Abstract

This paper briefly enlists the features of Sanskrit language and suggests the use of the same for Natural Language Processing studies and applications. The potential predicted has to be verified by computer experts. Certain advantages of Sanskrit mentioned may find use in some of the frontier areas of Computer Engineering Research, notably in AI and Knowledge-based systems.

The view-point expressed here is that a Sanskrit-based Compiler or interpreter may have to be developed to unearth the hidden treasures in Sanskrit technical literature. Other countries in the West also, of late, have undertaken similar studies and it would only be appropriate if Sanskrit gets the type of recognition that it so richly deserves in its own land even in areas of advanced technological research, for which it is undoubtedly suited intrinsically.

Hardware and Software aspects involved in translating the suggestions contained here is for the Computer experts to consider while the Author would remain at the disposal of such interested researchers/organisations for further inputs and/or discussions/clarifications etc.

I. Introduction

Transaction Processing involves Receipt, Storage, Manipulation/Processing, Transfer and Retrieval of Information. Electronic Data Processing has a wide range of applications these days owing to the enormous strides the field has taken in recent years, both in terms of hardware & software developed and available to common man. Speed, flexibility, computing power, peripherals and interfaces have virtually extended the scope of computers to almost every facet of human endeavour.

Routine (mechanical) processing functions have given way to intelligent processing systems and high level languages are nearing natural language-like structure ceaselessly. Object-oriented programming languages are turning even symbolic programming languages obsolete: Programs tend towards expert systems. Such is the galloping progress of Computer Engineering (hardware & software) that, it is the time to think of an almost as-old-as-the-world language of the Orient, namely, Sanskrit, in all its majestic glory, to be considered as a candidate for computer programming, in the fields of Natural Language Processing & Artificial Intelligence.

Be it knowledge representation or speech synthesis, natural language processing or machine translation, intelligent tutoring systems or unambiguous semantic extraction, study of complex mathematical problems or linguistics, in virtually any field, one can think of utilising the richness, strength, accuracy, efficiency, structure, flexibility and the extant works available in the Sanskrit language.

---

* *Aircraft Design Bureau, ADA (Systems),*
C.V. Raman Nagar, Bangalore-560 093

One of the prerequisites for such an ideal blend of the oldest and gramatically best structured natural language and the newest of the emerging topics of computer world is an appreciation of the real potential of both sides of the equation. In fact, recent efforts at bringing together Sanskrit and computers show encouraging results. In 1984, Dr. Rick Briggs of USA highly recommended Sanskrit as 'the' candidate for machine translation interlingua. In 1985 & 1986, conferences were held at Bangalore to correlate data on perception, knowledge representation & inference, as found in science and shastras through Sanskrit. Ever since, many institutions at home and abroad, have engaged in research in this area and have brought new ideas to the fore.

In a recent national conference on ''Samskritam and Computers'' held at New Delhi, the signal was clear that concerted efforts to enlist all interested persons from scholarly and computer backgrounds for a national directory to be prepared to ultimately create computer programming environment using Sanskrit extensively, was very desirable.

In a paper presented at the said conference, the author had dealt with the concept of using Sanskrit as the natural language for AI-related, particularly, knowledge based systems, citing the case of determining sentential import.

A few fundamentals of the suggestion would be described here. The fundamentals dealt are: Definition & characteristics of knowledge, Phonetics, Analysis of parts & forms of speech, Flexibility in word formation, structure of Grammar, Disambiguation rules/procedures/criteria, variety and richness of technical literature and Etymological & exegetical aspects, as found in oriental works.

II. Definition & Characteristics of Knowledge

Knowledge is of two types, i.e. knowledge about self and that about other (non-self) things. The former is termed as ''Consciousness'' while the latter is a quality of it. Generally, the latter is meant when we talk of knowledge since the former is not a subject of anything other than itself. The Consciousness is sentient, i.e. possesses knowledge, animate, i.e. takes on inanimate bodies, and always aware of itself as one in number and pleasant and favourable to itself. That is to say, for any living being, the self is known automatically always, the Self or life is most desirable and all actions are intended to make the Self happy. This Consciousness is also called variously as Awareness, I-ness, Self, Soul etc.

Knowledge is an attribute of sentient (conscious or living) being. It is a matter existing in the dual forms of particle and waves (like light). It manifests in many forms. Scriptures list many 'states' of knowledge like Cognition, Inference, Happiness, Sorrow, Wish, Love, Hate, Envy, Effort, Determination, Doubt, Mistake, Decision, Religious Belief, Courage, Shame, Intelligence, Fear, Emotion, Avarice, Ignorance, Arrogance, Tranquility, Comprehension, Retentivity, Recollection, Logic, Devotion and Science. Knowledge always goes with its subject, i.e., knowledge of 'what?'. The subject of knowledge may be any object described by a name and form.

