ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2020.0320

VERB IDENTIFICATION USING MORPHOPHONEMIC RULES IN TAMIL LANGUAGE

M. Mercy Evangeline and K. Shyamala
Department of Computer Science, Dr. Ambedkar Government Arts College, India

Abstract

Part of Speech (POS) tagging is a general process of classifying words into their parts, labeling them into different categories like Nouns, Verbs, Adjective, and Adverbs and so on. These different categories of words can be defined in a tagset. The tagset can be used for automatically assigning a word with a tag. Noun forms the first category of words generally found in any sentence and verb comes next. A verb actually describes the event occurrence or an action.in Tamil language, a verb form gets inflected by suffixes based on person, count, tense and voice. These suffixes are identified by reverse splitting method and the word is tagged as Verb. In this paper, tagging of words for Tamil language particular to Verb has been carried out. The implementation involves a rule based suffix stripping method for identifying verbs where suffixes are checked with the grammatical rules and are tagged as verb. The implementation proposed uses the traditional way of identifying a word based on grammatical rules in Tamil language, thus avoiding the process of transliteration. The implementation identifies a word based on grammatical rules, applies reverse splitting method and categorizes the words as VERB. The input is considered as Tamil words in their Unicode format, thus avoiding the process of transliteration. Most of the work done in the area of Text mining of Tamil documents mainly involves transliteration. Applying Tamil grammatical rules enrich the identification and tagging of words during morphological analysis as morphophonemic rules are considered. This is an advantage while tagging of words is considered for Tamil documents.

Keywords:

Morphological Analyzer, Tagging, Verb, Tamil Language, Classification, Identification

1. INTRODUCTION

In Tamil language, generally there are four kinds of words available. They are uriccol (*உரிச்சொல்*), peyarccol (*பெயர்ச் சொல்*), vi̲naiccol (*வினைச்சொல்*) and iṭaiccol (*இடைச்சொல்*) [8]. In this category, *உரிச்சொல்* are words which were mainly used in ancient poems. So excluding it, there are three main categories for the word classification in Tamil. Iṭaiccol (*இடைச்சொல்*) are words which don't have meaning of their own. Only when they are combined with other words they have a meaning. They can even change the meaning of the word to which they are added. Remaining are the two kinds: peyarccol (*பெயர்ச்சொல்*) and vi̲naiccol (*வினைச்சொல்*). Nouns (*பெயர்ச்சொல்* - peyarccol) form one of the important and significant word classes in Tamil language. According to S.A. Pillai, the noun classification generally available includes Animate and Inanimate nouns [13]. Within these main categories, we have subclasses available. The Animate noun have been further classified into Singular and Plural forms, Masculine, Feminine, Honorific and plural markers. The Inanimate nouns are classified as Mass nouns and Count nouns which include Singular and Plural forms. These Noun forms have different declensions rules which they can have.

Verb (*வினைச்சொல்* - Vi̲naiccol) forms the next category of words found in a sentence for Tamil Language. These are the words which describe the action. Every sentence will have a subject and the verb describes the action of the subject. Depending upon what the subject does, a verb can take different forms and tenses. As Tamil Nouns, Tamil verbs also get inflected by suffixes [17]. The different forms of suffixes for verb include suffixes for person, count, mood, tense and voice.

- Suffixes indicating person and number are indicated by the oblique case of the relevant pronoun. The suffixes indicating tenses and voice are formed by adding particles to the stem. The particles are generally defined as terms which cannot be generally inflected and it is also termed as word which when associated with another word will deliver a meaning.
- There are two voices indicated by Tamil grammar. The first one indicates that the subject of the sentence undergoes the action or it is the object of action defined by verb stem; the second one indicates that the subject refers to the action indicated by the verb stem.
- Verb is also inflected by three simple tenses - past, present, and future. These tenses are given by either simple suffixes or compound one.

1.1 DRAVIDIAN LANGUAGES

Dravidian Languages is considered to be a family of nearly 70 different languages spoken primarily in South Asia [10]. These languages are basically classified as South, South-Central, Central and North Central groups. They are spoken by more than 215 million people in India, Pakistan and Sri Lanka [10]. The 70 languages are subdivided into many groups.

India is generally considered as a geographical area with a varied variety of linguistically different languages. Out of these groups, the four major literary languages are Tamil, Telugu, Malayalam and Kannada and they are the official language of Tamil Nadu, Andhra Pradesh, Kerala and Karnataka. These four languages are also known as literary languages. Of these four literary languages Tamil is considered to be of the oldest language. It has been declared as a classical language of India, with three criteria's met; the origin is ancient, with an independent tradition and possessing considerable body of ancient literature.

