POS Tagging for the Primitive Languages of the World and Introducing a New Set of Universal POS Tagging for Sanskrit

Anupam Das, Bidisha Choudhury, and Shikhar Kumar Sarma

Abstract The digital structuring of a language depends on how their parts of speech are classified to make the language the most compatible for the user. The parts of speech (POS) of a language is defined as the word classifiers which classify a word more precisely for using it in the sentence properly for expressing the emotions or feelings of human beings through natural languages. The natural language processing in the research arena already classifies the exhaustive forms of POS of many languages in the world. Among all such languages, the POS of SANSKRIT language is also introduced by many eminent scholars. Here, an attempt is made to give the new set of POS of the SANSKRIT language with their proper definitions and explanations with examples and they are termed as Universal POS (UPOS). The set so defined that universally everybody can understand, and the further use of these tags can be made available to the researchers of the world. The new set is highly expected to be the most exhaustive form.

Keywords UPOS-universal POS · Primitive language · Word order · Compound

1 Introduction

In ancient times, human beings communicated their emotions and feeling or their instructions through gestures and some primitive oral sounds. Though it is difficult to give the exact period of time when the structured language came to the human society but linguists claimed around 10,000 years ago; the existence of the structured language evidenced through scriptures. The first structured language is really debatable as different linguists claimed differently. During the studies of linguists,

A. Das (✉)
Department of CSE, Royal Global University, Assam, India
e-mail: adas_arya@rediffmail.com

B. Choudhury
Department of Sanskri, Visva Bharati University, West Bengal, India

S. K. Sarma
Department of IT, Gauhati University, Assam, India

S. Fong et al. (eds.), *ICT Analysis and Applications*, Lecture Notes in Networks and Systems 517, https://doi.org/10.1007/978-981-19-5224-1_3

one thing came out as evidential that structured languages are used in their first appearances in texts formed and its use as formal verbal forms in contemporary era. Table 1 gives the five oldest languages claimed by many linguists without any further controversies [1].

The parts of speech (POS) of any language are the detailed word classifier. The best set of POS of a language can be claimed only when the POS covers the best possible sentence composition with its fullest meanings. The smallest entities used in each sentence are distinctly defined as the POS with their specific usages [1]. Table 2 gives the number of POS of the top 10 most popular languages of the world:

In the table, it is observed that most of the languages have their POS in very limited numbers with standard Chinese has 13 as maximum and Arabic has 3 as minimum. All these languages are structured by the grammarians with the proper

Table 1 Primitive languages and its existence

Ranks according to the evidence of scripts	Language name	Period of existence claimed	Evidence got as scripts	Development of parts of speech tags (POST)
1	Sanskrit	2nd millennium BC	2000 BC	2023 (including sub-tags)* introduced in this study
2	Greek	2000 BC	1500 BC	165 (including sub-tags)
3	Chinese	1500 BC	1250 BC	34 (including sub-tags)
4	Hebrew	1200 BC	1000 BC	23 (including sub-tags)
5	Tamil	2500 BC	300 BC	71 (including sub-tags)

Table 2 Primitive languages and their number of POS tags and number of SPEAKERS

S No.	Language name	No. of POS	Speakers
1	Standard Chinese	13 = 10 content words + 3 function words	1.1 billion
2	English	8	983 million
3	Hindustani	8	544 million
4	Spanish	9	527 million
5	Arabic	3	422 million
6	Malay	4	281million
7	Russian	9 = 4 main + 5 minor	267 million
8	Bengali	8	261 million
9	Portuguese	10	229 million
10	French	8	229 million

definitions of POS. The scopes of new classification of words in these languages are closed, and thus, they can be categorized as the closed languages. A language can be termed as open language, where the scope of inserting new set of POS is opened. The SANSKRIT falls under the open category language. There are many researchers tried to insert new set of POS into this language and successfully did it. In this paper, we tried another attempt to introduce a new set of POS called UPOS in the most structured language SANSKRIT. Here, we are not only inserting new set of POS but also digitizing all the POS with their respective tagging.

1.1 New Set of POS in Sanskrit

The POS in Sanskrit is crafted by many grammarians and linguists. There are certain issues which are not dealt with those works: like sub-tagging in main tags is not classified, and also sub-tag like case and case-ending was $b =$ not discussed. There are 100 main tags and total 2023 main tags along with their sub-tags are classified in this work.

