IJMS 2014 vol. 1(2):67 - 74

International Journal of Multidisciplinary Studies (IJMS)

Volume I, Issue I, 2014

Hierarchical Tag-set for Rule-based Processing of Tamil Language

Kengatharaiyer Sarveswaran and Sinnathamby Mahesan

Department of Computer Science

University of Jaffna, Sri Lanka.

ABSTRACT

Corpora are fundamental tools for Natural Language Processing. Part of Speech tagging provides more meaning to the corpora by annotating words. A tag-set used to annotate a corpus should be selected in such a way that it represents grammatical structure of the respective language. These tag-sets can be flat or hierarchical in structure. There are several efforts have been made in Tamil language to identify a tag-set. However, existing tag-sets have many shortcomings including inability of tagging all the words, inability to capture required syntactic information such as divisibility, too many numbers of tags in a set, flat in tag structure, and lack of extendibility. The scholar works Tolkāppiyam and Naṉṉūl clearly shows the grammatical classification of words. This paper proposes a new hierarchical tag-set with 10 labels for Tamil language in view of developing a morphological analyser by considering the existing limitations and using Tamil grammar. The morphological analyser can be used to extend the proposed tag-set easily with more grammatical information.

KEYWORDS: *POS tagging, Tag-set, Morphological analyser, Tamil grammar*

corresponding author Kengatharaiyer Sarveswaran, eMail: {sarves, mahesans}@jfn.ac.lk

1. INTRODUCTION

Corpus is a basic language resource for researchers in Natural Language Processing (NLP) for developing language technology applications. Words in corpus are normally annotated using a set of tags or Part of Speech (POS) labels to make them more useful for language processing; this process is called *Part of Speech tagging*. A Corpus may have text from single a language, called *mono-lingual* corpus, or multiple languages - *multi-lingual* corpus. Corpus like 'Penn Treebank' is called *parsed corpus*, which consists of parsed text with all the syntactic structure information.

Identifying appropriate tag-set for a corpus is a challenging task, which needs much prior studies about the language structure and thorough study about the purpose the corpus needs to be used for; identifying a tag-set for a general purpose corpus is even more challenging as it has all different kinds of text such as poems, colloquial write-ups, technical documentations, etc.

Like in other languages, several attempts have been made to identify a tag-set for Tamil language. Some tag-sets are developed for specific applications while others are claimed to be general purpose. Borrowing tags from one language may not be helpful to another language in some or all aspects; it may not capture the syntactic information of the intended language.

This paper surveys and critically analyses the existing tag-sets, and proposes a tag-set for Tamil language, specifically for developing morphological applications. This paper also shows that how the proposed tag-set helps tagging the words though small in size, and how further syntactic information can be obtained by using morphological analyser.

2. TAMIL LANGUAGE

Tamil, a member of Dravidian Language family, is a classical language that has the scholarly work called Tolkāppiyam (தொல்காப்பியம்) date back to 200BC and recorded as the oldest work in Tamil (George, 2000). Since then significant literatures have been written in Tamil. Tolkāppiyam is the known earliest work on Tamil grammar. Tamil grammar consists of five parts such as E̲luttu (letter), col (word), poruḷ (life-style/meaning), yāppu (form) and aṇi (method). The last two parts, yāppu and aṇi are the grammar for poetic writing (Renganathan, n.d.).

Another notable works on Tamil grammar is Nan̲n̲ūl (நன்னூல்) written in the 13th century. Nan̲n̲ūl was a derived work of Tolkāppiyam with refinements to include the need for the present context, and thus Nan̲n̲ūl is considered as the base of contemporary grammar (Shapiro & Schiffman, 1981).

Further, Tamil language is an agglutinative language in which affixes to a root word are used to mark several information including class, number, tense, gender and mood. It is relatively free from strict ordering of words in a sentence, called free word-order language; however, mostly written in Subject-Object-Verb order.

Pragmatics is very important for Tamil, because context heavily contributes to the meaning. To get the actual meaning of a sentence or phrase, the context needs to be considered to resolve any ambiguity. For instance, அவள் ஆண்டாள் (Avaḷ āṇṭāḷ) can be interpreted as "she is āṇṭāḷ" considering "āṇṭāḷ" as a proper name, and on the other hand it can also be interpreted as "she ruled (an empire)" taking "āṇṭāḷ" as complete-verb-of-past.