These Dravidian languages have a phonologically rich language system. Out of these languages, Tamil is considered to be one of the oldest languages dating to early Common Era. In 2004, Tamil was declared as a classical language based on three main criteria: the origin of the language is ancient, it follows an independent tradition and it possesses a considerable body of ancient literature [11]. Among these Dravidian languages, many linguistic features are noticeably common. Few thousand words are common, some grammatical forms are common and it goes on.

1.2 TAMIL LANGUAGE

Tamil takes a most distinctive place among all other Dravidian language because of its geographical expansion and its wide spread beyond the frontiers of India. Tolkappiyam is considered to be an ancient grammatical work in Tamil Language. It basically has three sections: first two discuss about the linguistics of the language and the third section deals with Tamil literature.

Nannool is another ancient work available in Tamil language. It is the most significant work available for Tamil after Tolkappiyam. Nannool is divided in five sections: the written language, spoken language, semantics, poetical language and rhetorical devices [7]. The last three divisions have been lost due to the changes in the growth of the language. The other two divisions, the written and spoken language are available now.

1.3 WORD CLASSIFICATION IN TAMIL LANGUAGE

According to Tolkappiyam, Tamil words are classified into four categories. They are Iyarcol (இயற்சொல்), tiricol (திரிசொல்), ticaiccol (திசைச்சொல்) and vatacol (வடசொல்). Iyarcol (இயற்சொல்) are words which very common in use and easily understood by everyone. They are again classified into peyar iyarcol (பெயர் இயற்சொல்) and Vinai iyarcol (வினை இயற்சொல்). Tiricol (திரிசொல்) consists of words which are used in Poetry. Ticaiccol (திசைச்சொல்) are words which are borrowed languages like Urdu, Portuguese, which are spoken in other regions. Vaṭacol (வடசொல்) are words which borrowed from Sanskrit. All the words which are borrowed from Sanskrit have to follow certain rules specifically to be used in Tamil. They have to strictly confirm to the phonetic system of Tamil language and have to be written in the Tamil script.

2. LITERATURE SURVEY

Kengatharaiyer et al. [1] have designed a Morphological Analyzer and Generator for Tamil language. It initially covers for Tamil verbs only. A Finite State Transducer was used to develop the Generator [1]. This work was developed as an extension for computation grammar using Lexicon Functional Grammar. According to this work, rule based approach like the above mentioned one proves to be very effective that going on formal machine learning alternatives. This also provides good accuracy. The system was developed as three-level web based system for handling issues which can be expected when dealing with agglutinative language like Tamil. It shows a higher level of accuracy for simple verbs in Tamil.

Dokkara et al. [2] have proposed an engine for morphological generation of verbs inflection required for a given word form in Telugu Language [2]. It is based in finite state technique, a computational method to generate inflectional verb forms. The input is given in the form of an XML file which has all the rules pertaining to a verb which is applicable at the sentence level or word level. This input file is used by the engine for generating sentences which are well-formed grammatically in Telugu language. In this methodology, when a word is tested, the engine identifies the morphophonemic group, secondly the category of verb inflection. From this it identifies the phonetic alterations for the particular word depending upon the group and its class. Then it adds the tense mode and personal suffix. Finally the verb inflectional form for the given word in generated. This engine was tested on 508 verbs and it has been identified that 83% of the words were extracted correctly and the remaining percent didn't fall into a particular verb category as they were not included in the grammar reference which was considered for this work.

Nimal et al. [3] have proposed a Morphological analyzer for Verb and Noun in particular to Malayalam Language [3]. In this work, a Rule and Dictionary based approach is adopted along with suffix stripping concept. In this method, a dictionary of root word and all forms of its inflection are maintained. These inflections are obtained by using morphophonemic changed available for available language. For a given input, first it is transliterated into English word. Then, it is checked for the availability in dictionary. If not available, multiple suffix stripping process is applied. During this process of suffix stripping, sandhi rules are used and the resulting root word is checked for availability in the dictionary. Then the word is retransliterated back to Malayalam. If not available in the dictionary, again the suffix extraction method is adopted.