1.2 Analysis of UPOS in Sanskrit

According to the sub-features of all the above tags are calculated, then the total number of UPOS will be as shown (Figs. 1, 2, 3, 4, 5, 6, 7).

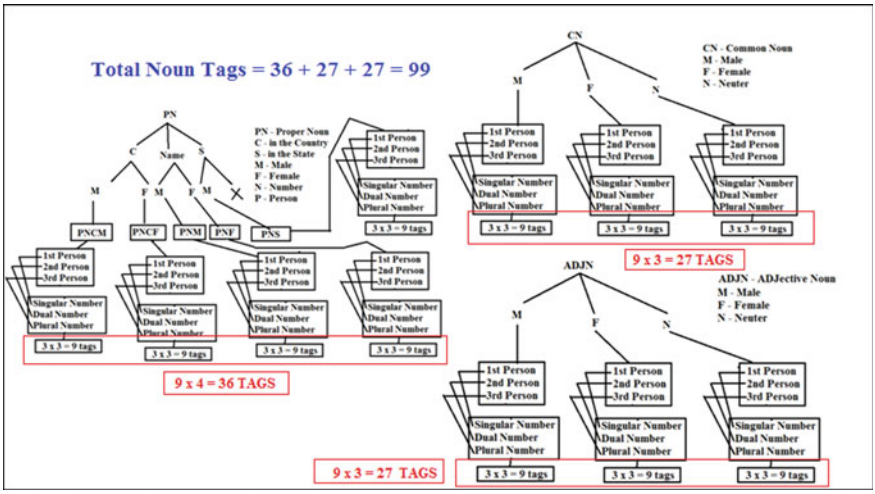

Fig. 1 Noun tags

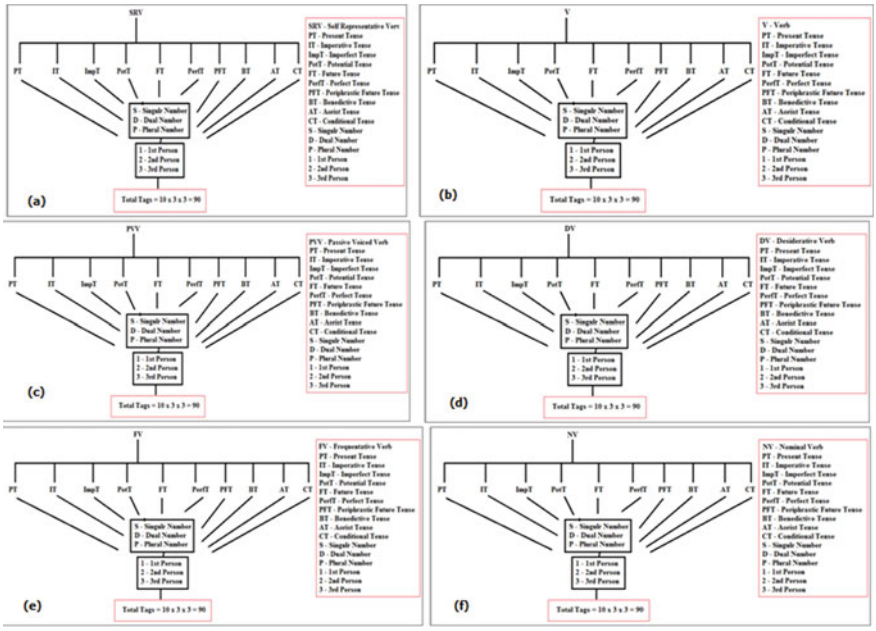

Fig. 2 Verb tags

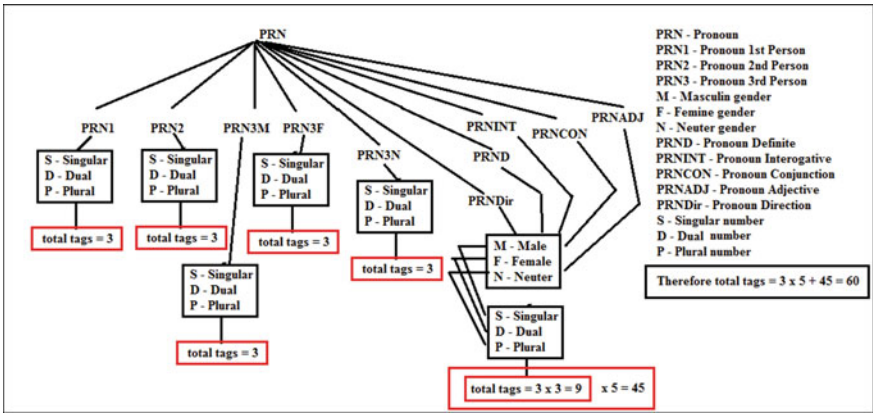

Fig. 3 Pronoun tags

2 Comparative Study of the Previous Work

Table 3 gives a comparison of the previous works with this present work: [2–10].

