Recognition Of Kalippa Class Of Tamil Poetry

Rajeswari Sridhar, Rajkiran R, Narendra Kumar N J, Giridhar M

rajisridhar@gmail.com, rajkiran2507@gmail.com, billanaren@gmail.com, mgiridhar11@gmail.com

Department of Computer Science and Engineering,

Anna University, Chennai - 600025, INDIA

Abstract: The aim of this paper is to recognize the Kalippa class of Tamil poetry which comes under the 'Pa' super class of poetry. Context Free Grammar is useful to identify the structure of a poem, which is initially converted into an intermediate representation, converting sounds of a similar group into one common symbol. This is followed by 'asai' computation, 'seer' computation and 'thalai' computation. We achieved an overall accuracy of identification of Kalippa class of 80%.

Index Terms: Kalippa, CFG, Seer, Thalai, Adi

1. INTRODUCTION

The structure of Tamil grammar for both prose and poetry is derived from the complex set of rules postulated in Tolkappiyam, an ancient text [1]. Tholkappiyam is split into three categories namely Ezhuthadhigaaram (rules for word formation in Tamil), Solladhigaaram (syntax of the Tamil language) and Poruladhigaaram (meaning of Tamil language). Two more sections are 'Yappu' (metrics) and 'Ani' (Figures of Speech) [1].

We aim to identify the subclass of 'Pa' class of poetry called 'Kalippa' using the rules mentioned under the 'Yappu' section of Tolkappiyam.

The paper is organized as follows: Section 2 deals with general classification of poetry, and specific to Kalippa, Section 3 delves into the previous work done in the same area in English and Tamil, Section 4 defines the CFG used for classification and architecture of the system, Section 5 deals with results and analysis using the designed system, and finally concluding with Section 6.

2. CLASSIFICATION OF TAMIL POETRY

'Pa' is the representation of 'osai' (harmonics). Based on the different harmonics, 'Pa' is classified into 4 as: Venpa, Aasiriyappa, Vanchippa and Kalippa. Most poems conform to one particular 'Pa' type, though poems which have a mixture of types are not unheard of in Paipadal and Kalithogai [1].

Figure 1 shows the classification of the ‘Pa’ class of poetry. In this paper we consider the identification of the Kalippa class of poetry. Kalippa unlike the other forms of 'Pa' is broken down into 6 parts. They are:

1. Tharavu 2. Thaazhisai 3. Araagam 4. Ambhodhrangam 5. Thanichol 6. Suridhagam

The resultant complexity of Kalippa makes it difficult for a machine to classify the type of Kalippa since the demarcation between the parts is usually through meaning and context.

Each of the parts has its own structure and rules:

Tharavu: It is the first part of a Kalippa, and consists of a minimum of 4 lines and a maximum of 12 lines per part. There may 0, 1 or 2 occurrences of Tharavu in a Kalippa verse. Tharavu predominantly contains Pulimangai and Karuvilangai cheers.

Thazhisai: The second part of a Kalippa, Thazhisai along with Tharavu forms the opening part of a poem introducing the topic. As a rule, Thazhisai must contain lesser number of lines than the corresponding Tharavu. According to Tamil

literature, the minimum is set at 2 and maximum is set at 11 lines. There can be 3, 6 or 12 occurrences of Thazhisai in a Kalippa verse.

Araagam: This is the third part and it is optional. Karuvilam seers are majority in Araagam. 4-word lines are the most commonly found, whereas other line structures are also possible. Upper limit is set at 8 lines.

Ambhodharangam: Literally means "waves", and it is named because of the diminishing size of the lines which starts with 4 worded lines and gradually diminishes to 3-worded and 2-worded lines.

Thanichol: A single word. This marks the beginning of the poem's conclusion.

Suridhagam: It is the final part of a Kalippa verse and follows the same structure as a Tharavu.