'Illumining' with an agent or object is an essential characteristic of knowledge. So also are unsurpassed Speed, Subtlety, Lightness and self-luminence while presenting objects. Knowledge is non-sentient, i.e, it does not possess knowledge. That is to say that knowledge does not know itself. The two concepts of self-illuming and sentience are to be clearly understood. A lamp is self-illumining in that it does illuminate itself as well as other objects. But it does not know itself, which is sentience, and hence we say lamp is non-sentient.

Knowledge is akin to a lamp in this sense. In contrast, Consciousness is sentient, i.e. it knows itself or it has knowledge as an attribute.

Since as yet there are no ways invented of creating life, we cannot aim to create knowledge. In AI, therefore, a simulation of an intelligent human activity can be attempted. For this also, Analysis in Metaphysics holds valuable clues. Here, all objects (things with name and form) are classified into two categories as Instruments or Sources of knowledge and Subjects of knowledge. The instruments of knowledge are of three types, viz. Sensory Perception, Inference and Verbal Testimony. The first of these could be simulated for applications like Robotics, while the last two are useful in AI research. The representation and usage of these according to Oriental works is well worth a try.

III. Phonetics

One of the significant advantages of Sanskrit is that the grammer ensures total precision and guards against ambiguity, mis-spelling and mis-pronounciation as the meanings are bound to get altered otherwise. The real advantage is that since the correlation between written and spoken forms is one is to one, the two forms of input can be exchangeably used. The analysis of alphabets (characters) is based on sound production from well-defined places of utterance and hence, comprehensively and clearly covers all possible cases.

Sound emanating from distinct places of utterance have been identified, e.g., Throat (guttural), Palate (palatals), Head (cerebrals), Teeth (dentals), Lips (labial), Nose (nasal), root of the tongue and a combination of these. Besides origin of sound, all syllables have other attributes like effort (internal and external) required to pronounce, timing (short, long and prolated), accessories/instruments necessary, accents (acute, grave and circumflexed), connotation (vowel, consonant, conjunct), etc., fix a given literal with its meaning. Besides this, all valid words have proper derivation/deduction from finite set of well-grouped verb-roots and noun-bases so that what is meant is uniquely determined and accuracy ensured. This concept is not opposed to that of flexibility in word formation as will be explained later in this article.

Speech synthesis can possibly benefit by this feature immensely since accent, frequency, emphasis and timing oriented discrepancies associated with other natural language speech inputs are absent here. Context and intention related choice of the right meaning when more than one possible meanings can be taken is to be treated separately as in any other natural language case, but even here, there are some precise guidelines available as will be clear later.

IV. Analysis of Parts & Forms of Speech

The Annexure brings out the various categories of words in Sanskrit and accordingly, a matrix of all possible, valid word forms can be automatically generated by formalisation of the grammar rules. This is an area for quite in-depth study and research through application programs, best written in Sanskrit, if the present suggestion is realisable. A detailed discussion of the parts of speech is not called for here, and about the forms also, it is enough to note that there are nine distinct types identified (as in Annexure) and this greatly enhances the utility of a natural language for computer interaction. The grammer has simple, effective rules for conversion between various forms and sentences.

V. Flexibility in Word-Formation

All valid word-forms have two significant parts, namely, the stem or substrate (verb-root for verbs and nounbase for noun-forms) and the affixes (prefix and/or suffix). These are also indicated in the Annexure. With respect to the admissible combinations of substrates and affixes, there are elaborate but clear rules specifying these with the attendant changes in the meanings denoted, the latter being derivable strightaway. This is the admirable blend of rigid framework but flexible application, often likened to a golden ingot that lends readily to a myriad forms of ornaments through malleability and ductility while retaining strength. It is remarkable that across the Universe, there cannot be distortions in Sanskrit either written or spoken, according to Paninian Grammar and violations of it will be transparent at once to the linguists.

VI. Structure of Grammar

Panini's structure of Sanskrit Grammar has won the admiration of linguists of all ages and regions. The grammar is sound based in that as explained earlier, throat, tongue, head, teeth, lips, nose etc. are identified as origins of various literals and in dealing with such analytically formed sounds, other parameters enumerated earlier are used effectively.