Sivaneasharajah et al. [4] have implemented a Morphological Analyzer/Generator for Tamil Language [4]. It is based on Two-Level Morphology system. For this work, Orthographic rules have been considered along with Lexicon. These rules have been written as regular expressions, by applying only finite state operations. For this implementation, seven different forms of Noun class and 11 different class forms for Verb are considered. It also includes automatic transliteration scheme for encoding and decoding. It makes use of transducers for identifying both Noun and Verb. These transducers work on basis of Orthographic rules for Tamil Language.

Kumar, R. et al. [5] have implemented a Rule based Machine Aided Translation (MAT). Verb classification was mainly considered to create a bilingual English Malayalam dictionary for the MAT [5]. In Malayalam language verb can be formed with ten suffix attachments. Splitting them is complicated process. Thus the analysis of Verb in Malayalam is a complex task. For this implementation, Verb Classification done was considered for the MAT.

3. IDENTIFICATION OF VERBS USING MORPHOPHONEMIC RULES IN TAMIL LANGUAGE

Tamil inhabits a distinctive position in Dravidian language as it expands beyond the frontier of India [16]. It is not only native to Tamil Nadu, but also Tamil people living in Ceylon, Burma , Malaysia, Singapore and some other places in south Asia. Tolkāppiyam (தொல்காப்பியம்) is considered to be the ancient grammatical work available in Tamil. Words are classified as Iyarcol (இயற்சொல்), tiricol (திரிசொல்), ticaiccol (திசைச்சொல்) and vaṭacol (வடசொல்) in Tamil Language based on the nature and place. Iyarcol இயற்சொல் refers to the words in common use like Maram -மரம், vantāṉ - வந்தான், malai - மலை, kaṭal - கடல். Iyarcol (இயற்சொல்), tiricol (திரிசொல்) refers to words which are used very specifically in poetry like Āḻi - ஆழி, ceppiṉāṉ - செப்பினான், kiḷḷai - கிள்ளை, tattai -தத்தை, cukam - சுகம். Ticaiccol (திசைச்சொல்) refers to words which are regional to the language

[9]. It includes words which have come from languages which are spoken in other directions of Tamil Nadu. Some examples include Ācāmi (ஆசாமி), cāvi (சாவி), pāḻi (பாழி). Vatacol (வடசொல்) refers to the words which were borrowed from Sanskrit. It mainly constitutes words which were formed with letters which were common to Tamil and Aryan language, special characters and specific characters to these languages. Some example includes hari (ஹரி), pakṣi (பக்ஷி), kaṭiṉam (கடினம்).

Another way of classification of word in Tamil language are Nouns (பெயர்ச்சொல்), Verb (வினைச்சொல்), Preposition and postposition (இடைச்சொல்), Adjective (உரிச்சொல்) [12]. The nouns generally indicate person, gender, number and the animate and inanimate grouping of things. The verb generally includes action done by subject term. It generally refers to a words action or profession. The categories of the subject term will fall under the category of person, animate and inmate. This is indicated by Tiṇai (திணை). (Tiṇai) திணை is generally classified into Uyartiṇai (உயர்திணை), aḫṟiṇai (அஃறிணை). Uyartiṇai - உயர்திணை includes words denoting persons and aḫṟiṇai - அஃறிணை includes words describing animate, inanimate and neuter category.

Persons are indicated using three genders, Masculine, Feminine and neuter. It also includes Plural forms representing many in numbers. In Tamil language, masculine and feminine represents singular form [16]. Persons are also represented as first person, second person and third person depending upon the way a person is addressed in the sentence.

Verb classification for this implementation is based on the different form of tense markers with the stem words. Verb also includes words which represent human being/person, animate and inanimate things [6].

The Table.1 gives the representation for all the three tenses for a word. The tense include the present, past and future tense. It also gives an overall identification of a word under different categories of person which includes masculine, feminine, honorific and irrational. It also shows the form of word in singular and plural forms under each category of tense.

The Fig.1 gives a representation for Singular forms of person - Masculine, Feminine, Rational and Singular form for irrational subjects. The representation gives the different state of the subject with respect to tenses like present, past and future tense. The various indications of the FSA are discussed below.

The 1st and 2nd Singular form of the subject end with -Ēṉ (ஏன்) and āy (ஆய்). Singular form of person in Masculine, Feminine and Rational end with - āṉ (-ஆன்), āḷ (-ஆள்), -ār (-ஆர்), which are termed as paṭarkkai viṉaimuṟṟu vikuti (படர்க்கை வினைமுற்று விகுதி) for Masculine, Feminine and Rational form.