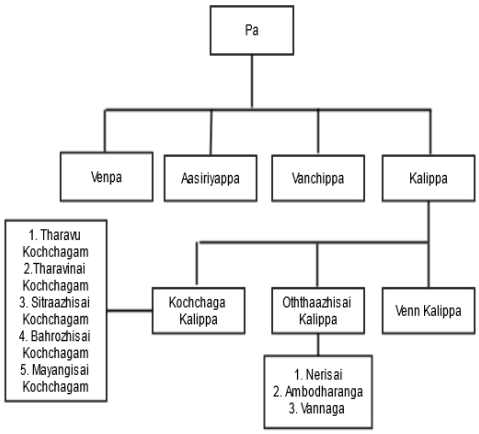


Figure 1 Classification of the 'Pa' class of poetry

3. SURVEY OF EXISTING WORK

The study and attempt of poetry classification has been performed both in English and in Tamil. The Venpa class of Tamil Poetry has been identified [2] [3] and classified by researchers using a CFG [4]. Research work is being done in classifying forms of 'Pa' after identifying them. Similarly in English, differentiating between prose and poetry based on shape, metric and rhyme using Bayes' Rule and Multi Layer Perceptron has been done by researchers [5]. Other class of 'Pa' namely Venpaa along with its sub-classes, Aasiriyappa and Vanjippa has also been done by constructing and optimizing CFG rules [6].

In this paper, we aim to contribute to the existing identification algorithms by incorporating Kalippa, which is the rarest among the types of 'Pa' and is the most difficult in identifying. The difficulty arises due to the various sections of Kalippa and the complexity in identifying these sections.

4. ALGORITHM

The algorithm for the identification of Kalippa is split into 3 stages.

1. *Input parsing and intermediate representation*: The input in Tamil is parsed and converted into an intermediate representation based on sound duration so as to facilitate further stages.

2. *Seer and Adi identification*: Using the existing rules for Venpaa, the seer and adi for each line are computed from the intermediate representation.

3. *Thalai computation*: Using the output of the previous block, we identify Thalais in the input poem and identify it as Kalippa, if it matches the required criteria (Kalithalai).

Unicode representation of Tamil characters: In this work, for implementation, each Tamil character is interpreted as one or two consecutive Unicode characters.

The following sub-sections describe the 3 steps described above.

4.1 Vowel-Consonant Tokenization

Tamil is a phonetic language in which alphabets are formed by combining vowels and consonants. There are 12 vowels and 18 consonants, thus resulting in 216 characters. Therefore, the first step in identifying the poem class is to segment the input using Tamil grammar rules, involving classification of alphabets into long or short alphabets based on the vowel and consonant rule.

The tokenizer performs the function of separating into short and long vowels as explained below:

Vowels and consonant-vowel compounds in Tamil alphabet have been classified into ones with short sounds (kuril) and the ones with long sounds (nedil). A sequence of one or more of these units optionally followed by a consonant can form a ner asai (the Tamil word asai roughly corresponds to syllable) or a Nirai asai depending on the duration of pronunciation. Ner and Nirai are the basic units of meter in Tamil prosody.

The input is tokenized as a sequence of vowels and consonants (kuril/nedil/ottru) by using the rules of the Tamil grammar based on short or long vowels. After identification this is written in an intermediate file and is used for identifying the 'asai' which in turn is used for the next phase namely, seer analysis.

4.2 Asai Determination

In Tamil language the 'Asai' is defined according to the following rules:

Ner asai:
1. Single Kuril
2. Single Kuril followed by Ottru
3. Single Nedil
4. Single Nedil followed by Ottru

Nirai asai:
1. Double Kuril
2. Kuril followed by Nedil
3. (1) and (2) followed by Ottru

In addition, the Ner asai and Nirai asai can be combined together in groups of 2, 3 or 4 and each combination has a name of its own. The occurrence of two asais either alone or in a combined fashion is called the seer, which is categorized in accordance with the rules mentioned below that are referred from a lookup file using hash table data structure thereby, reducing the input access time

Hence by referring to these rules and identifying from the intermediate file, each word in the poem is classified into 'asai' and further organized as 'seer'. Using this seer, each word is also assigned the corresponding name and is stored in the intermediate file.

