

CLASSIFICATION OF TAMIL POETRY USING CONTEXT FREE GRAMMAR USING TAMIL GRAMMAR RULES

Subasree Venkatsubhramaniyen¹, Subha Rashmi² and Rajeswari Sridhar³

¹Department of Computer Engineering,
College of Engineering Guindy, Anna University

ABSTRACT

Context Free Grammar is a prime tool for specifying rules to verify the syntax of any language. This paper aims to classify Tamil poems into its subclasses by framing their CFGs. Initially, tokenization is done where vowels (short syllables) and consonants (long syllables) are identified, followed by 'asai' analysis, 'seer' analysis. We then compute 'Thalai', which is essential to distinguish among the 'Paas'. We verify the found 'Paa' with the determined 'osai'.

KEYWORDS

Venpaa, Aasiriyapaa, Kalipaa, Vanchipaa, Sub-classification of 'Paas', & CFG

1. INTRODUCTION

Tamil grammar and poetics are ancient and unique disciplines interwoven into a complex system, the beginnings of which are a legend. Tolkappiyam [1], which is a very old literature dated 8000 BCE is a great treatise that encompasses the entire basic structure for Tamil grammar consists of 3 parts, viz Eluttu (phonology), Chol (morphology and syntax) and Porul (poetics). The above classification was extended further with an addition of 2 more parts, namely yappu(metrics) and ani(figures of speech).

In this paper, we focus on identifying categories of 'Paa' and its 'osai' which falls under yappu. The 'Paa' is used to classify and categorise poetry. The paper is organised as follows: Section 2 discusses a broad classification of poem, Section 3 discusses on some existing work in the areas of poetry classification both English and Tamil, Section 4 briefs the architecture of our system, Section 5 on Results and analysis and finally section 6 concludes the paper.

2. CLASSIFICATION OF TAMIL POEMS

'Paa' is a structural representation of 'osai' (sound). 'Paas' are featured by the variations in their musical pronunciation, leading to the 4 broad categories, 'Venpaa', 'Aasiriyapaa', 'Kalipaa' and 'Vanchipaa'. Most of the poems use more than one meter in their lines. To name a few works, Kuruktokai 122 is entirely in the aasiriyam meter, Purananuru 11 is mostly in the vanchi meter but ends in the aasiriyam, Paripatal 5 demonstrates a mixture of aasiriyam and vanchi meters in

lines 17-22, Paripatal 18 has a mixture of venpaa and aasiriyam in lines 49-56 , Kalittokai 11 has a mixture of kali and aasiriyam in lines 1-4, Paripatal 5 mixes venthalai and kalithalai in line 52. Similarly, Kalittokai 38 mixes venthalai and kalithalai in line 15 . Therefore, it is evident that an entire composition(paa or paattu) can be in just one meter or may contain sections which are in different meters.

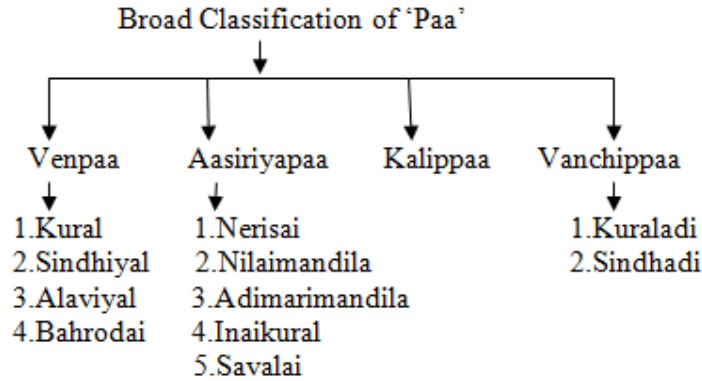


Figure 1. Paa classification

3. LITERATURE SURVEY

Researchers have classified and identified one of the 'Paas', namely Venpaa. In their work, Venpaa identification [2] was modeled based on CFG [6]. In the work done by the authors [2], CFG is represented as rules and based on the rules Venpaa identification is being carried out. The authors dealt only with identifying a given poem as Venpaa, but not its classification. English poem recognition [4] makes an attempt to classify a given English text as either poem or prose based on poem feature recognition techniques. Bayes rule and Multilayer Perceptron (MLP) were used for the classification task. Their approach of poem recognition was primarily based on the three poetic features namely shape (structural features), meter and rhyme. Another interesting research paper on Metaphor identification and Analysis [5] revolves around identifying metaphorically used words, thereby developing taxonomy of the propositional structure of metaphors.