Moreover, depending on the grammar we practise, words may be classified in dif-ferent ways. For example, according to Tolkāppiyam a word should not begin with a letter ச, சை or சௌ amongst many others (Naccin̲ārkkin̲iyār, 1937). However, a number of words do have these letters as the first letter as can be seen even in classic literatures like Caṅkam literatures. Sticking to Tolkāppiyam grammar to tag a Tamil text, the words like சங்கம் (Caṅkam), சமைப்பான் (Camaippān̲), for example, will have to be classified as foreign words, whereas Nan̲n̲ūl allows us to classify them as a noun and a verb respectively as Tamil words (Caṅkara namaccivāyappulavar, 1957).

3. TAG-SET AND POS TAGGING

Part of Speech (POS) tagging plays a critical role in NLP application such as machine translation, question answering system, and spelling & grammar checker (Petrov, Das & McDonald, 2012) (Pandian & Geetha, 2008). In POS tagging, words are tagged in such a way that it shows significant amount of syntactic information about the word and its neighbours (Mohanty, 2005). Selection of appropriate labels or POS tags that provide syntactic information about language is the first task of POS tagging of a corpus. The number of labels used in POS tagging is determined by the syntactic complexity of a language and the purpose for which the corpus is built.

Since the syntactic structures drastically vary among languages, especially among different families of languages, sets of labels that are used in POS tagging also vary among languages. The analysis of the tag-sets underlying various corpus shows that the majority of tag-sets are very fine-grained and language specific (Petrov, Das & Mcdonald, 2012), because of this the POS tagging is also referred to as grammatical tagging (UCREL, 1987). A set of tags for a language not only depend on the grammar of the language, but also on the purpose for which the tags are going to be used. Identifying comprehensive set of tags for a language is a challenging task for Tamil language, which has undergone several revisions in the past. It is critical to identify an appropriate set of tags by considering the language grammar and the purpose (Mohanty, 2005).

In addition, tags should be chosen in such a way that they should not lead to ambiguity when a word occurs in different contexts [6].

Researchers have come up with different numbers of tags for different languages, ranging from 11 for Russian to 294 for Chinese (Petrov, Das & McDonald, 2012). Tagging a corpus by hand need enormous man power. Therefore, tagging is usually automated using different techniques. However, larger number of tags will negatively influence the accuracy of the POS tagging. On the other hand, a smaller number of tags will be less useful and may not provide adequate information (Mohanty, 2005).

A tag-set can have two structures, namely, hierarchical and flat (Baskaran et al., 2008). In flat structure, tag-sets list down the categories applicable for a language without any provision for modularity or feature reusability.

On the other hand, in hierarchical structure, tag-sets are structured relative to one another and providing flexibility for customisation according to the language and application. In addition, the hierarchical approach will help to easily extend the tag-set for future needs. For instance, in hierarchical approach, a word can be classified as verb or noun initially and then if the word is a verb then it can be further checked for class, tense, and mood as necessary (Baskaran et al., 2008).

In many cases, tag-sets are designed on the basis of morphological information, such as person, number, gender, tense, aspect, modality, case, and the like (Mohanty, 2005). However, applications like morphological analysers can be used to identify grammatical categories mentioned above instead of tagging every piece of information. For instance, in Tamil, the verb *படித்தான்* (Paṭittāṉ) can be divided using a morphological analyser as: *படி+த்த்+ஆன்*. Here *படி* is the root verb, *த்* indicates that it is a past tense word and *ஆன்* indicates that it is singular and the actor is a masculine (Nuhman, 1999). However, some primary information like whether a word can be divided or not and the primary grammatical category such as noun or verb should be fed to the morphological analyser. With the use of such knowledge further information can be obtained by a morphological analyser. Since the accuracy of a POS tagger is degraded when the number of tags increased, an application like morphological analyser can be used to overcome the problem. Therefore, in addition to the basic grammatical category decided by a POS tagger, extended syntactic information can be obtained by using morphological analyser.

Further to representing syntactic information of a language, tag-set should be able to denote punctuation marks such as period, comma, question mark, exclamation mark, *etc.* (Taylor, Marcus & Santorini, 2003). However, with tag-sets proposed by (Pandian & Geetha, 2008) (Dhanalakshmi, Kumar, Shivapratap, Soman & Rajendran, 2009) only a selected set of notations like dot and comma can be tagged. It may not help to identify, for instance, exclamation or questions or the parts in parentheses.

