CHAPTER 6

DEVELOPMENT OF RULE BASED TOOLS FOR TAMIL LANGUAGE

6.1 INTRODUCTION

Generation of Treebank needs laborious man power, time, effort, expertise, bootstrapping, accuracy, consistency and automated tools as mentioned in section 5.5. It needs robust morphological analysis, POS tagging, phrasing and marking dependency relations. Some of the tools reported in the literature for morphological analysis, POS tagging and phrasing are limited to certain domains, corpora and applications. They are not suitable for the large scale generation of Treebank. When morphological rules are rich in a language, electronic resources, like the corpora, are not needed in high volume for training process. Even if there is a resource deficiency, the rule based morphological analysis and POS tagging can be done with a reasonable accuracy (Sandipan Dandapat et al 2007, Singh et al 2006, Aniket Dalal et al 2007). Generation of new tags, stemming process, creating lookup tables, dynamic memory learning and disambiguation are some of the problems to be handled in an effective manner. This chapter focuses on the development of rule based tools for morphological analysis, POS tagging and phrasing.
6.2 ISSUES IN RULE BASED MORPHOLOGICAL ANALYSIS AND POS TAGGING

There are some issues in the generation of rule based morphological analysis and POS tagging.

- Sometimes, a pair of words with the same or different words, is used as echo words. In that case, a single word does not have any meaning. They are to be referred in the lookup table. Some of the examples are worth mentioning:

  लाला सुंदरी (sound of laughter)
  श्री श्री (sound of flowing water)
  (here and there)
  (wander and loiter)

- Nouns may take forms combined with verbs and participles. In morphological analysis and POS tagging, words should be checked for verbal and participle noun patterns before verbal patterns with stemming and checking for suffix patterns. Some of the examples are,

  (PARTICIPLE NOUN) (man who is doing)
  (VERBAL NOUN UNTENSED) (doing)
  (PARTICIPLE NOUN or VERB)
  (which has been finished or finished)

  [To identify whether it is participle noun or verb, the next word has to be checked for verbal pattern]

- To differentiate a verb from a verbal participle noun, location is important. If the word is located at the end of a sentence, then it
is a verb since Tamil is a head final language. Some of the examples are,

\[
\text{VERB} \quad \text{(a bird flew)}.
\]

\[
\text{PARTICIPLE NOUN} \quad \text{(the one which flew fell down)}
\]

- Just like verbal and participle nouns, adjective nouns also have the same suffix patterns. After the removal of noun suffix patterns, partial stems are to be checked for adjective patterns in the lookup table like verbal patterns. Some of the examples are,

\[
\text{PARTICIPLE NOUN} \quad \text{(a man who is doing)}
\]

- Verbal stem + pronoun suffix pattern

\[
\text{PARTICIPLE NOUN} \quad \text{(one which is finished)}
\]

- Verbal stem + pronoun suffix pattern

\[
\text{ADJECTIVE NOUN} \quad \text{(a man who is good)}
\]

- Adjective stem + pronoun suffix pattern

\[
\text{ADJECTIVE NOUN} \quad \text{(a man who is intelligent)}
\]

- Adjective stem + pronoun suffix pattern

- Verbal participle and adjective nouns may take the combined form with prepositional cases. First verbal participle or adjective needs to be separated from pronoun with prepositional case. Noun category is selected from POS of the remaining part which contains pronoun with prepositional suffix. Some of the examples are,

\[
\text{PARTICIPLE NOUN} \quad \text{(to the man who is doing)}
\]

\[
\text{ADJECTIVE NOUN} \quad \text{(the man who is intelligent)}
\]
In case of continuous and perfect tense in Tamil, auxiliary and verbal patterns (and) are to be identified and processed separately. For example,

(have done)
- +

(am doing)
- +

(beautiful) - + .ஒல்
(developing) - + .ங்
(developed) - + + .ங்
(to be developed) - +
(not developed) - + .அ்

- Affixes in the form of prefixes need to be handled.