In this work, we were motivated by the CFG representation of Venpaa and have extended the rules for the sub-classification of Venpaa. In addition we have also extended and designed CFG for other Paa variations namely Aasiriyapaa and Vanchippaa.

4. ALGORITHM

The algorithm for our Paa identification system is given in Fig. 1. In this work, for implementation, each Tamil character is interpreted as one or two consecutive Unicode characters. The following sub-section discusses each module of the Venpaa classification system.

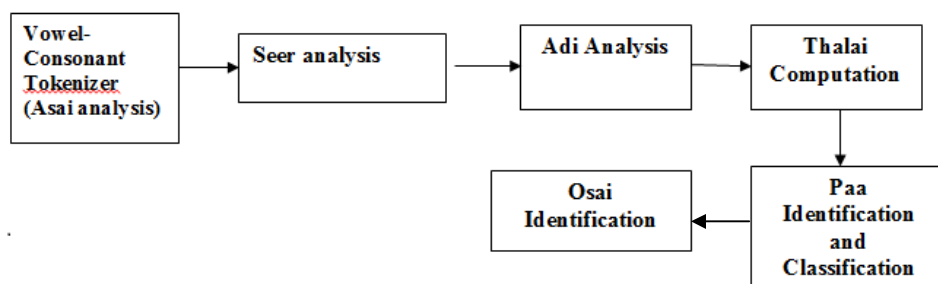


Figure 2. Block diagram of Paa analyzing system

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/ndeil/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:

Table 1. Asai rules.

Ner asai	Nirai asai
1. Single kuril 2. Single kuril followed by ottru 3. Single nedil 4. Single nedil followed by ottru	1. Double kuril 2. Double kuril followed by nedil 3. Double nedil 4. Kuril followed by nedil 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 and a sample of which is indicated in Table 1. 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.

Table 2. Asai seer (seer with one asai)

Asai Seer	i)Naal - Nerasai
	ii)Malar - Niraiasai
	iii)Kasu- Nerpu (Nerasai followed by ukaram)
	iv)Pirappu – Niraipu (Niraiasai followed by ukaram)

Table 3. Iyar seer (seer with two asais)

Asai	Seer
Ner Ner	Thema
Nirai Ner	Pulima
Nirai Nirai	Karuvilam
Ner Nirai	Kuvilam

Table 4. Uri seer (seer with three asais)

Asai	Seer
Ner Ner Ner	Themankai
...	...
Ner Nirai Nirai	Kuvilankani

Table 5. Asai seer (seer with four asais)

Asai	Seer
Ner Ner NerNer	Themanthanpu

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, and by comparing with Table 1, each word is also assigned the corresponding name and is stored in the intermediate file.

4.3. Thalai Computation

Tamil is a phonetic language in which alphabets are formed by combining vowels and consonants. The occurrence of connected feet (seer) in poetry is called ‘thalai’. As can be observed from Table 1, every seer has a fixed ending, namely ‘maa’, ‘vilai’, ‘kai’, ‘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. Table 2 shown below represents this classification, grouped based on the poetry class.

Table 6. Thalai rules

Thalai	Rules
i)Nerondriya Aasiriyathalai	Ma cheer before Ner asai
ii)Niraiondriya Aasiriyathalai	Vila cheer before Nirai asai
iii)Iyarcheer Venthalai	Ma cheer before Nirai asai and Vila cheer before Ner
iv)Vencheer Venthalai	Kaai cheer before Ner asai
v)Kalithalai	Kaai cheer before Nirai asai
vi)Ondriya Vanchithalai	Kani cheer before Nirai asai
vii)Ondratha Vanchithalai	Kani cheer before Ner asai

These rules are later mapped and used for Thalai identification.