4.3 Thalai Computation

The occurrence of connected feet (seer) in poetry is called 'thalai'. Every seer has a fixed word ending namely, 'maa', 'vila', 'kaai', 'kani' with which we can compute the feet of the poetry which will then be used for the assortment, also taking prefix

of the seer into account. These rules are later mapped and used for Thalai identification. We use these Thalais that are identified for the recognition of Kalippa.

4.4 Kalippa Identification

Based on the rules of Kalippa as discussed in the previous section we constructed a Context Free Grammar.

The CFG designed in the processing of the input verse to identify Kalippa is as follows:
G={V, T, P, S}
V={CHEER, EERASAI, MOOVASAI, EETRU CHEER, NAAL, MALAR, KAASU, PIRAPPPU, THEMAA, PULIMAA, KARUVILAM, KOOVILAM, THEMAANGAAI, PULIMAANGAAI, KARUVILANGAAI, KOOVILANGAAI, NER, NIRAI}
T={KURIL, NEDIL, OTRU}
P={
<CHEER> → <EERASAI> | <MOOVASAI>
<EETRU CHEER> → <NAAL> | <MALAR> | <KAASU> | <PIRAPPU>
<EERASAI> → <THEMAA> | <PULIMAA> | <KARUVILAM> | <KOOVILAM>
<MOOVASAI> → <THEMAANGAAI> | <PULIMAANGAAI> | <KOOVILANGAAI> | <KARUVILANGAAI>
<THEMAA> → <NER> <NER>
<PULIMAA> → <NIRAI> <NER>
<KARUVILAM> → <NIRAI> <NIRAI>
<KOOVILAM> → <NER> <NIRAI>
<THEMAANGAAI> → <THEMAA> <NER>
<PULIMAANGAAI> → <PULIMAA> <NER>
<KARUVILANGAAI> → <KARUVILAM> <NER>
<KOOVILANGAAI> → <KOOVILAM> <NER>
<NAAL> → <NER>
<MALAR> → <NIRAI>
<KAASU> → <NER> <NER>
<PIRAPPU> → <NIRAI> <NER>
<NER> → <KURIL> | <NEDIL> | <NER> <OTRU>
<NIRAI> → <KURIL> <KURIL> | <KURIL> <NEDIL> | <NIRAI> <OTRU>
<KURIL> → {VOWELS OR COMPOUNDS WITH A SHORT SOUND}
<NEDIL> → {VOWELS OR COMPOUNDS WITH A LONG SOUND}
<OTRU> → {CONSONANTS, WHICH HAVE AN EXTREMELY SHORT SOUND}
}

5. RESULT AND ANALYSIS

Input:
வாணெடுங்கண் பனிகூர வண்ணம்வே றாய்த்திரிந்து
தோணெடுந்தன் தகைதுறந்து துன்பங்கூர் பசப்பினவாய்ப்
பூணொடுங்கு முலைகண்டும் பொருட்பிரிதல் வலிப்பவோ?
சூருடைய நெடுங்கடங்கள் சொலற்கரிய என்பவால்
பீருடைய நலந்தொலையப் பிரிவாரோ பெரியவரே?
சேணுடைய கடுங்கடங்கள் செலற்கரிய என்பவால்
நாணுடைய நலந்தொலைய நடப்பாரோ நலமிலரே?

சிலம்படைந்த வெங்கானம் செலற்கரிய என்பவால்
புலம்படைந்த நலந்தொலையப் போவாரோ பொருளிலரே?
எனவாங்கு
அருளெனும் இலராய்ப் பொருள்வயிற் பிரிவோர்
பன்னெடுங் காலம் வாழியர்
பொன்னெடுந் தேரொடு தானையிற் பொலிந்தே!