The 3rd Singular form of Irrational subject comes with a suffix -atu or -um (-அது or -உம்). Kiṟu (-கிறு), t (த்), v (வ்) represents the tense form of the subject.

Table.1.Verb representation for different tense and person

Person	Present	Past	Future
1st Singular	-kiṟēṉ -கிறேன்	-tēṉ -தேன்	-vēṉ -வேன்
2nd Singular	-kiṟāy	-tāy	-vāy
	-கிறாய்	-தாய்	-வாய்
3rd Singular Masculine	-kiṟāṉ -கிறான்	-tāṉ -தான்	-vāṉ -வான்
3rd Singular Feminine	-kiṟāḷ -கிறாள்	-tāḷ -தாள்	-vāḷ -வாள்
3rd Singular Honorific	-kiṟār -கிறார்	-tār -தார்	-vār -வார்
3rd Singular Irrational	-kiṟatu -கிறது	-tatu -தது	-yum -யும்
1st Plural	-kiṟōm -கிறோம்	-tōm -தோம்	-vōm -வோம்
2nd Plural	-kiṟīrkaḷ -கிறீர்கள்	-tīrkaḷ -தீர்கள்	-vīrkaḷ -வீர்கள்
3rd Plural Rational	- kiṟārkaḷ -கிறார்கள்	-tārkaḷ -தார்கள்	-vārkaḷ -வார்கள்
3rd Plural Irrational	-kiṉṟaṉa -கின்றன	-taṉa -தன	-vaṉa -வன

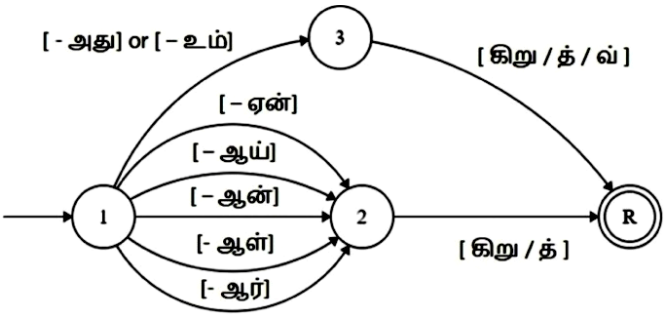

Fig.1. FSA for Singular form of representation for different tense

The Fig.2 representation shows the different form of subject in Plural. -kal (-கள்), -aṉa (-அன), -ōm (-ஓம்) represent the Paṉmai viṉaimuṟṟu (பன்மை வினைமுற்று). -kiṉṟu/t/v (-கின்று/த்/வ்), -kiṟu/t/v (கிறு/த்/வ்) are tense representation of the subject (Kālaṅkāṭṭum iṭainilai - காலங்காட்டும் இடைநிலை).

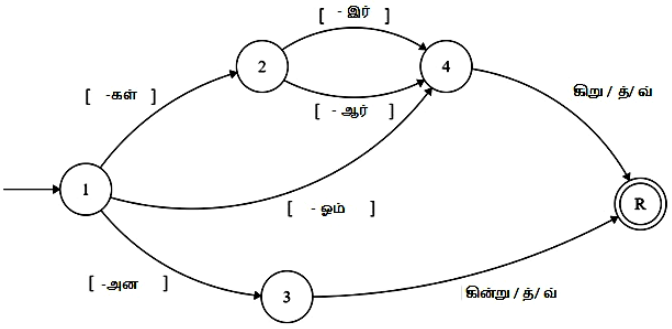

Fig.2. FSA Representation for Plural Form of Subject for different Tense

In Fig.3, a word is checked for 1st Singular form of verb. It shows the different forms taken in tense form which includes Present, Past and Future tense. The suffixes are checked for suffix patterns using reverse splitting method. If it falls under anyone of these categories they are defined as verb.

For example,

Ceykiṟēṉ - ceytēṉ - ceyvēṉ (*செய்கிறேன் - செய்தேன் - செய்வேன்*) represents 1st singular form for the root word cey (*செய்*) in all the three tenses.

