International Journal of Sanskrit Research 2023; 9(4): 104-106

International Journal of Sanskrit Research

अनन्ता

ISSN: 2394-7519
IJSR 2023; 9(4): 104-106

www.anantaajournal.com
Received: 23-06-2023
Accepted: 27-07-2023

Abhimanyu Rao
Shiv Nadar School, New Delhi, India

Corresponding Author:
Abhimanyu Rao
Shiv Nadar School, New Delhi, India

Exploration of the ideal sanskrit grammar for natural language processing: An analysis of laukik and Vedic Sanskrit

Abhimanyu Rao

DOI: https://doi.org/10.22271/23947519.2023.v9.i4b.2179

Abstract
Prior to the writing of this paper, there has been a lot of research into the applicability of Sanskrit as a language for programming. This further expanded into research about how applicable it is for Natural Language Processing (NLP). Reasons cited include its syntax and almost mathematical grammar, along with its lack of exceptions. This made Sanskrit a major candidate in the field of natural language processing. However, since then, research in the field has not progressed further, and we are yet to see Sanskrit being used in natural language processing. Furthermore, Sanskrit is a language with a rich history, dating back several thousand years, and it has evolved a lot since its conception. This gave way to even more possibilities of the usage of Sanskrit, since there were ancient and modern forms of the language. Therefore, researching the ideal version of Sanskrit for usage in Natural Language processing was chosen as a topic, and Vedic(ancient) and Laukik(modern) versions of the language were compared. After conducting research, it was found that Laukik Sanskrit seems to be a more refined version of the language. It has eliminated several ambiguities, and streamlined the grammar. This led to the belief that Laukik Sanskrit is more suitable for Natural Language Processing.

Keyword: NLP, sanskrit, grammar, vedic, laukik

Introduction
What kind of Sanskrit is better to use as an NLP language? To answer this question, we must first understand what NLP is. Natural language processing, or NLP, is a method through which humans hope to make technology understand and interact with language in the same way as humans themselves. This means that they wish to forgo two things; confusions and miscommunications, and specialised knowledge of Artificial intelligence communication. An example of this can be when one looks at their own virtual assistants, like Siri. Often, when asking a long or confusing question, there is a very high chance of misinterpretation by Artificial Intelligence [1]. This leads to unhelpful, or otherwise unrelated responses. The search for a proper language for NLP comes from this issue. If a suitable language is found, the usage of that language could prevent the aforementioned issues.

This is where Sanskrit comes into consideration. For years now, it has been argued that Sanskrit makes an ideal language for an NLP. There are a few reasons for this. Firstly, Sanskrit has a more "mathematical" syntax. This means that there is not much interpretation needed to correctly modify Sanskrit words. They generally follow very simple structures with a root word, prefixes and suffixes. All these structures have generalised rules which stem from one of three gender and singular-plural choices. This makes it a language that would be easier to decode computationally than something like English, which is full of exceptions. There is also a very beneficial lack of 4 main ambiguities in the Sanskrit language; Semantic, lexical, pragmatic, or structural [2].

Lexical ambiguities arise in a language when one word can have several meanings, which makes an exact understanding of a sentence difficult. Especially as artificial intelligences are not very good with understanding of context and rely much more on the words themselves. Sanskrit has a much lower lexical ambiguity than most other languages [3]. Secondly, there is a pragmatic ambiguity.

This occurs when there is a lack of clarity in sentences like “A loves his wife and so does B”. In this sentence, it is unclear as to whether B loves his own wife or A’s wife. Sanskrit clears that up by ensuring that possessives are very distinct and so such confusions don’t occur [4]. Semantic ambiguity is an ambiguity about what the speaker is referring to. This ambiguity is battled by Sanskrit's strict rules of tense and positioning, which mean that it is always clear what is being referred to in sanskrit. Lastly, a structural ambiguity can occur when two or more words together form a structure that could refer to multiple things, like seeing a crane fly could mean either seeing the bird, a crane, fly, or see a crane fly, which is an insect [4].

