# Advanced Tamil POS Tagger for Language Learners

**M. Rajasekar, A. Udhayakumar**

***Abstract - In the emerging technology Natural Language Processing, machine translation is one of the important roles. The machine translation is translation of text in one language to another with the implementation of Machines. The research topic POS Tagging is one of the most basic and important work in Machine translation. POS tagging simply, we say that to assign the Parts of speech identification for each word in the given sentence. In my research work, I tried the POS Tagging for Tamil language. There may be some numerous research were done in the same topic. I have viewed this in different and very detailed implementation. Most of the detailed grammatical identifications are made for this proposed research. It is very useful to know the basic grammar in Tamil language.***

***Keywords- Natural Language Processing, Machine Translation, Parts of Speech Tagging, POS Tagger for Tamil.***

## I. INTRODUCTION

The Part of Speech (POS) Tagging is an important process in the field of Natural Language Processing. In the computational linguistics part-of-speech tagging also called as grammatical information tagging is the process of assigning grammatical tag to every word of the given sentence. POS Tagging is one of the harder process in Natural Language Processing. Because some words have more than one grammatical tag (POS tag) in some different places. Example, book will come as noun in one place and comes as verb in another place.

*The Book (noun) is on the table and Ramu book(verb) the tickets for Robo 2.*

Most of the NLP researchers have already tried the POS tagger by implementing different concepts. In English language, commonly there are nine parts of speech. noun, pronoun, verb, adverb, adjective, preposition, article, conjunction, and interjection. In viewing the previous research approaches about POS Tagging, the part of speech is distinguishing from 42 to 150 for English Language. The POS Tagging is an important process in natural language parsing, machine translation, speech reorganization, information retrieval and other computational linguistics development.

## II. POS TAGGING IN TAMIL

Tamil is one of the Dravidian languages and longest surviving languages in the world. It has very classical literature, has been documented for over 2000 years. And Tamil is a morphologically very rich. Tagging a grammatical information to a word is very complex. Because the word structure is very much complex. The words are in Tamil made with a root word with or without one or more affixes.

**Revised Manuscript Received on August 01, 2019**

**Dr. A. Udhayakumar,** Professor and Controller of the Examinations at Hindustan Institute of Technology and Science, Chennai, India,

**M. Rajasekar,** Research Scholar at Hindustan Insitute of Technology and Science, Chennai, India.

To make a POS tagger for Tamil language is very challengeable. The mail challenges in Tamil POS Tagging are solving complexity in word structure and ambiguity of words [1].

## III. OBJECTIVES

The main objectives are to make an improved POS tagger for Tamil Language Learners. We made an analysis on Tamil classical grammar, collected actual part of speech in Tamil language and used it for POS Tagging. Some of other goals are:

- To provide machine aided POS Tagger in Tamil with improvement.
- To make a tool to help the students to learn Tamil grammar easily
- To make a helpful tool for Tamil language learners.
- To make the computational advancement in Tamil linguistic research

## IV. RELATED WORKS

Various concepts already exist for POS Tagger in Dravidian languages. For Tamil language A rules-based POS Tagger was developed by Arulmozhi et al, 2004[2]. A POS Tagger for Classical Tamil was developed and tested by R. Akilan, et al, 2012[3]. A POS Tagger and Chunker for Tamil was developed by Dhanalakshmi V et al, 2013[4]. And a Hybrid POS Tagger for Tamil was developed by Arulmizhi et al, 2006[5]. This system is developed by using HMM technique and a rule based system. These existing concepts are mainly focused on some similar methods, mostly rule-based. There are some generalized tag sets are also developed. Namely AUKBC, Vasuranganathan tag set, CIIL tag set, and Amrita POS Tag set. These all tag sets are developed with focus on English general tag sets. We have concluded some problems with these tag sets.

1. Every tags are generated as English language tags only.
2. Tag sets are not defined as deep, though in Tamil language the grammatical information is much varied when comparing with English tag sets.
3. The Tag sets are limited; it is not describing the Tamil words in detailed.