- Ceykiṟēṉ (*செய்கிறேன்*) - cey + kiṟu + ēṉ [*செய் + கிறு + ஏன்*]
- Ceytēṉ (*செய்தேன்*) - cey + t + ēṉ [*செய் + த் + ஏன்*]
- Ceyvēṉ (*செய்வேன்*) - cey + v + ēṉ [*செய் + வ் + ஏன்*]

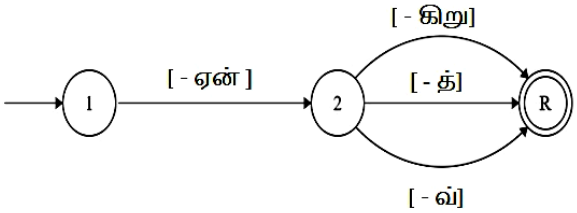

Fig.3. FSA Representation of 1st Singular Form

In Fig.4, word is checked for 1st Plural form of verb. It shows the different forms taken in tense form which includes present, past and future tense. Reverse splitting method is used for checking the pattern. If it falls under anyone of these categories they are defined as verb.

For example,

- Ceykiṟōm - ceytōm - ceyvōm (*செய்கிறோம் - செய்தோம் - செய்வோம்*) represents 1st plural form for the root word *செய்* in all the three tenses.
- Ceykiṟōm (*செய்கிறோம்*) - cey + kiṟu + ōm [*செய் + கிறு + ஓம்*]
- Ceytōm (*செய்தோம்*) - cey + t + ōm [*செய் + த் + ஓம்*]
- Ceyvōm (*செய்வோம்*) - cey + v + ōm [*செய் + வ் + ஓம்*]

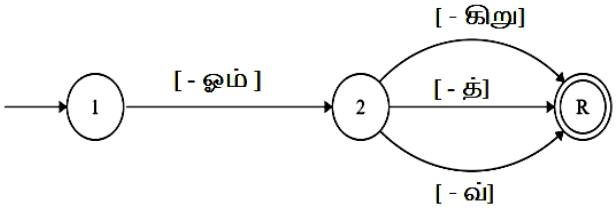

Fig.4. FSA Representation of 1st Plural Form

In Fig.5, a word is checked for 3rd form of verb representation for Masculine, Feminine and Honorific person. It shows the different forms taken in tense form which includes Present, Past and Future tense. The suffixes are checked for suffix patterns starting from the last character. If it falls under anyone of these categories they are defined as verb.

For example,

- Ceykiṟāṉ - ceytāṉ - ceyvāṉ (*செய்கிறான் - செய்தான் - செய்வான்*) represents masculine form for the root word *செய்* in all the three tenses.

Ceykiṟāṉ (*செய்கிறான்*) - cey + kiṟu + āṉ (Paṭarkkai āṇpāl)

- [*செய் + கிறு + ஆன்* (*படர்க்கை ஆண்பால்*)]

Ceytāṉ - cey + t + āṉ [*செய்தான் - செய் + த் + ஆன்*]

Ceyvāṉ - cey + v + āṉ [*செய்வான் - செய் + வ் + ஆன்*]

In the above example, cey (*செய்*) represents the verb root (Viṉaippakuti -*வினைப்பகுதி*), Kiṟu, t, v (*கிறு, த், வ்*) represents the period or tense for the root word (Kālam kāṭṭum iṭainilai - *காலம் காட்டும் இடைநிலை*) and - Āṉ (-*ஆன்*) represents Paṭarkkai uyartiṇai āṇpāl viṉaimuṟṟu vikuti (*படர்க்கை உயர்திணை ஆண்பால் வினைமுற்று விகுதி*).

- Ceykiṟāḷ - ceytāḷ - ceyvāḷ (*செய்கிறாள் - செய்தாள் - செய்வாள்*) represents feminine form for the root word cey (*செய்*). - Āḷ (-*ஆள்*) represents Paṭarkkaip peṇpāl viṉaimuṟṟu vikuti (*படர்க்கைப் பெண்பால் வினைமுற்று விகுதி*), cey (*செய்*) represents the verb root (vinaippakuti - *வினைப்பகுதி*), - Kiṟu, -t, -v (-*கிறு, -த், -வ்*) represents the period or tense for the root word (Kālam kāṭṭum iṭainilai - *காலம் காட்டும் இடைநிலை*).
- Ceykiṟār - ceytār - ceyvār (*செய்கிறார் - செய்தார் - செய்வார்*) represents honorific form for the root word cey (*செய்*), - Ār (-*ஆர்*) represents Paṭarkkaip palarpāl viṉaimuṟṟu vikuti (*படர்க்கைப் பலர்பால் வினைமுற்று விகுதி*), cey (*செய்*) represents the verb root (Viṉaippakuti - *வினைப்பகுதி*), Kiṟu (*கிறு*),t (*த்*), v (*வ்*) represents the period or tense for the root word (Kālam kāṭṭum iṭainilai - *காலம் காட்டும் இடைநிலை*).