Reasons such as its lack of major changes, “mathematical” syntax, and lack of ambiguity, be it semantic, lexical, pragmatic, or structural, have made it a frontrunner in the research of Natural Language processing. In fact, Zoho, an Indian billion-dollar company, has become the first of its kind to use Sanskrit based software [5]. However, despite multitudes of research being done in the field of using sanskrit as an NLP, not much has been done to compare the different types of the language as a factor of effectiveness for the same purpose [6, 7]. This paper aims to analyse the two types of sanskrit, वेदिक and लौकिक and come to a conclusion as to which works better to facilitate natural language processing.

Methodology

The aim of this research is to understand the impact of different types of Sanskrit grammar, Vedic and Laukik, on its suitability for Natural Language Processing (NLP).

In the discussion section the differences between Vedic and Laukik Sanskrit are discussed on the basis of the nature of grammar and phonetics.

In order to gather, synthesise and evaluate the findings of all relevant data on the subject and to present a thorough overview of the current state of knowledge on this subject, this study used a systematic literature review approach as a research methodology [8]. To reduce bias and ensure reliability and validity of the review, this procedure involves using a systematic and open approach to locate, assess and summarise all important papers.

Results and Discussion

As discussed previously, the usage of Sanskrit as an NLP is a very likely scenario. Sanskrit has a very strict syntax, with minimal exceptions, making it a good choice for NLP usage. There is a structured combination system for words to exact a combined meaning, known as sandhi. This structured process with a defined set of rules means that it is easy to assume the meaning of a combined word simply by knowing the meanings of its parts. This is not so for several languages, including English, where compound words have their own meanings. Take the example of the word Grandparents. It is a combination between the words grand, that means magnificent, and parents. The combination means the parents of the parents, which does not mean the same as its root words. On the other hand, take the example of the sanskrit word दुश्शसन, which is a combination of the words दुः, which means difficult, and शसन, which means control/rule. The word दुश्शसन means difficult to control/rule, which is exactly the sum of what the root words mean. This means that the word meanings database for sanskrit need not be as extensive for sanskrit, since the meanings of several words can simply be deduced by knowledge of their parts.Additionally, most words in sanskrit have fixed roots with variable prefixes and suffixes that can add to the meaning. This makes it very systematic to analyse the meaning of Sanskrit words, once again indicating that knowledge of the rules and basic root words can allow for significant understanding of the language [9].

However, this paper is not about the benefits of Sanskrit as a language for Natural language processing. It aims to discuss the differences between Vedic and Laukik Sanskrit, and ultimately come to a conclusion as to which one is better suited to be used for Natural language processing. To understand this, this paper will be discussing the lexical differences between the two languages.

Firstly, vedic sanskrit has one additional tense form known as लेत लकार. This form of tense is no longer used in laukik sanskrit. On one hand, this indicates lack of necessity, since this is not needed for the language to exist, as it can function without it. On the other hand, the availability of more conjugations can lead to more versatile communication for natural language processing, indicating that it might be beneficial for an extra tense to be available [10].

The concept of svarabhakti existed in vedic sanskrit. The concept of svarabhakti, also known as epenthesis, was used in vedic sanskrit as well. This concept, still used in several languages, including English, is the insertion of small sounds between two letters. Take the example of the word thimble, which is pronounced ‘θɪm.bəl’, despite there being no vowel between b and l [11]. This concept was no longer used in laukik sanskrit. Since Sanskrit is generally spelled phonetically, the existence of svarabhakti would likely hinder the natural language processing, since it would be more difficult to discern the spelling of the word by the sound alone.

Vedic sanskrit also had an additional vowel ऌ, which still exists, but is rarely used in laukik sanskrit in words like ऌत्, meaning simple future [10].

Vedic sanskrit included one extra articulation style of vowels known as plut swar [10]. Plut swar are not used in laukik sanskrit, since there are no words that involve their usage. ओ३म् is an example of a word that uses plut swar, but that writing style is merely to depict the pronunciation, with it being symbolised by ‘ॐ’.