- Prepositions may come with cases also. They are to be separated and processed. For example,

(over) – (PRP) + ꞏ(Case)

- Prepositions may also come along with WH questions. They are to be separated and processed. Some of the examples are,

(to whom) - +
(from where)- +

- Benefactive case (preposition - , ) and adverbial suffix .அ் lead to ambiguity. To resolve it, priority should be given to prepositional case. Some of the examples are,
6.3 ISSUES IN PHRASING

While framing rules for phrasing, many issues are encountered. A few special rules are considered here to address the issues for some peculiar cases:

- During first level phrasing, if first constituent element is adverb (AV) or noun (NN), and second element is a verbal noun or noun with preposition case (PR), they are combined to form prepositional phrase (PP).

  \[(PP (ADV \quad) (NVTB \quad))\]
  \[(PP (NNSM \quad) (NNSMB \quad))\]

- If an adverb (AV) or preposition (PR) or verb phrase (VP) is followed by verbal noun (NN), they are combined to form noun phrase (NP).

  \[(NP (ADV \quad) (NVUT \quad))\]
  \[(NP (NNSNB \quad) (NVUT \quad))\]

- When similar categories appear in sequence, they are combined with the same phrasal tag till they end. This happens for sequence of nouns or verbs

  \[(NP \quad (NP \quad (NNSN \quad) (NNSN \quad)) \quad (NNSN \quad))\]
  \[(VP \quad (VP \quad (VP \quad (CVPP \quad) (CVPP \quad)) \quad (OPT \quad))) \quad (VTSNF \quad))\]
During second level phrasing, if a prepositional phrase (PP) is followed by an adjective (AJ) of type ADJ, the adjective (ADJ) should be combined with the next word.

\[(PP \ (DET \ \ \ ) \ (NNSNL \ \ \ )) \ (NP \ (ADJ \ \ \ ) \ (NNPN \ \ \ )) \ (VP \ (VTPNA \ \ ) )\]

When a prepositional phrase (PP) is followed by an adjective (AJ) of type ADJAP, ADJFP, ADJNP and ADJPP, they should be combined with the preceding prepositional phrase (PP) to form adjective phrase (ADJP).

\[(NP \ (ADJP \ (PP \ (DET \ \ \ ) \ (NNSNG \ \ \ )) \ (ADJAP \ \ \ )) \ (NNPN \ \ \ )) \ (PP \ (PRP \ \ ) ) \ (VP \ (VTPNA \ \ ) )\]

6.4 IMPLEMENTATION OF RULE BASED MORPHOLOGICAL ANALYSER AND POS TAGGER FOR TAMIL LANGUAGE

The formation of words in Tamil is based on a well-defined set of rules. Adverbs ( – fast) and adjectives ( – beautiful) have a closed set of suffix (, ) patterns. Similarly direct prepositions ( – in), conjunctions ( – or), pronouns ( – he) and interjections ( – alas) also have a closed set. For all these closed set of words, morphological analysis is not needed and POS tags are applied directly with the help of the lookup table. But, for adverbs and adjectives, morphological analysis is done by separating suffixes.

Adjective noun ( – the man who is good), verbal noun ( – one who has done) and participle noun ( – one which is done) are some of the patterns which have pronouns ( , ) in
them. They are identified and separated by substring match. POS tagging is done after separating pronouns, adjectives or verbs or participles, and case endings (— + + — to the one who has done).

Perfect and continuous tense words (, , ) are to be separated from words by substring match. POS tagging is done with separated word patterns (— + — have gone, — + — had gone, — + — was going). Checking pronoun with preposition is done by substring match with case endings, and pronoun and case ending patterns (— + — to him). POS tagging is done based on case endings which are considered as inflectional prepositions. Similarly, for other noun based inflections, case endings are identified and morphological analysis and POS tagging are done with case endings (— + — by the cat).

Verbal inflections are checked with 80 possible patterns (Rajendran et al 2003) and some more newly suggested patterns. Morphological analysis on verbs is done with verbal suffix patterns described in Table 5.3. POS tagging on verbs is done with suffix patterns which cover number, person, gender, tense, voice, honorific, question, etc. Numbers are checked with lookup table and partial pattern matching (— + — thousand and hundred).