## V. BUREAU OF INDIAN STANDARDS (BIS) TAG SETS

The Bureau of Indian Standards (BIS) Tagset has authorized a common tagsets for Parts of Speech Tags for Indian Languages on 2010[6]. Most of the experts in the area of Natural Language Processing have involved generating this tagsets. The research works related to the POS Tags must follow these BIS Tagsets. We are also followed and generated the main tags from this BIS Tagsets. The BIS Tagsets for Tamil is shown below.

| S. No | Main Tag | Sub Tags |
|---|---|---|
| 1. | Noun | Common, Proper, Nloc |
| 2. | Pronoun | Personal, Reflective, Relative, Reciprocal, Wh-word |
| 3. | Demonstrative | Deictic, Relative, Wh-word |
| 4. | Verb | Finite, Non-Finite, Verbal Participle, Relative Participle Verb, Conditional Verb, Infinitive Verb, Gerund, Verbal Noun, Auxiliary |
| 5. | Adjective | |
| 6. | Adverb | |
| 7. | Preposition | |
| 8. | Conjunction | Coordinator, Subordinator |
| 9. | Particles | Default, Classifiers, Interjection, Intensifier, Negation |
| 10. | Quantifiers | General, Cardinals, Ordinals |
| 11. | Residuals | Foreign, Symbol, Punctuation, Unknown, Echo words |

**Table 1. BIS Tagsets for Tamil**

## VI. PROPOSED TAG SETS

We need a tag sets to give fully grammatical information for Tamil Literature. It should be in basic level, to satisfy all the grammar rules in Tamil language. This stimulates me to develop our own HITS POS Tagset for Tamil Language. The proposed Tagsets for Tamil language are as follows:

| S. No | Tag | Description | Details | Details in Tamil |
|---|---|---|---|---|
| 1. | <NW> | Natural Word | Iyar Chol | □□□□□ □□□ |
| 2. | <IDW> | Indirect Word | ThiriChol | □□□□□ □□□□□ |
| 3. | <DW> | Direction Word | ThisaiChol | □□□□□ □□□□□ □ |
| 4. | <FW> | Foreign Word | VadaChol | □□□□□ □ |

**Table2. Noun Tags (Literature view)**

| S. No | Tag | Description | Details in English | Details in Tamil |
|---|---|---|---|---|
| 1. | <NT> | Noun of Things | PorulPeyar | □□□□□□□□□□□ |
| 2. | <NP> | Noun of Place | Idappeyar | □□□□□□□□□ |
| 3. | <NDY> | Noun of Date/Year | Kaalappeyar | □□□□□□□□□□ |
| 4. | <NPS> | Noun of Parts | ChinaiPeyar | □□□□□□□□□□□ |
| 5. | <NQ> | Noun of Qualities | Kunappeyar | □□□□□□□□□□ |
| 6. | <NA> | Noun of Action / Verbal Noun | ThozhilPeyar | □□□□□□□□□□□ |

**Table 3. Noun Tags (Grammar view)**

| S. No | Tag | Description | Details in English | Details in Tamil |
|---|---|---|---|---|
| 1. | <FP> | First Person | Thanmai | □□□□□ |
| 2. | <SP> | Second Person | Munnilai | □□□□□□□□ |
| 3. | <TP> | Third Person | Padarkai | □□□□□□□□ |
| 4. | <SMS> | Superset Male Single | Aanpaal | □□□□□□□ |
| 5. | <SFS> | Superset Female Single | Penpaal | □□□□□□□□ |
| 6. | <SP> | Superset Plural | Palarpaal | □□□□□□□□ |
| 7. | <SUS> | Subordinate | Ondranpaal | □□□□□□□□□□ |
| | | Single | | |
| 8. | <SUP> | Subordinate Plural | Palavinpaal | □□□□□□ □□□□ |

**Table 4. Pronoun Tags**

| S. No | Tag | Description | Details in English | Details in Tamil |
|---|---|---|---|---|
| 1. | <DV> | Direct Verb | TherinilaiVinaimutru | □□□□□□□□ □□□□□□□□□□ |
| 2. | <IV> | Indirect Verb | KurippuVinaimutru | □□□□□□□□ □□□□□□□□□□ |
| 3. | <VF> | Verb Finite | Vinaimutru | □□□□□□□□□□ |
| 4. | <VIF> | Verb Infinite | Vinaieccham | □□□□□□□□□□ □ |
| 5. | <PRT> | Present Tense | Nigazhkaalam | □□□□□□□□□□ |
| 6. | <PT> | Past Tense | IrandhaKaalam | □□□□□□□□□□ |
| 7. | <FT> | Future Tense | EthirKaalam | □□□□□□□□□□ |