The usage of suffixes is another interesting aspect of vedic sanskrit. Vedic sanskrit includes many synonymous suffixes, where countless suffixes added the same meaning to a word. For example, the suffix तुमुन् has 14 synonymous suffixes, and क्त्वा has 3 [10]. This adds a redundancy, since if one meaning can be provided by several words, those words are unnecessary, and only one of those words is truly needed. Having several such words can unnecessarily clutter a database.

Vedic sanskrit also possessed the usage of the लङ् and लुङ् लकार in any tense, while in laukik sanskrit, those are specific to past tense [10]. For example, while the sentence “पुरुषः अहसत्” clearly means “The Man Laughed” (simple past tense) in laukik sanskrit, in vedic sanskrit, the tense would not be clear, and additional classifiers would have been necessary. Having tense specific conjugations makes it easier for a program to discern the tense of a verb. Vedic sanskrit lacked clear rules for the reflexive verbs, while laukik sanskrit has clear rules for these [10].

Conclusion

This paper discusses the suitability of two different versions of sanskrit for natural language processing. It should be noted that while both versions of sanskrit share much of their grammar, the differences between the two are clear enough to imply that one would be preferable over the other for usage in Natural Language Processing.

Laukik Sanskrit appears to be the clear choice to pursue for Natural Language Processing. There are fewer ambiguities in this version of the language, making it simpler for a computer to comprehend and manipulate the language. This lack of ambiguities leads to a more logical flow of language, and since computers operate on language, such a logical flow will lead to more natural linguistic synthesis.

Unfortunately, this research was not without limitations. The sources needed were few and far between, which led to major setbacks and several delays in the writing of this paper. This has led to dependency on certain sources for much of the paper, and the lack of comparable research has led to the belief that there is still much to be done with the topic.

Future opportunities of research could begin with a deeper comparison of the two languages. It could be followed by a practical test of the hypothesis, which can only be possible once an NLP program that operates solely on Sanskrit has been made.

Acknowledgement

I would like to thank my mentor and my parents, Sapna Khajuria and Rajshekhar Rao for supporting me. My Sanskrit teacher, Mr. Deepak Dabral, for inciting a passion for the language in me and for supporting me in writing this paper. I would also like to thank Mrs. Mahesh Kakkar, Professor Ranjit Behera, and Ms. Madhuchandadi Anondodhoni for supporting my endeavours.

References

1. Hoy MB. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. Medical Reference Services Quarterly; c2018 Jan 12.
2. Deshpande DS, Kulkarni N. A Review on various approaches in Machine Translation for Sanskrit Language. International Journal of Future Generation Communication and Networking. 2020 Jan;13(2):113-120.
3. Kak SC. The Paninian approach to natural language processing. International Journal of Approximate Reasoning; c1987, 1(1). https://www.sciencedirect.com/science/article/pii/0888613X87900077
4. Bathulapalli C, Desai D, Kanhere M. Use of Sanskrit for natural language processing. International Journal of Sanskrit Research. 2016;2(6):78-81.
5. India's first billion dollar company using Sanskrit to build machine translation software. Infomance; c2023 Jun 22. Retrieved July 29, 2023, from https://www.infomance.com/zoho-indias-first-billion-dollar-company-using-sanskrit-to-build-machine-translation-software/
6. Saxena S, Agrawal R. Sanskrit as a Programming Language and Natural Language Processing. Global Journal of Management and Business Studies. 2013;3(10):1135-1142.
7. Huet G. Towards Computational Processing of Sanskrit. French Institute for Research in Computer Science and Automation; c2003.
8. Snyder H. Literature review as a research methodology: An overview and guidelines. Journal of Business Research; c2019, 104. https://doi.org/10.1016/j.jbusres.2019.07.039
9. Lowe JJ. The syntax of Sanskrit compounds. Language. 2015 Sept;91(3):71-115.
10. तिवारी, ड. भ. (2012). भाषा विज्ञान.
11. THIMBLE | Pronunciation in English. (2023, July 26). Cambridge Dictionary. Retrieved July 31, 2023, from https://dictionary.cambridge.org/pronunciation/english/thimble