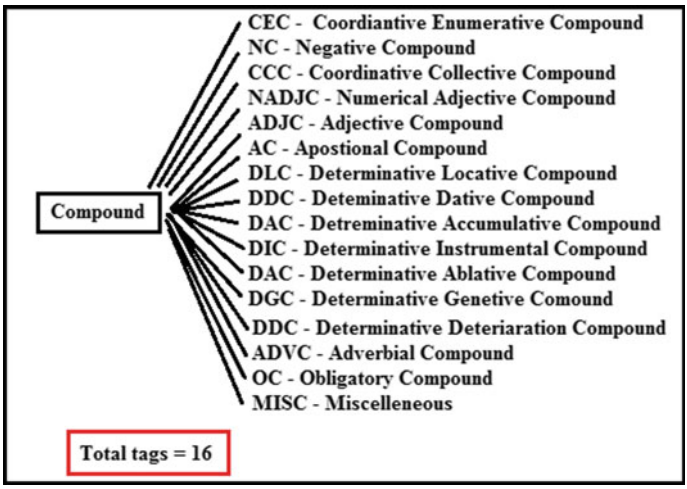

Fig. 4 Compound tags

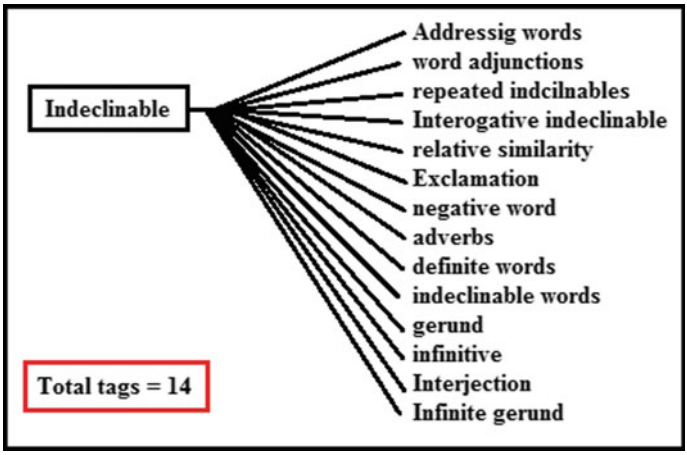

Fig. 5 Indeclinable tags

3 Analysis of Primitive Languages

3.1 Word Order

From Table 4, it is observed that the word order is most flexible in Sanskrit language with 83.33% and then Tamil language with 50%. Though the most of the structural form of used sentences are SOV in Sanskrit and Tamil languages, both these