**Table 5. Verb Tags**

| S. No | Tag | Description | Details in English | Details in Tamil |
|---|---|---|---|---|
| 1. | <PSM> | Participle Male | an,aan | □□□, □□□ |
| 2. | <PSF> | Participle Female | l, aal, i | □□□, □□□, ஐ |
| 3. | <PPH> | Participle Plural Human | ar, aar pa, maar | □□□, □□□, ㄴ □□□□ |
| 4. | <SNHP> | Single Non-Human Participle | Thu | □□ |
| 5. | <PNHP> | Plural Non-Human Participle | Thu un | □□, □□□ |

**Table 6. Participle Tags**

| S. No | Tag | Description | Details in English | Details in Tamil |
|---|---|---|---|---|
| 1. | <AWD> | Attrib. Word Doubler | IrattaiKilavi | □□□□□□ □□□□□ |
| 2. | <AWC> | Attrib. Word Chains | Adukkuthtodar | □□□□□□□□□ □□□□□ |
| 3. | <AWCO> | Attrib. Word Coining | PuNarchi | □□□□□□□□□ |
| 4. | <ACO.AD) | Attrib. Word Coning, Addition | Thondral | □□□□□□□ □□□□□□□□□ |
| 5. | <ACO.AL> | Attrib. Word Coning, Alteration | Thirithal | □□□□□□□ □□□□□□□□□ |
| 6. | <ACO.DL> | Attrib. Word Coning, Delete | Keduthal | □□□□□□□ □□□□□□□□□ |

**Table 7. Attribute Tags**

| S. No | Tag | Description | Details in English | Details in Tamil |
|---|---|---|---|---|
| 1. | <DL> | Demonstrative Letters | SuttuEzhuthukkal | □□□□□□□□□□□□□□□ □□□ |
| 2. | <IL> | Interrogative Letters | VinaaEzhuthukkal | □□□□□□□□□□□□□□□□ □□□ |

**Table 8. Special Letters Tags**

| S. No | Tag | Description | Details in English | Details in Tamil |
|---|---|---|---|---|
| 1. | <PC> | Punctuate.Comma | KaaLpulli | □□□□□□ □□□□ |
| 2. | <PSC> | Punctuate. Semi colon | Aaripulli | □□□□□□ □□□□□ |
| 3. | <PCO> | Punctuate. Colon | Mukkalpulli | □□□□□□ □□□□□□ □□ |
| 4. | <PFS> | Punctuate. Full Stop | Muttruppulli | □□□□□□ □□□□□□ □□ |
| 5. | <PQM> | Punctuate. Question Mark | Vinaakkuri | □□□□□□ □□□□ |
| 6. | <PEXM> | Punctuate. Exclamation Mark | Viyappukkuri | □□□□□□ □□□□□□ □ |
| 7. | <PDQ> | Punctuate. Double Quotation | IrattaiMerkolKuri | □□□□□□ □□□□□□ □□□□□□ |
| 8. | <PSQ> | Punctuate. Single Quotation | OttraiMerkolKuri | □□□□□□ □□□□□□ □□□□□ |
| 9. | <PB> | Punctuate. Bracket | Adaipukkuri | □□□□□□ □□□□□□ □ |
| 10. | <PHM> | Punctuate. History Mark | Varalaatrukkuri | □□□□□□ □□□□□□ □□ |
| 11. | <PHY> | Punctuate. Hyphen | OttraiSamakkuri | □□□□□□ □□□□□□ □ |
| 12. | <PPS> | Punctuate. Plus Sign | Siluvaikkuri | □□□□□□ □□□□□□ |
| 13. | <PSM> | Punctuate. Star Mark | Natchatthirakkuri | □□□□□□ □□□□□□ □□ |
| 14. | <PBR> | Punctuate. Braces | IrattaiInaippukkuri | □□□□□□ □□□□□□ □□□□□□ □ |

**Table 9. Punctuation Tags**

These tags sets are defined in details of Tamil Grammar as completely. These tags may come as single or combined. There are 52 root tags in HITS Tagset. The HITS Tagset is mostly focused on Tamil literature. It covers most of the grammatical definition in Tamil language.