4.3. Paa Identification

As already indicated in Figure 1 of Section 3, in this paper we discuss classifying a poem into Venpaa, Aasiriyapaa and Vanchipaa. We consider identification of Venpaa first.

4.3.1 Venpaa Identification

The general rules for a Venpaa which we derived from literature [3] are that a typical Venpaa should belong to Venthalai and it is further emphasized that no other thalai is permitted. There is a restriction on the maximum number of lines allowed for a poetry to be Venpaa and it is 12. Besides restricting the number of lines, the concluding line needs to be sindhadi and the remaining lines need to be alavadi.(A sindhadi is one in which the line is made up of 3 seer while an alavadi is a line made up of 4 seers). Another prominent characteristic of Venpaa is its musical nature which is termed as Cheppalosai. In this work, we identify poem as Venpaa and its sub-class and which is performed based on the following classification. Kural Venpaa, Sindhiyal Venpaa, Alaviyal Venpaa, Bahrodai Venpaa obey rules of Venpaa, but Savalai Venpaa do not strictly bind to the grammar of Venpaa. Table 3.1 gives Venpaa sub class rules.

Table 6. Venpaa rules

Kural Venpaa	<ul style="list-style-type: none"> • Must have exactly 2 lines and first must be Alavadi and the second line must be Sindhadi. • Last seer of last line must be a Naal, Malar, Kasu or Pirappu as given in Table 1.1.
Sindhiyal Venpaa	<ul style="list-style-type: none"> • Must have three lines • First 2 lines should be Alavadi and the last should be Sindhadi
Alaviyal Venpaa	<ul style="list-style-type: none"> • It must have four lines • First 3 lines should be Alavadi and the last line should be Sindhadi.
Bahrodai Venpaa	<ul style="list-style-type: none"> • It must have minimum of five lines and maximum of twelve lines.
Savalai Venpaa	<ul style="list-style-type: none"> • Result of combining two Kural Venpaas

4.3.2 Aasiriyapaa Identification

It is the right poem class to express the emotions in love, bravery in war etc. Aasiriyapaa as explained in [3] is characterized by Aasiriyathalai. Occurrences of other thalais are also permitted. It must have minimum of three lines. Unlike Venpaa, there is no limitation on the maximum number of lines. The last asai of last line must end with any one of Ae, O, En, Ee, Aa, Aai,Ai. Agavalosai delineates the musical sound of Aasiriyapaa. Due to the ambiguity in the class of Aasiriyapaa as discussed above, identifying it by a machine becomes difficult. Table 3.2 gives sub class Aasiriyapaa rules.

Table 7. Aasiriyapaa rules

Nerisai Aasiriyapaa	The penultimate line must be Sindhadi and the remaining lines must be Alavadi
Nilaimandila Aasiriyapaa	All lines must be of Alavadi
Adimari Mandila Aasiriyappa	Meaning of the poem remains unaltered even if the lines are interchanged
Inaikkural Aasiriyapaa	The first and the last line must be of Alavadi and the inbetween lines may be of Sindhadi or Alavadi

4.3.3 Vanchipaa Identification

Vanchipaa is characterised by the rules as derived from [3] that it must have Vanchithalai and it must have minimum of three lines. It must end with 'vanchipa thanisol' (singli word) and Aasiriyasurithagam (The penultimate line must be of three seers and the last line must be of four lines). The musical sound of vanchippa is Thoongalosai. Vanchipaa is classified as follows:

Table 8. Vanchipaa rules

Kuraladi Vanchipaa	It must have two cheer lines of three asai that end with kani.
Sindadi Vanchipaa	It must have three cheer lines of three asai that end with kani.