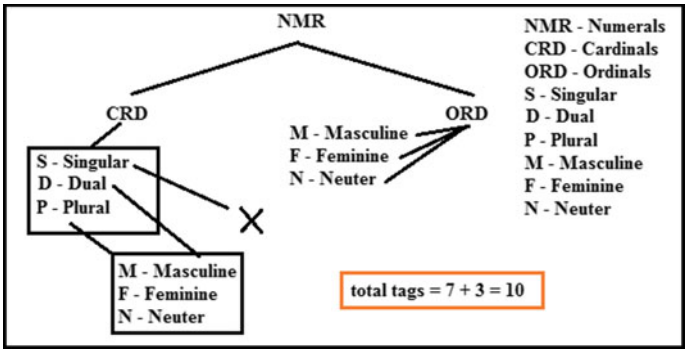

Fig. 6 Numeral tags

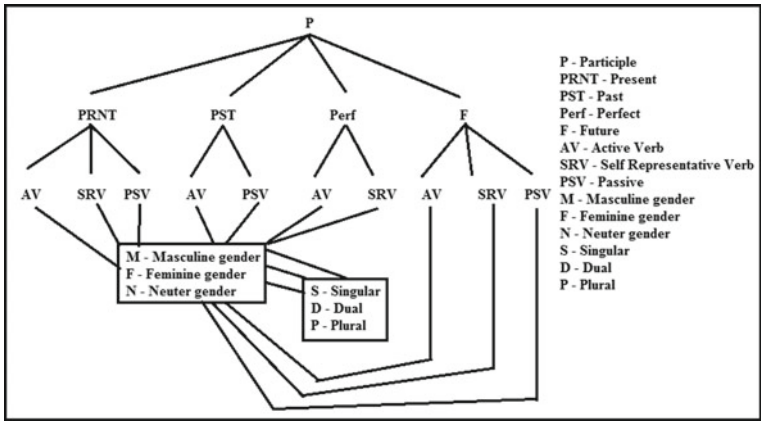

Fig. 7 Participle tags

languages are also compatible in the other word order formats mentioned in the table. But other languages are not compatible with other word order formats.

3.2 The Compound

The compound as defined by the linguistics as a combination of words which are carrying a special meaning. Also it can be defined as a lexeme that consists of more than one stem. The rule for forming a compound is compounding, composition or nominal composition. Table 5 gives the compound for the first five oldest languages.

Table 3 Comparative study of POS in Sanskrit

Development of POS	Institute contribution	Total tags described
JPOS in 2007	JNU	134 (main and sub)
Tirupati POS in 2013	Tirupati University	134 (main)
SAN POS in 2015	Gujrat University	51(main)
IL-POST	Microsoft Research India Lab	28 (main)
Sastra Univ Post for Sanskrit in 2015	Sastra Univ Thanjavar India	11(main)
ILMT	IIIT Hyderabad	26(main)
CPOS IN 2017	JNU	134(main)
Oliver Hellwig in 2016 (new)	Dusseldr of University	136(main)
SANS post in 2019	IIT Kharagpur	130(main)
UPOST in 2022	Royal Global University, Visva Bharati Uinversity and Gauhati University	2023(main and sub)

Table 4 Word order comparison of primitive languages

Languages	No. of POS	SOV	SVO	VSO	VOS	OVS	OSV	WO (%)
Sanskrit	2023	Yes	Yes	Yes	Yes	Yes	No	83.33
Greek	165	Yes	No	No	No	No	No	16.66
Chinese	34	No	Yes	No	No	No	No	16.66
Hebrew	23	No	No	Yes	No	No	No	16.66
Tamil	71	Yes	Yes	No	No	No	Yes	50.00

Table 5 Comparison of compounds in primitive languages

S No.	Name of the languages	No. of compounds
1	Sanskrit	16
2	Greek	9
3	Chinese	5
4	Hebrew	4
5	Tamil	7

3.3 The Basis of Grammar in Primitive Languages

The basis of grammar is based on the factors of a language like (Table 6).

Table 6 Basis of grammar of Sanskrit and other primitive languages

Basis of language	Sanskrit	Greek	Chinese	Hebrew	Tamil
Character set (alphabet)	42	24	Around 6500 mandrains	22	30
No. of vowels	9	5	Not specified	5	12
No. of consonants	33	19	Not specified	17	18
Number	3	2	2	3	2
Order of sentences	5	1	1	1	3
Tense	10	3	Not specified	Not specified	3
Verb mood	4	3	3	2	3
Gender	3	3	Not specified in grammar but preserved in tone	2	2
Part of speech (POS)	2023	165	34	23	71

3.4 The Main and Sub-Tag Classification

In this work, the main tags are classified in its maximum form, and the sub-tags are thoroughly classified, and finally, there are 100 main tags and its associated sub-tags are classified and in results it is found total 2023 tags (Table 7).