After Asai recognition:
nNn Nnn nnn nNn
nNn NNn nnn NNn
nNn Nnn NNn NN
nNn NNn NNn nN
nNn NNn Nnn NNn
nNn NNn NNn nN
nNn NNn Nnn NNn
NNn nnn NNn nN
NNn NNn nnn NNn
Nnn
NN Nn NN Nn
nN nn nN
nN nN nnn Nn
n-Ner;N-Nirai

After Seer identification:
Koovilangai Pulimangai Themangai Koovilangai
Koovilangai Karuvilangai Themangai Karuvilangai
Koovilangai Pulimangai Karuvilangai Karuvilam
Koovilangai Karuvilangai Karuvilangai Koovilam
Koovilangai Karuvilangai Pulimangai Karuvilangai
Koovilangai Karuvilangai Karuvilangai Koovilam
Koovilangai Karuvilangai Pulimangai Karuvilangai
Karuvilangai Themangai Karuvilangai Koovilam
Karuvilangai Karuvilangai Themangai Karuvilangai
Pulimangai
Karuvilam Pulima Karuvilam Pulima
Koovilam Thema Koovilam
Koovilam Koovilam Themangai Pulima
Kaiseers : 34
Kaniseers : 0
Thalai identification:
Venthalai: 19
Aasiriyathalai: 5
Kalithalai: 23
Vanchithalai: 0

Result: Qualifies as Kalippa (due to the absence of Kaniseers and majority of Kalithalais)

Analysis of results: We were able to obtain 80% accurate identification for Kalippas using our algorithm. The errors in identification are caused by split in words, which are done to facilitate the understanding of meaning of the poem. Such

splitting of words, interferes with the seer identification and hence thalai computation. The presence of special characters such as commas, question marks and exclamation characters do not affect the input. Other reasons for the incorrect identification are due to the vagueness in the interpretation of Thazhisai and Araagam section of the Kalippa. In our work we have considered Thazhisai to be in the range of lines between 2 and 11, whereas occurrences of 3, 6, 12 are possible. This incorrectness percolated to Araagam also. For araagam section we have set the upper limit as 8 lines whereas it could be less or more. This could be corrected by modifying the CFG and also use a regular expression to handle individual sections which could be used in the construction of CFG.

6. CONCLUSION AND FUTURE WORK

In this work we identified the Kalippa class of poetry by constructing a CFG that describes the various sections of Kalippa. We achieved an identification accuracy of 80%. There exists only one major source of Kalippa namely the Kalithogai unlike the other types of 'Pa', which is a testament to the complex nature of Kalippa and its structure. The CFG could be further augmented to recognize certain unique aspects of the sub-types of Kalippa and classify them. In addition, CFG and hence the identification algorithm could be also modified to minimize the errors caused due to split words, by course of prefix matching.

REFERENCES

- Tolkappiyam in Tamil Unicode format http://www.projectmadurai.org/pm_etexts/utf8/pmuni0100.html
- Balasundaram L, Ishwar S, Sanjeeth Kumar Ravindranath, ”Context Free Grammar for Natural Language Constructs-An implementation of Venpa class of Tamil Poetry ”, Proceedings of Tamil Inayam, pp.128-136, 2003
- K.V. Madhavan, S. Nagarajan and Rajeswari Sridhar, “Rule based classification of Tamil poems”, International Journal of Information and Education Technology, Vol. 2, No. 2, pp. 156 – 158, 2012.
- http://en.wikipedia.org/wiki/Context-free_grammar
- Hamid R. Tizhoosh, Farhang Sahba, Rozita Dara, ” Poetic Features for Poem Recognition: A Comparative Study”, Journal of Pattern Recognition and Research (2008) 24-39
- S. Subha Rashmi, V. Subasree, Rajeswari Sridhar, "Classification of Tamil Poetry Based On Constructing Context Free Grammar Using Tamil Grammar Rules", accepted for publication at ICCSEA, 2013.