## VII. ARCHITECTURE OF TAMIL POS TAGGER

As we discussed about the proposed POS Tagger for Tamil, the overall system architecture of POS Tagger is shown in the following:

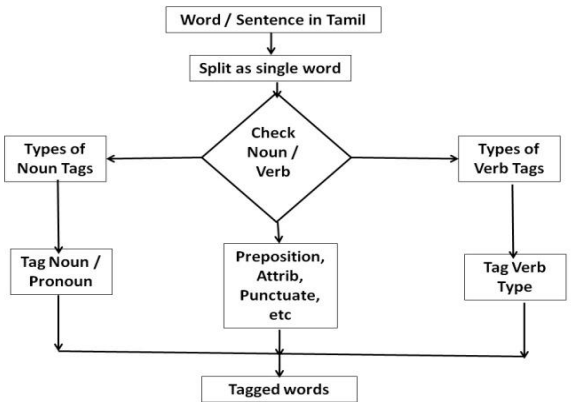


**Figure 1. Overall Architecture**

In the above figure the POS Tagger architecture is showed. At first we have to give the word or sentences in Tamil, as input. The system will split the sentences into separate number of words for its future process. Then it checks whether it is noun or verb or other components in the grammar. Then it will forward the words into its own process. Then each of the POS tagging will be done with its own tagging machine. Finally we get the exact output for the given words or sentences.

### A. System Description:

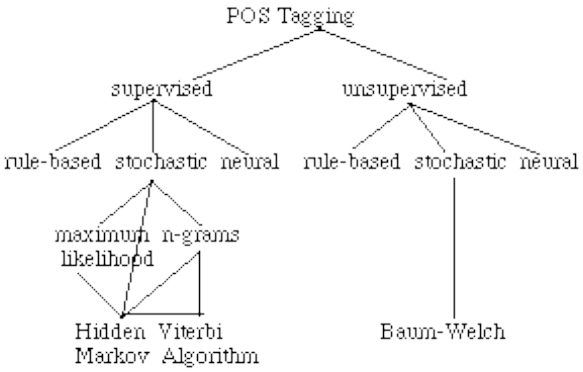


**Figure 2. Approaches in POS Tagging**

There are three types of approach in POS Tagger development. 1. Rules Based 2. Stochastic and 3. Hierarchical approach. From these three types of approach, we have preferred the rules based approach to design the overall system. The steps followed in the core system are,

*Step 1:* The system gets input from the end user as word or sentences.

*Step 2:* it will find the input is word or sentence by checking the whole input with the corpus annotation. If it is there means it will show the Tagged information of the given word. If it is not available in the corpus, it will go for chunking process.

*Step 3:* In the chunking phase, it will split words from the given sentences.

Then it will check word by word from the corpus annotation.

*Step 4:* In this phase, every word will be checked at first with noun corpus. Then it will go for Verb corpus. Then it will go for other adjective, adverb, all other corpus. If the tag set is found in anyone of the corpus it will finish the checking process for that particular word. Finally it will show the tagged words with the tag sets.

### B. Tagger Development:

We have developed a POS Tagger End user environment to Interact with the POS Tagger. It is purely based on Embedded with the Web technologies. It can be used in any kind technological devices. We have used the HTML with PhP Script as development core, and the MS Access as the data storage. The front end user interface has Tamil keys as in webpage. The front end view is shown in the following figure.


**Figure 3. Front End**

This POS Tagger front view is very much comfortable for the users they can easily type Tamil words.


**Figure 4. Front end 2**

### C. Output of the POS Tagger:

By using the user friendly POS Tagger, we can easily type Tamil words, as well as the result of the Tagged set of words for the given input. The following Figure shows that the output of the given words.


**Figure 5. Output of POS Tagger**

### D. Corpus Development:

To produce this POS Tagger system, we need to develop such a huge parallel corpus in Tamil – English language, with its appropriate POS Tagsets. I have developed the Parallel corpus contains around 1.8 lakhs of root words with POS Tagsets. When we pass to Morphological Analysis phase these root words will generate 15 times more morphemes with its POS tagsets. But we have focused on detailed grammatical tagsets for the Tamil Words in our corpus. The Morphological Analysis of a particular word is following process for POS Tagging. Noun and Verb are have been regenerated as morphs. It will be available as Root + Prefix + Infix + Suffix + Stem +Etc. based on the Tense, Person, it will vary from one to another.