Table 7 Main and sub-tag classification of Sanskrit

S No.	POS category	Main POS tags	Sub-POS tags	Total
1	Noun	11	63	693
2	Verb	6	90	540
3	Pronoun	10	63	630
4	Indeclinable	14	–	14
5	Participle	10	9	90
6	Numerical	4	3	10
7	Compound	16	–	16
8	Nipatan	1	–	1
9	Prefix	1	–	1
10	Designation	1	–	1
11	Other language	1	–	1
12	Doubttag	1	–	1
13	Punctuation	25	–	1
Total		100		2023

Table 8 Common and uncommon sub-tags along with their main tags

S No.	POS name	Main tags	Sub-tags				
			Gender	Person	Number	Tense	Case and case-ending
1	Noun	11	X	3	3	X	7
2	Verb	6	X	3	3	10	X
3	Pronoun	10	3	1	3	X	7
4	Indeclinable	14	X	X	X	X	X
5	Participle	10	3	X	3	X	X
6	Numerical	4	3	X	X	X	X
7	Compound	16	X	X	X	X	X
8	Nipatan	1	X	X	X	X	X
9	Prefix	1	X	X	X	X	X
10	Designation	1	X	X	X	X	X
11	Other language	1	X	X	X	X	X
12	Doubttag	1	X	X	X	X	X
13	Punctuation	25	X	X	X	X	X

3.5 The Main Tags with Their Common and Uncommon Sub-Tags Classification

There are 100 main tags are identified in this work and sub-tags are enlisted with the corresponding main tags (Table 8).

4 UPOS Interface Designed for Creating the Database for Sanskrit Language

To create a database for the words of Sanskrit, an interface is designed for preparing the corpus.

The following snapshots are the interface of UPOS (Fig. 8).

5 Observations and Findings

i. According to the evidence, Sanskrit is the oldest language.
ii. The most popular languages fall in the category of closed language, whereas the ancient languages including Sanskrit are open languages.