All technical literature in Sanskrit have a fundamental set of 'Aphorisms', which are short, pithy, versatile sentences that capture the concepts thoroughly. These are termed as "Sutras". (A Sutra by definition, is a statement that has minimum literals, i.e. very brief, unambiguous, purposeful, universal & versatile, i.e. can apply to a large number of cases by generalisation, free from errors, inadequacy, fillers, etc.). Grammar rules are also in this 'Sutra' style which greatly condenses the amount of instructions or information to be given to precisely convey a particular aspect. Hierarchy, Sequence, Precedence, Priority, Normal & Exception rules, Grouping, use of Designators or Operators (control characters), explicit instructions for interpretation in particular cases, generalisation, Chaining, Restriction, Extension, convention etc. impart the needed flexibility while maintaining the control on validity of word-formation.

The Grammar rules are contained in eight chapters with four quarters per chapter. They total a little less than 4000 in number, the longest of them containing no more than 90 alphabets! But within this, the total array of all possible, valid word-formations, are covered comprehensively with no ambiguity whatsoever. The verb-roots are separately listed and number a little above 2000. All declensions of noun-forms are categorised into about 225 different types based on gender and ending character (vowel, consonant), including special forms. Each noun take on 24 forms (3 numbers and 8 cases), viz. Nominative, Vocative, Accusative, Instrumental, Dative, Ablative, Possessive or Genitive and Locative. The numbers are singular, dual and plural. While an exhaustive list of Nouns has to be obtained from other sources (like the famous Lexicon- 'Amarakosa' etc.), Panini has laid down general rules for determination of gender etc.

Verb-roots are grouped into ten categories each having a given group suffix, besides verb forming suffix, added to the roots to form verbs. Clasification of verbs is quite elaborate and there are six tenses and four moods in which verbs can be expressed. Besides, there are many special forms derivable from verb-roots. Verbs used to denote action for self and for others have different connotations, as also those that go with either or both the types. There are 3 numbers as in the case of Nouns and verb-roots can take one of 3 persons (analogous to cases), i.e. first, second and third (I, you and It). Verbs are invariant with gender.

There are other interesting word-forms such as Governance or functional Clauses, the combined and compound types with the attendant rules for operation on them. These are the ones that require utmost

ingenuity to formalise and represent in Computer. Also, these are intention-dependant and require context information for interpretation. This is a very challenging area for research and has a potential to achieve maximum brevity possible in a given situation. Certain word forms that are invariant with gender, cases, numbers, roots, etc. are called Indeclinables and these are also dealt with effectively in Panini's Rules. The emphasis here is not on details but to highlight the scope for computerised research to benefit from. After all, there are celebrated works analysing very critically these sutras and we should be able to adopt them usefully before it is late. The advantages of traditional form of study in Sanskrit could then be had by all.

VII. Disambiguation Rules/Procedures/Criteria

The absence of syntax in Sanskrit is a definite plus point in its favour. The semantics also can be extracted by well laid out procedures. Here, apart from Grammar, the rules of Syllogism (a branch of study dealing with logic) and Mimamsa (study of Scriptural Texts) are utilised. Mechanism of associating meanings with words is dealt with in detail and guidelines for establishing meanings at word, sentence and discourse levels are given. Accordingly, various criteria like Expectancy, Proximity, Compatibility, Primary and Secondary Denotation, Implication, Context, Suggestion, Relevance, Place, Time, Commencement, conclusion, Repetition (Emphasis), Novelty, Utility or Objective, Figurativeness, Propriety etc. are indicated for getting at the meanings intended to be conveyed.

Within the rules of a particular branch of study also, rules to guide priority, conflict handling, exceptions, special cases etc. are well-defined to ensure precision and accuracy. Also, there are numerous illustrations provided in later works to explain the fundamentals contained in Sutras, which are invariably authored by great Sages. These works include commentaries, treatises, notes and expositions. The extent of technical literature available as of today also is by no means small and one should impartially analyse them for what they are worth with an open mind. When the same language is used for such a study also, it would be very appropriate and original, as the ideas can be captured precisely.