The CFG for all Paas is given below:

$G=(V,T,P,S)$

$V=\{VENPAA,KURAL_VENPAA,SINDHIYAL_VENPAA,ALAVIYAL_VENPAA,BAHRODAI_VENPAA,SAVALAI_VENPAA,AASIRIYAPPA,NERISAI_AASIRIYAPPA,NILAIMANDILA_AASIRIYAPPA,INAIKURAL_AASIRIYAPPA,EETRADI_1,EETRADI_2,VANCHIPPA,KURALADI_VANCHIPPA,SINDHADI_VANCHIPPA,KURALADI,ALAVADI,THANICHOL,SINDHADI,CHEER,EETRUCHEER,EERASAI,MOOVASAI,THEMAA,PULIMAA,KARUVILAM,KOOVILAM,THEMAANGAAI,PULIMAANGAAI,KOOVILANGAAI,KARUVILANGAAI,NAAL,MALAR,KAASU,PIRAPPU,NER,NIRAI\}$

$T= \{KURIL, NEDIL, OTTRU\}$

$S= \{VENPAA, AASIRIYAPPA, VANCHIPPA\}$

P is given as follows

<VENPPA> → <KURAL_VENPAA> | <SINDHIYAL_VENPAA> | <ALAVIYAL_VENPAA>
 | <BAHRODAI_VENPAA> | <SAVALAI_VENPAA>
 <KURAL_VENPAA> → <ALAVADI> <SINDHADI>
 <EETRUCHEER> → <NAAL> | <MALAR> | <KAASU> | <PIRAPPU>
 <SINDHIYAL_VENPAA> → <ALAVADI> <ALAVADI> <SINDHADI>
 <ALAVADI> → <CHEER> <CHEER> <CHEER> <CHEER>
 <ALAVIYAL_VENPAA> → <ALAVADI> <ALAVADI> <ALAVADI> <SINDHADI>
 <BAHRODAI_VENPAA> → <ALAVADI> {5, 11} <SINDHADI>
 <SAVALAI_VENPAA> → <KURAL_VENPAA> <KURAL_VENPAA>
 <AASIRIYAPPA> → <NERISAI_AASIRIYAPPA> | <NILAIMANDILA_AASIRIYAPPA> |
 <INAIKURAL_AASIRIYAPPA>
 <NERISAI_AASIRIYAPPA> → <ALAVADI> {2,} <EETRADI_1>
 <EETRADI_1> → <CHEER> <CHEER> <EETRUCHEER> //SINDHADI
 <CHEER> → <ORASAI> | <EERASAI> | <MOOVASAI>
 <NILAIMANDILA_AASIRIYAPPA> → <ALAVADI> {2,} <EETRADI_2>
 <EETRADI_2> → <CHEER> <CHEER> <CHEER> <CHEER> //ALAVADI
 <INAIKURAL_AASIRIYAPPA> → <ALAVADI> {<ALAVADI> | <SINDHADI>} {1,}
 <EETRADI_2>
 <VANCHIPPA> → <KURALADI_VANCHIPPA> | <SINDHADI_VANCHIPPA>
 <KURALADI_VANCHIPPA> → <KURALADI> <THANICHOL> <AASIRIYAPPA>
 <KURALADI> → <CHEER> <CHEER>
 <SINDHADI_VANCHIPPA> → <SINDHADI> <THANICHOL> <AASIRIYAPPA>
 <SINDHADI> → <CHEER> <CHEER> <EETRUCHEER>
 <THANICHOL> → <EERASAI> | <MOOVASAI>
 <CHEER> → <EERASAI> | <MOOVASAI>
 <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}

4.5 Osai Identification

A paa is characterised by its feet connection(thalai) and also by its sound(Osai). In this work we verify the identified paa by determining its Osai also. Osai can be classified as Cheppalosai, Agavalosai, Thoongalosai and Thullalosai which are described below:

4.5.1 Cheppalosai

It is the musical sound of Venpa kind of poems. Its sound is due to venthalai. After identifying the Venpaa and its sub-class we used the thalai that has been identified and determine whether it belongs to the category of Cheppalosai as Ozhugisai , Enthisai or Thoongisai Cheppalosai. Venpaa is said to possess the musical sound ‘Enthisai Cheppalosai’ if the occurrence of venseer venthalai exceeds that of iyarseer venthalai and ‘Thoongisai Cheppalosai’ if the vice-versa occurs. Another osai called ‘Ozhugisai Cheppalosai’ attributes to that Venpaa that contains equal number of iyarseer and venseer venthalais[3].