In some corpora, words are also tagged to mark information like whether all the letters are in upper case, or whether the word appears in a title or in body. For example, in Brown corpus, words occurring as constituents of titles are given their normal tag with the addition of the hyphenated tag –TL (Francis & Kucera, 1979).

4. TAG-SET FOR TAMIL

Several researches have been carried out on POS tagging and many tag-sets have been identified for Tamil (Sankaran et al., 2008) (Umaraj, 2012) (Baskaran et al., 2008). Many of these tag-sets are claimed as a tag-set for general purpose corpus that can be used for various kinds of NLP applications, and some are proposed for specific NLP applications. Also, there are efforts proposing language independent universal tag-sets. Tag-sets that have been identified from conference papers and journals obtainable via the Internet are taken for the discussion in this section.

4.1 Tag-sets for Tamil language

Dhanalakshmi, Kumar, Shivapratap, Soman & Rajendran (2009) proposed 32 POS tags for Tamil. Authors claim that this has been proposed by considering the following two problems found in the existing tag-sets: One, the existing tag-sets were large in size, and this led to more ambiguity in tagging. Therefore, the success rate would be less. The second problem arises due to the inclusion of tags for grammatical features like tense, gender, mood, etc. Authors claim that this makes the tagging process complex. By considering these problems, a set of 32 tags have been proposed by these authors. The proposed tag-set is flat in structure. This does not provide a way to tag brackets, symbols, foreign words, and punctuation marks like exclamation marks. Another key problem of this tag-set lies in its ability to distinguish between the noun and its inflected forms. For example, this tag-set takes *தமிழ்க்* in *தமிழ்க்* கொடி as noun. However, it is an inflected noun. This issue is found in most of the available tag-sets.

Madhu, Vijay & Ashish (n.d.) have proposed a tag-set with 12 tags for Tamil, and authors claim that these 12 are the most frequent categories of words. Notable feature of this tag-set is having a tag called "others", which can be used to tag word unknown in Tamil.

Selvam & Natarajan (2009) have identified more than 600 tags for Tamil language by considering finer details of Tamil grammar such as 11 cases of nouns, all different possible tenses, genders, and mood along with verbs. This is an extensive set of flat tags, and further analysis may even increase the number of tags. A rule-based approach has been used to do the tagging. Writing rules for extensive flat list of tags will be a tedious task, and some of these details could also be obtained by using morphological analysers.

Many existing tag-sets have the influence of tag-sets of other languages like English. For instance, the tag 'proper noun' is not used in typical Tamil grammar; instead it has two categories called 'given name' (iṭukuṟip peyar), and 'rationale name' (Kāraṇap peyar). For the rule-based processing of a language, native grammatical annotations are important; the borrowed tag would have less value.

Existing tag-sets for Tamil are designed in a flat structure capturing only coarse-level categories (Sankaran et al., 2008). Based on the analysis of the existing POS tag-sets and the results, it is clear that hierarchical tag-sets are flexible and would provide good accuracy while having the option to extend it further (Baskaran et al., 2008). Instead of having a large number of independent tags, a hierarchical tag-set contains a small number of categories at the top level, each of which has a number of sub-tags that can be arranged in a hierarchical or tree form, and it can be made flat, if so needed for ease of any processing.

4.2 Tag-sets for Indian languages and universal tags

India, IIIT Hyderabad (2006) has released a POS tag-set for Indian languages that contains 21 tags. This is a hierarchical tag-set, which consists of three tag-sets that contain many sub tags.

Baskaran S. *et al*. (2008) have proposed a hierarchical tag-set for Indian languages, specifically for Dravidian languages and Indo-Arian languages. In this research the tag-sets have been classified into different levels, namely categories, which are obligatory tags, types, and attributes. 31 tags have been identified on the second level as types and 18 attributes on the third level. The important aspect of this work is that only 11 tags are must, and then tagging can be done in finer levels as necessary. When go finer in granularity, the tags become more language specific.

Petrov, Das & McDonald (2012) have identified 12 language independent universal tags, including a tag for punctuation and a tag for all the unknown, foreign words. This work mainly focuses on mapping or merging tags in different languages to come up with parallel corpus. This is a flat scheme.