There may be ambiguities with regard to nouns and verbs. So disambiguation module is needed for the identification of nouns and verbs when multiple tags are assigned to a word. In disambiguation module, nouns
are distinguished from verbs by the POS of preceding word like determiner or adjective. Verbs are distinguished by the POS of preceding word like adverb or verbal participles. If this is not resolved, the most frequently applied tag is used by matching it with the dictionary which contains word, tag and frequency. Dynamic memory learning technique is used to add new nouns, verbs, adjectives and adverbs in various lookup tables when words do not match with entries of their respective lookup tables. The proposed rule based morphological analyzer and POS tagger are shown in Figure 6.1.
Figure 6.1 Rule Based Morphological Analyzer and POS Tagger
6.5 IMPLEMENTATION OF RULE BASED PHRASING TOOL FOR TAMIL LANGUAGE

In this process, the input sentence and its sequence of tags generated by the rule based morphological analyzer and POS tagger are given as inputs to the phrasing tool. For the sequence of tags, their respective categories are fetched from the lookup table. Then, zero level phrasing is applied with categorical information to form bracketed tag-word pairs. Starting from first level phrasing to N level phrasing, phrasal tags are applied according to the rules and their priorities. The final annotated sentence with root tag is obtained from this phrasing tool as output. The collection of annotated sentences forms a Treebank which will be the resource for training the phrase structure parsing model. The phrasing tool is shown in Figure 6.2.

![Figure 6.2 Rule based Phrasing Tool](image-url)
6.6 EXPERIMENTS

The rule based morphological analyzer and POS tagger are built successfully based on the proposed morphological and POS tagging rules. From CIIL corpus, 455,504 sentences from 686 documents (which contain 2,652,370 words) are taken as input and their POS sequences are obtained. Among the 455,504 sentences, 1,000 sentences are selected randomly and the gold standard is prepared. Also another 1000 sentences from the Bible corpus are selected and gold standard is prepared. The rule based morphological analyzer and POS tagger are tested for both the test cases (1000 sentences each from CIIL and the Bible corpora). The rule based phrasing tool is also built successfully based on categorical information and proposed phrasing rules. For the same 1,000 sentences from CIIL corpus, the gold standard for phrasing is obtained as Treebank. The gold standard prepared for rule based morphological analyzer and POS tagger in CIIL corpus is given as input to the rule based phrasing tool and thus phrased sentences are obtained as Treebank.

6.7 RESULTS AND DISCUSSION

Next, POS sequences of those 1,000 sentences are tested with gold standard and an accuracy of 85.56% is achieved. The generated phrase structure Treebank is tested with its gold standard and a phrasing accuracy of 75.11 % is achieved. The details are shown in Table 6.1.
Table 6.1 Results of Rule Based Tools

<table>
<thead>
<tr>
<th>Details</th>
<th>Rule based Morphological Analysis and POS tagging</th>
<th>Rule based Phrasing tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus Used</td>
<td>CIIL Corpus</td>
<td>CIIL corpus</td>
</tr>
<tr>
<td></td>
<td>Bible Corpus</td>
<td></td>
</tr>
<tr>
<td>No. of sentences for annotation</td>
<td>455,504 (CIIL Corpus)</td>
<td>1,000 (same CIIL sentences in gold standard)</td>
</tr>
<tr>
<td></td>
<td>1000 (Bible Corpus)</td>
<td></td>
</tr>
<tr>
<td>No. of sentences in Gold standard</td>
<td>1,000 (same CIIL test set)</td>
<td>1,000</td>
</tr>
<tr>
<td></td>
<td>1000 (Bible)</td>
<td></td>
</tr>
<tr>
<td>No. of sentences in test set</td>
<td>1,000 (CIIL random sentences)</td>
<td>1,000 (parsed sentences)</td>
</tr>
<tr>
<td></td>
<td>1000 (Bible corpus)</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>85.56% and 83 %</td>
<td>75.11%</td>
</tr>
</tbody>
</table>

The performance of hand annotation with bootstrapping using phrase structure parsing model reported in section 5.5 is better in POS tagging than the rule based morphological analyzer and POS tagger. But it requires multiple cycles of bootstrapping and manual corrections to improve the parsing model. But in rule based approach, manual efforts are not needed and Treebank generation is much simpler compared with bootstrapping model.

The reliable parsing models can be trained with a large volume of annotated sentences obtained from rule based approaches with manual corrections. Still, a better performance can be achieved if the exact root word and categorical information are provided. The rule based morphological analyzer and POS tagger can be improved through statistical methods like POS projection and morphological induction from English through sentence aligned corpora and by leveraging lemmatization in English (David Yarowsky and Grace Ngai 2001, David Yarowsky et al 2001).
6.8 SUMMARY

The rule based morphological analyzer, POS tagger and phrasing tool for Tamil have been successfully developed based on detailed linguistic and morpho-syntactic studies. In addition, the various issues behind morphological analysis, POS tagging and phrasing in Tamil have been addressed. These tools have been tested with their respective gold standards and a better performance is achieved.