4.5.2 Agavalosai

It is the musical sound of Aasiriyapaa kind of poems. Its sound is due to aasiriyathalai.

Similar to Venpaa, the maximum occurrence of thalai determines its musical sound. ‘Enthisai Agavalosai’ features nerondriya aasiriyathalai, ‘Thoongisai Agavalosai’ features niraiondriya aasiriyathalai, and ‘Ozhugisai Agavalosai’ for the equal occurrences of above mentioned aasiriyathalais[3]. We designed rules and confirm the poem as Aasiriyapaa based on Agavalosai.

4.5.3 Thoongalosai

It is the musical sound of Vanchipaa kind of poems. It is due to Vanchithalais. ‘Enthisai Thoongalosai’ is specific to those vanchipaas that have highest number of ondriya vanchithalai, ‘Agaval Thoongalosai’ for those that have highest number of ondriya vanchithalai, and ‘Pirinthisai Thoongalosai’ for those that have equal number of the above said vanchithalais[3]. The presence of Vanchithalai in the poem is confirmed by Thoongalosai.

5. RESULTS AND ANALYSIS

The algorithm proposed in this paper works for all Thirukkural(1330) that are venpaa and all other ancient poems But, poems containing hyphenated words especially Venpaa do not produce expected results,, since the CFG has not been designed for it. Besides, poems that belong to the categories like ‘marabu kavithai’, ‘haiku’, ‘puthu kavithai’ are not guaranteed to work as per the proposed algorithm.

Table 9. Test cases

No of documents		Correctly identified	Not identified
Venpaa	1330(thirukural)+60(other)	1330+40	20
Aasiriyapaa	48	48	-
Vanchipaa	13	13	-

The above table gives the details of the system tested for various inputs.

6. CONCLUSIONS

The project is designed to identify the 3 Pa (Venpaa, Asiriyapaa, Vanchipaa) and further sub-classification of the identified Pa has been done. The corresponding Osai associated with the Paa is also determined. It can be further extended to identify and classify kalipaa. In addition, CFG could be modified to include hyphenated words to increase the efficiency. A suggestion system can be developed for a new poet so as to obey these grammar rules.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

- [1] <http://en.wikipedia.org/wiki/Tolk%C4%81ppiyam>
- [2] 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
- [3] K. Rajagopalachariyar, "Ilakkana-vilakkam/yappiyal", Kanappan publications
- [4] Hamid R. Tizhoosh, Farhang Sahba, Rozita Dara, "Poetic Features for Poem Recognition: A Comparative Study", Journal of Pattern Recognition and Research (2008) 24-39
- [5] Peter Crisp, John Heywood, Gerard Steen, "Metaphor Identification and analysis, classification and quantification", Language and Literature February 2002 11: 55-69
- [6] http://en.wikipedia.org/wiki/Context-free_grammar

Authors

Subasree Venkatsubramaniyen is a Computer Science student from College of Engineering, Guindy. Her interests include Algorithms, Data Structures and Machine learning.



Subha Rashmi is a Computer Science student from College of Engineering, Guindy. Her interests include Language Technologies, Data Mining, and Compiler Theory.



Rajeswari Sridhar has a Ph.D in Computer Science and engg. She has more than 18 publications and international conferences. Her include language technologies, NLP, Information Retrieval, Music Signal Processing, Data Structures and Compilers. She is currently working as Asst. Professor (Senior Grade) at College of Engineering, Guindy.