5. NEW TAG-SET

A new tag-set for Tamil language has been proposed in this research after analysing existing efforts. The new tag-set has 10 labels shown in Table 1. These tags have been identified in view of developing a rule-based morphological analyser, and using which the tagging can be further extended. For instance, grammatical features like tense, class, mood *etc*. can be identified by using the morphological analyser.

The new proposed tag-set has been derived considering the scholar works Naṉṉūl that categorises the words clearly in hierarchical manner for contemporary usage. Fundamentally, words have been classified into four types, namely, noun, verb, conjugation (Iṭaiccol) and attributive (uriccol) in Naṉṉūl. Conjugation words do not stand alone and attributive words are not very common in regular writings. Naṉṉūl classifies words into two categories known as divisible (Pakupatam) and indivisible (pakāppatam). Even though this is important for rule-based processing, especially for morphological analyser, it is not addressed in any of the existing tag-sets.

The proposed tag-set is also designed in such a way that it can be extended in hierarchical manner as shown in [Fig. 1]. Labels for nouns (‘N’) and verbs (‘V’) are defined in the tag-set. At the top-level all the nouns and their inflected forms will be marked as ‘N’ and all the verb forms will be marked as ‘V’. Further refinement will be performed in the next hierarchical level. For example, finite words will have additional tag ‘F’ in the second level, as this is required for morphological analyser, and for identifying adverbs. Moreover, the divisibility of a word is marked using ‘D’.

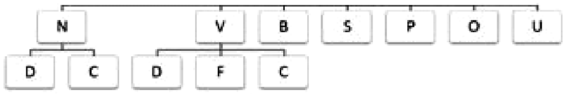

Fig. 1. The hierarchical structure of the proposed tag-set

Inflected nouns such as தமிழ்க் will get tag 'N' on the top level and 'D' on the second level. The ‘borrowed words’ are tagged using ‘B’. Words in other scripts also will be tagged using this label. ‘Compound words’ are marked separately using ‘C’ label, because this information is also important for morphological analyser. All other known word types such as conjugation, attributive are tagged as ‘other’ by label ‘O’. If a word is colloquial, poetic or unknown, then it is marked as ‘Unknown’ with label ‘U’.

English Corpus like Brown uses hyphenated tags combining two more tags for a word. Several tags can be merged together using hyphen without introducing a new tag. This idea is used in the tag-set being proposed in this paper. The hyphenated tags will also be useful to identify the super class when granular level is dealt with. The following examples shows the hierarchical tagging with the use of hyphenated tags:

- *கண்ணன் \N கதவைத் \N-D திறந்துகொண்டு \V-C போனான் \V-F . \P*
- *மாணவர்கள் \N-D தொழில்நுட்பத்தில் \N-C நாட்டம் \N கொண்டவர்களாக \O உள்ளனர் \V-F . \P*
- *இன்று \N டீசல் \B விலை \N 10 \S ரூபாய் \B . \P*

Table 1. Proposed tag-set

Tag	Label	Description
Nominal word	N	Nouns, including arbitrary nouns, rationale nouns. Example: கண்ணன், கண்ணனின்
Verbal word	V	Finite verb, Adjective, Adverb, Participial noun *etc.* Example: வந்தான், வந்து, வந்தவன்
Finite word	F	Finite words, example Finite verbs. Example: வந்தான்
Divisible word	D	Divisible nouns and verbs are marked with D tag. Example: வந்தான் – Divisible verb, கண்ணனின் – Divisible noun
Compound word	C	Words that can be further categorised into more meaningful words are annotated as Compound Word. Example: தெரியாதிருந்தான்
Borrowed word	B	Words that do not satisfy the Tamil grammar but exist in Tamil text are categorised as borrowed-word. Example: டீசல்
Symbol	S	Tamil notations such as யாசூபுஊஉமீஞு, notations like date, time and notations used in other languages are annotated with this label.
Punctuation	P	All the punctuation marks, including ? ! ” ‘ . – will be annotated using this Label
Other	O	All other Tamil words, including conjugation, attributive are tagged using label.
Unknown	U	All the unknown words are tagged with this. Later, all these unknown words can be easily grabbed and reviewed easily.

