABLE II. SANSKRITISED COGNATE WORDS FOR NUMBERS

Sanskrit Word	Numerals	Greek Word	Latin Word
"éka"	1	"en"	"unus"
"dvá"	2	"duo"	"duo"
"trí"	3	"tria"	"tres"
"catúr"	4	"tessera"	"quattuor"
"páñca"	5	"pente"	"quinque"
"sás"	6	"hexa"	"sex"
"saptá"	7	"hepta"	"septem"
"astá"	8	"okto"	"octo"
"náva"	9	"ennea"	"novem"
"dása"	10	"deka"	"decem"

Sir Jones, whilst studying and learning Sanskrit, visualized the fact that Sanskrit is the originator of all languages. In his book "In the Sanskrit Language", Jones (1786) wrote of how he observed that Sanskrit had a powerful semblance and similarity to Latin, Greek and other important European languages. He also realized the fact that they were not only having the same origin but were related or associated with other Celtic, Gothic and Persian languages.

A big revolutionary thought emerged from his learning that separated language from religion. It also prepared the solid foundation for a better scientific, logical and controlled approach towards linguistics. His ultimate findings are being considered very vital to the mankind. It is being considered equivalent to the scientific invention/ discoveries made by great scientists such as Einstein, Darwin, Newton and Galileo. In table 3, we provide the list of few English words derived from Sanskrit (as English originated and evolved by borrowing and using words from Latin/Greek/Arabic [8].

TABLE III. ENGLISH WORDS ROOTED FROM SANSKRIT LANGUAGE

Core / Root/ Original Words in Sanskrit Language	Word in English Language	Same Word Used in Greek (Gr)/ Arabic (Ar)/ Latin (La)
"Matr (means Mother)"	"Mother"	"Mater (La)"
"Jan (means Generation)"	"Gene"	"Genea (Gr)"
"Aksha (means Axis)"	"Axis"	"Axon (Gr)"
"Navagatha (means Navigation)"	"Navigation"	"Navigationem (La)"
"Sarpa (means Snake)"	"Serpent"	"Serpentem (La)"
"Naas (means Nose)"	"Nose"	"Nasus (La)"
"Anamika (means Anonymous)"	"Anonymous"	"Anonymos (Gr)"
"Naama (means Name)"	"Name"	"Nomen (La)"
"Ashta (means Eight)"	"Eight"	"Octo (La)"
"Barbara (means Foreign)"	"Barbarian"	"Barbaria (La)"
"Dhama (means House)"	"Domicile"	"Domus (La)"
"Danta (means Teeth)"	"Dental"	"Dentis (La)"
"Dasha (means Ten)"	"Deca"	"Deca (Gr)"
"Madhyam (means Medium)"	"Medium"	"Medium (La)"
"Kaal (means Time)"	"Calendar"	"Kalendae (La)"
"Kri (means to Do)"	"Create"	"Creatus (La)"
"Mishra (means Mix)"	"Mix"	"Mixtus (La)"
"Ma (means Me/My)"	"Me"	"Me (La)"

“Pithr (means Father)”	“Father”	“Pater (La)”
"Bhrathr (means Brother)”	“Brother”	“Phrater (Gr)”
“Loka (means Place)”	“Locale”	“Locus (La)”
“Maha (means Great)”	“Mega”	“Magnus (La)”
“Makshikaa (means Bee)”	“Mosquito”	“Musca (La)”
“Mrta (means Dead)”	“Murder”	“Mortis (La)”
“Nakta (means Night)”	“Nocturnal”	“Nocturnalis (La)”
“Pancha (means Five)”	“Penta”, “Five”	“Pente (Gr)”
“Parah (means Remote)”	“Far”	“Pera (Gr)”
“Patha (means Path)”	“Path”	“Pathes (Gr)”
“Raja / Raya (means King)”	“Royal”	“Regalis (La)”
“Sama (means Similar)”	“Similar”	“Similis (La)”
“Sapta (means Seven)”	“Seven”	“Septum (La)”
“Smit (means Smile)”	“Smile”	“Smilen (La)”
“SthaH (means Situated)”	“Stay”	“Stare (La) “
“Svaad (means Tasty)”	“Sweet”	“Suavis (La)”
“Tha (means That)”	“That”	“Talis (La)”

X. CONCLUSIONS

In conclusion, we can say that Sanskrit being well programmed natural language is the most suitable language for Soft Computing areas in Artificial Intelligence and Natural Language Processors. It can be used as high-level language to write programs and to give instructions to advanced robots which are more likely to understand Sanskrit better. We also need highly powerful and robust architecture computers for Sanskrit language testing and implementation. Depth knowledge and learning of Sanskrit must be encouraged for AI & NLP programmers. It also requires huge amounts of money, research and man-power.

XI. FUTURE SCOPE & APPLICATIONS

We must explore the opportunities to develop the Sanskrit language and grammar on priority basis for sustainable & wholesome growth and development of Indian cultural heritage, more willingly than promoting a foreign language. Besides these are interesting research areas to explore;

- WorldNet need for Sanskrit language.
- Sanskrit texts conversion from text to speech.
- Web based Hypertext archive/files/records for Sanskrit Literature.
- Easy Panini Grammar for English language.
- Panini Grammar system validity checking through computing.
- Good quality search engine for Sanskrit E-library.

REFERENCES

[1] Rick Briggs, *Knowledge Representation in Sanskrit and Artificial Intelligence*, DOI: http://dx.doi.org/10.1609/aimag.v6i1.466

[2] Vaishali Ravindranath, *Sanskrit - the most suitable language for computer linguistics*, ISBN 9789381992968 Article number TRF_123.

[3] 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.

[4] Deeptanshu Jha, Dr. Rashmi Jha & VarunVarshney, *Natural Language Processing and Sanskrit* **,** International Journal of Computer Engineering & Technology, Volume 5, Issue 10, October 2014, ISSN 0976-6367(PRINT) ISSN 0976-6375 (ONLINE), pp. 57-63.

[5] Divya Teja & Sasidhar Kothuru*, Sanskrit in Natural Language Processing,* International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X ,Volume 5, Issue 3, March 2015, pp. 596-600.

[6] http://artint.info/html/ArtInt_290.html

[7] http://www.notablebiographies.com/supp/Supplement-Fl-a/Jones-William.html

[8] http://www.hitxp.com/articles/culture/sanskrit-greek-english-latin-roman-words-derived-pie-proto-indo-european-language/

[9] https://en.wikipedia.org/wiki/Sanskrit_nouns