For Example,
The noun, Kannan will be generated as,

1. Kannan Ai
2. Kannan Aal
3. Kannan ukka
4. Kannan ukkaga
5. Kannan udaya
6. Kannan in
7. Kannan idam

## VIII. TESTING OF POS TAGGER

The developed POS Tagger has been tested with some set of words for its accuracy. Some of the examples were given below:

□□□□□□□□ ➔ □□□□□□□□□□□ ➔ <NT>
□□□□□□□□□□□□➔□□□□□□□□□ ➔ <NP>
□□□□□□□□□□➔□□□□□□□□□□□ ➔ <NA>
□□□□□□□ ➔ அ➔ □□□□□□□□□□□□ ➔<DL>

Like this we have tested around 10,000 root words for its accuracy. It shows 97.04% of accuracy when compared with manual POS Tagging for the same words.
When comparing with other POS Tagger for Tamil we have tagged more number of words with its correct form of POS tagsets. We have improved with deep grammatical definitions for Tamil words.

## IX. RESULTS AND ANALYSIS

The POS Tagger for Tamil language is developed as a try to help the Tamil Language Learners to understand the Grammatical POS Tagging. The proposed method is implemented with the set of tags assigned manually. The system will check each word in the given sentence and find out the exact Tag. The is tested with set of documents contains the following number of words. The evaluation result of our POS Tagger is shown in the following table. We have evaluated as states. The analysis of the evaluation is given in the chart.

| Word Type | Noun / Pronoun | Verb / Adverb | Attributes / Preposition / Others | Punctuation |
|---|---|---|---|---|
| Tested | 4578 | 3967 | 1098 | 45 |
| Correct | 4423 | 3812 | 997 | 42 |
| Accuracy | 96.61 | 96.09 | 90.80 | 93.33 |

**Table 10. Test and Evaluation**

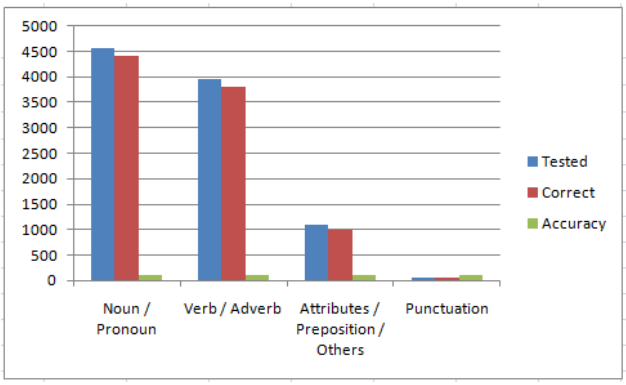


**Figure 6. Analysis Chart**

## X. CONCLUSION

This paper describes the improved POS tagger for Tamil language efficiently. In the corpus around 1.8 lakh words has been used. The system tested and compared with manual POS Tagging.

It is a good accuracy about 96%. This will be help for further development in Machine Translation approach.

## REFERENCES

1. Dhanalakshmi V, Anandkumar M, Shivapratap G, Soman, K P, Rajendran S, *Tamil POS Tagging using Linear Programming*, In International Journal of Recent Trends inEngineering, May 2009, 1(2):166-169.
2. Arulmozhi et al , A rule based POS tagger for Tamil, 2004.
3. R. Akilan, E.R. Naganathan , POS Tagging for Classical Tamil Texts, International Journal of Business Intelligent, Vol: 1 No.1, January-June 2012.
4. Dhanakshmi V et al, POS Tagger and Chunker for Tamil Language, 2013.
5. Arulmozhi P and Sobha L, A hybrid POS tagger for Tamil using HMM technique and a rulebased system, 2006.
6. Nithish Chandra, Sudhakar Kumawat and Vinayak Srivastava, Various Tagsets for Indian Languages and their performance in POS Tagging, Proceedings of 5th IRF Conference, Chennai, 23rd March 2014.

## AUTHORS PROFILE

M. Rajasekar is a Research Scholar at Hindustan Insitute of Technology and Science, Chennai, India. He has been working as a System Analyst since 2008. He has completed his MCA from Anna University, Chennai. He has been working in the research area of Natural Language Processing.

Dr. A. Udhayakumar is a Professor and Controller of the Examinations at Hindustan Institute of Technology and Science, Chennai, India, He has been teaching MCA subjects as well as mathematics since 1989. He has been published more than 40 research paper in Mathematics and Computer applications. He is a specialist in Stochastic Optimization Problems.