6. CONCLUSION

Analysis on existing tag-sets showed the need for a new, extendable tag-set for Tamil language, especially for the rule-based processing of Tamil language. According to the need, a hierarchical tag-set has been proposed in this paper in view of developing a morphological analyser. The tag-set has 10 labels and this can be extended easily. More importantly, the proposed tag-set is derived from the Tamil grammar itself without borrowing from other languages.

Future works

Currently, words are tagged by hand. An automatic tagger that facilitates for hierarchical tagging needs to be developed. Further, now we are working on a morphological analyser for Tamil language in view of automating the task of tagging.

REFERENCES

MOHANTY, R.K. (2005) Parts-of-Speech Tagging. In *2nd Asian Regional Training on Local Language Computing*. Cambodia, 23 June 2005. Cambodia: PANLC.

UCREL. (1987) *CLAWS part-of-speech tagger for English*. [Online] Available from: http://ucrel.lancs.ac.uk/claws. [Accessed: 30th August 2013]

PETROV, S., DAS, D. & McDONALD, R. (2012) A Universal Part-of-Speech Tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul, 21-27 May 2012. Istanbul: LREC. pp. 2089-2096.

PANDIAN, S.L. & GEETHA, T.V. (2008) Morpheme based Language Model for Tamil Part-of-Speech Tagging. In *polibits*. 38. p. 19-25.

FRANCIS, W.N. & KUCERA, H. (1979) *Brown Corpus Manual*. [Online] Available from: http://clu.uni.no/icame-/manuals/BROWN/INDEX.HTM. [Accessed: 20th August 2013].

SANKARAN, B. et al. (2008) A Common Parts-of-Speech Tagset Framework for Indian Languages. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech, 28-30 May 2008. Marrakech: LERC. pp. 1331-1337.

UMARAJ, K. (2012) Issues while developing annotated corpora for Modern Tamil. In *Proceeding of 11th International Tamil Internet Conference*. Chidambaram, 28-30 December 2012. Chidambaram: TI2012. pp. 112-115.

GEORGE L,H. (2000) *Statement on the Status of Tamil as a Classical Language*. [Online] Available from: http://southasia.berkeley.edu/tamil-classes. [Accessed: 10th June 2013].

RENGANATHAN, V. (n.d.) *Tamil Language, History and Literature*. [Online] Available from: http://www.southasia.sas.upenn.edu/tamil/lit.html. [Accessed: 10th September 2013].

SHAPIRO, M.C. & SCHIFFMAN, H.F. (1981) *Language and society in South Asia*. Missouri: South Asia Books.

NUHMAN, M.A. (1999) *Basic Tamil Grammar*. Colombo: Readers' Association.

TAYLOR, A., MARCUS, M., SANTORINI, B. (2003) The Penn Tree Bank: An Overview. Abeillé, A. (ed.). *Treebanks*. Text, Speech and Language Technology, Vol. 20. Netherlands: Springer.

BASKARAN S., et al. (2008) Designing a Common POS-Tagset Framework for Indian Languages, In *The 6th Workshop on Asian Language Resources*. Hyderabad, 11-12 January 2008. Hyderabad: ALR. pp. 1331-1337.

DHANALAKSHMI, V., KUMAR, A., SHIVAPRATAP, G., SOMAN, KP. & RAJENDRAN, S. (2009) Tamil POS Tagging using Linear Programming, *International Journal of Recent Trends in Engineering*. Vol. 1, No. 2. pp 166-169.

MADHU, R., VIJAY, C. & ASHISH, P. (n.d.) *An Attempt at Multilingual POS Tagging for Tamil*. [Online]. Available from: http://pages.cs.wisc.edu/~madhurm/CS769_final_report.pdf. [Accessed: 20th June 2013].

INDIA. IIIT HYDERABAD. (2006). *POS Tag Set for Indian Languages*. India: IIIT Hyderabad.

SELVAM, M., NATARAJAN, A.M. (2009) Improvement of Rule-based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques. *International Journal of Computers*. Issue 4, Volume 3. pp. 357-367.

Nacciṉārkkiṉiyār. (1937) *tolkāppiyam (eḻuttatikāram) urai*. Chunnakam: The Thirumakal Press.

Caṅkaranamaccivāyappulavar. (1957) *naṉṉūl viruttiyurai*. ceṉṉai: vittiyānupālaṉa accakam.