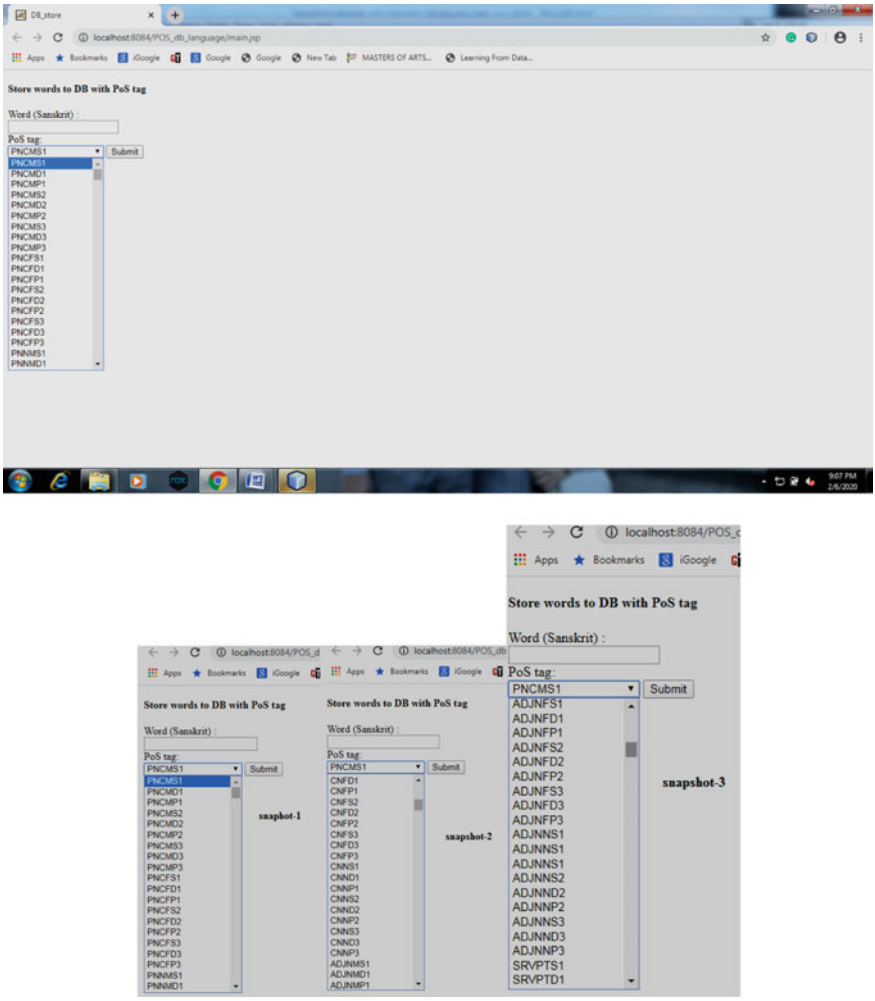

Fig. 8 Snapshots of the interface

iii. The popular languages have boundary-based POS, whereas ancient languages have much more scope of inserting new formation of words in a sentence.
iv. In other previous works, POS tagging: main and its associated sub-features are not discussed except in JNU works.
v. On the other hand, these discussed sub-tags in the previous researches were not clearly analyzed and categorized. As a result, it causes huge confusion in making corpus of the language.
vi. Due to this kind of POS set, word overlapping might occur frequently.
vii. With incomplete POS set, the training data for machine learning will be insufficient, and thus, machine could not perform cent percent.

viii. Word orders in primitive languages are restricted in one or two forms, but in Sanskrit, the word order is more flexible as it supports five forms out of possible six forms.
ix. Sanskrit is considered as the free flow sentence formation language as it supports more than 80% of word order as compared to 50% in Tamil and less than 17% in other primitive languages.
x. The lexeme of Sanskrit language is more because it contains more number of compounds as compared to other oldest or primitive languages.
xi. The communication in human languages is more structured when the time factor or tense is incorporated during their verbal formation of sentences. It is observed in Sanskrit, Greek and Tamil tenses are defined, whereas in Chinese and Hebrew languages, the tense forms are not specified.
xii. Among Sanskrit, Greek and Tamil, Sanskrit presents wide range of tense formation as compared to the Greek and Tamil.
xiii. In Table 7, it is given the more précised form of main and its associated sub-tags.
xiv. Table 7 also identifies more accurately the sub-tags along with their main tags.
xv. Table 7 also helps to prepare the UPOS in more accurate form.
xvi. The most important fact observed here is the sub-tags category case and case-ending which was not discussed in all other previous works.
xvii. To prepare the most efficient corpus of any language depends on the proper definitions and classifications of sub-tags.
xviii. The Sanskrit language can be considered as the most structured language as it contains elaborate classification of main and sub-POS tags.
xix. It is also observed that all the main tags do not have common sub-tags. This phenomenon was not properly shown in the previous works.
xx. Table 7 shows the sub-tags classifications with their main tags. The common and uncommon sub-tags are clearly given in this table.

6 Future Work

Using the database created in this work, a further study will be carried out for making corpus of Sanskrit language which will be a treasure for the researchers of the near future. The corpus will be made available to the world by making them open access to the world.

7 Conclusion

The present work attempted to analyze the primitive languages and developed an interface for making a good database for the Sanskrit language using the new set

of POS called UPOS. The study also pointed out the basic features of the primitive languages of the world. Some interesting findings are observed about the POS set of Sanskrit during the discussion. Hope the present work will give some insight to the primitive languages of the world and also provide an enhance elaborative set of POS set to the linguists.

References

1. https://www.fluentin3months.com/most-spoken-languages/
2. SanskritTagger, a stochastic lexical and POS tagger for Sanskrit. In: Hellwig O
3. A Thesis (2005) The effects of part-of-speech tagsets on tagger performance. In: MacKinlay A, of Melbourne University
4. Hellwig O (2002) Sanskrit und computer. Ph.D. thesis, Freie Universit¨at Berlin
5. Huet G (2007) Shallow syntax analysis in sanskrit guided by semantic nets constraints. In: International workshop on research issues in digital libraries
6. TagMiner: a semisupervised associative POS tagger elective for resource poor languages. In: Rani P, Pudi V, Sharma DM
7. Banko M, Moore RC (2004) Part-of-speech tagging in context. In: Proceeding of COLING
8. Bharati A, Misra Sharma D, Bai L, Sangal R (2006) AnnCorra: annotating corpora guidelines for POS and chunk annotation for indian languages. Tech. Rep. TRLTRC-31, Language Technologies Research Centre, IIIT, Hyderabad
9. Bhatt R, Narasimhan B, Palmer M, Rambow O, Sharma DM, Xia F (2009) A multi-representational and multi-layered treebank for Hindi/Urdu. In: proceeding of the third linguistic annotation workshop, pp 186–189
10. Part-of speech tagger for Sanskrit: a state of art survey. In: Adinarayanan S, Sri Naren N. J Assistant Professor, SASTRA University Thanjavur