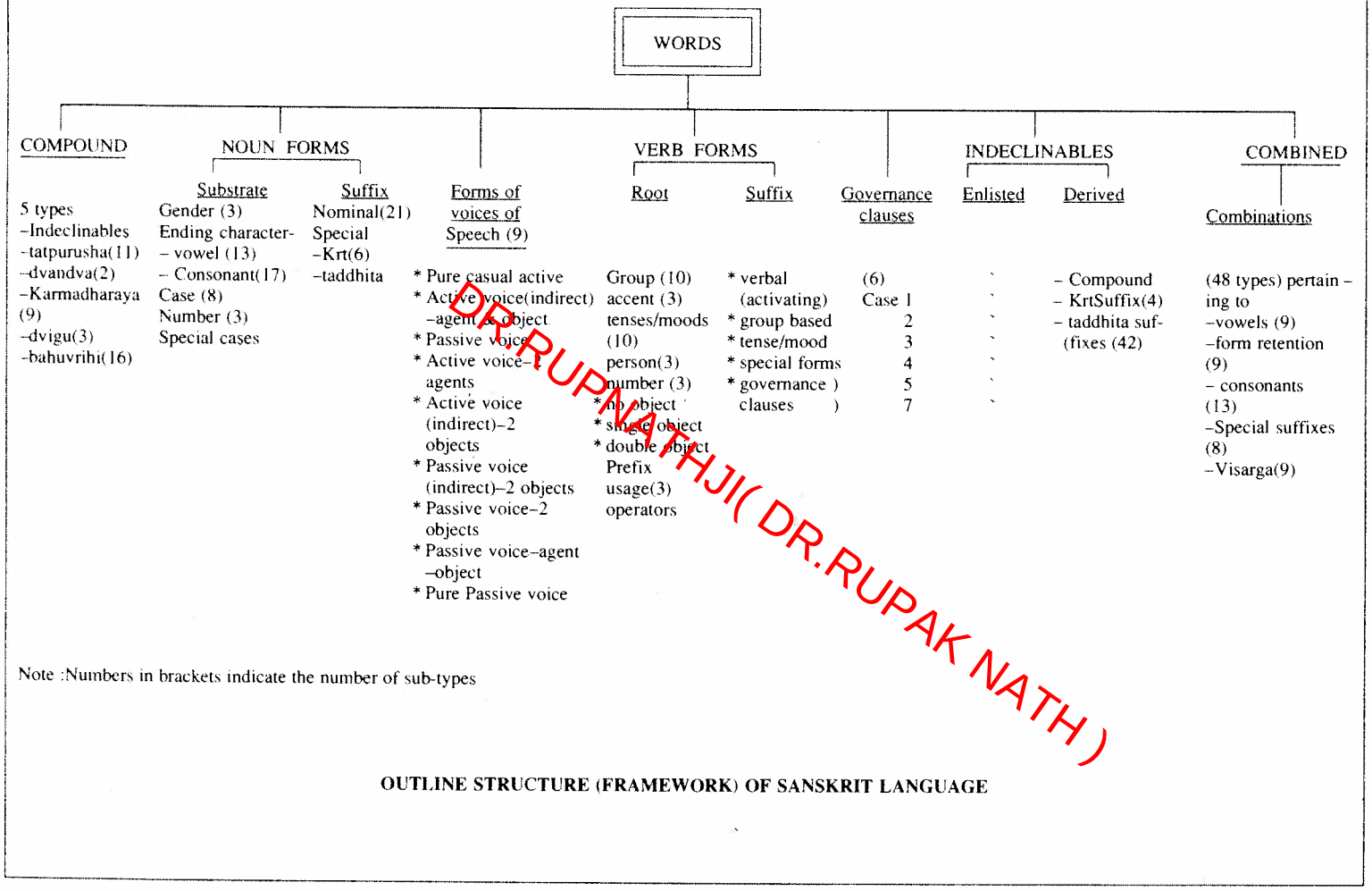
VIII. Variety and Richness of Technical Literature

Technical literature in Sanskrit comprises of 14 branches of Learning. These are the four Vedas (Scriptures), Rig Veda, Yajur Veda, Sama Veda and Atarvana Veda, the six vedic auxiliaries, namely, Phonology, Grammar, Prosody (Metrics), Etymology, Astronomy and Ceremonial Directory (Ritualry), Study of Vedic Texts, Syllogism (Logic), Epics and Codes of Moral Rectitude. There is a great treasure of knowledge contained in these in an efficient and streamlined manner. Discounting controversial claims and exaggerations, we need to dispassionately study these and take what is worth.

There are other branches of study dealing with matters like Vedic mathematics (which is currently undergoing a great resurgence), Astrology, Chemistry, Medicine, Aeronautics, Metallurgy, Warfare & Weaponry etc. which are less popular and more depleted in terms of available, useful literature, but neverthless require a tool like the one suggested to tap the contents properly. It is not an easy task either, but concerted efforts have no barriers that are unsurmountable.

Besides technical literature, there are numerous literary works that enrich the language particularly and Linguistics in general. For intelligent Tutoring etc. this may prove quite useful. Thus a study of Sanskrit literature using Sanskrit in a Computer seems to deserve consideration.

Annexure



Note :Numbers in brackets indicate the number of sub-types

OUTLINE STRUCTURE (FRAMEWORK) OF SANSKRIT LANGUAGE