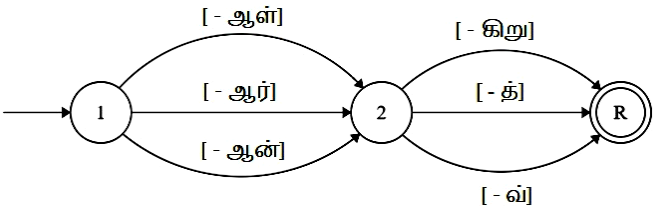

Fig.5. FSA representing 3rd Singular form for Persons including Masculine/Feminine/Honorific

Masculine and Feminine form of verb cannot take plural form. They only fall into singular category. Rational and irrational forms take the plural form. The Fig.6 and Fig.7 shows the plural form taken by Rational and irrational words. The suffixes are checked for suffix patterns and if it falls under anyone of this category they are defined as VERB.

For example, rational verb can be represented as follows:

- Ceykiṟārkaḷ (*செய்கிறார்கள்*) - cey + kiṟu + ār + kal [*செய் + கிறு + ஆர் + கள்*]
- Ceytārkaḷ (*செய்தார்கள்*) - cey + t + ār + kal [*செய் + த் + ஆர் + கள்*]
- Ceyvārkaḷ (*செய்வார்கள்*) - cey + v + ār + al[*செய் + வ் + ஆர் + கள்*]

-kal (-*கள்*) represents the Īṟu (*ஈறு*) (vikuti - *விகுதி*) of the word and in this example it indicates the plural form of the word, -Ār (-*ஆர்*) indicates that the root is representing third person (patarkkaip palarpal - *படர்க்கைப் பலர்பால்*), - Kiṟu, -t, -v (-*கிறு, -த், -வ்*) represents the tense markers (Kālam kāṭṭum iṭainilai - *காலம் காட்டும் இடைநிலை*) and cey (*செய்*) is the word identified as root (Viṉaiyaṭi -*வினையடி*) and tagged as verb.

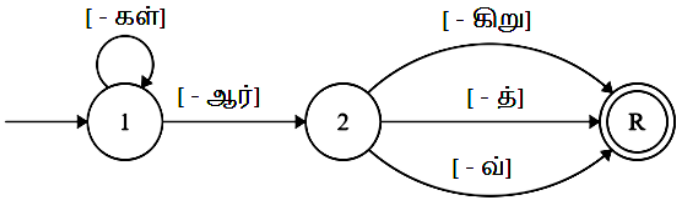

Fig.6. FSA representing 3rd Plural form for rational words

For example, irrational verb can be represented as follows:

- Ceykiṉṟaṉa (*செய்கின்றன*) - Cey + kiṉṟu + aṉa [*செய் + கின்று + அன*]
- Ceytaṉa (*செய்தன*) - Cey + t + aṉa [*செய் + த் + அன*]
- Ceyvaṉa (*செய்வன*) - cey + v + aṉa [*செய் + வ் + அன*]

-ana (-*அன*) represents the Īṟu (*ஈறு*) (vikuti - *விகுதி*) of the word which is defined as Aḥriṇaip paṉmai viṉaimuṟṟu vikuti (*அஃறிணைப் பன்மை வினைமுற்று விகுதி*), -Kiṟu, -t, -v (-*கிறு*, -*த்*, -*வ்*) represents the tense markers (kalam kattum itainilai - *காலம் காட்டும் இடைநிலை*) and cey (*செய்*) is the word identified as root and tagged as verb.

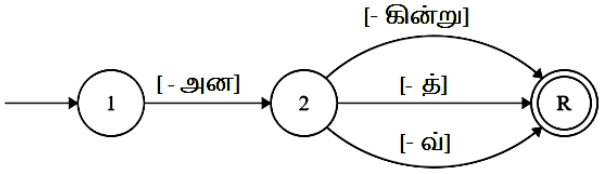

Fig.7. FSA representing 3rd Plural form for irrational words

Iṭainilai (*இடைநிலை*) always represents the tense of the word. Gener ally it may be present, past or future tense. T, ṭ, ṟ, iṉ (*த், ட், ற், இன்*) always indicates the past tense marker. Āniṉṟu, kiṉṟu, kiṟu (*ஆநின்று, கின்று, கிறு*) indicates the present tense marker. Āniṉṟu (*ஆநின்று*) is not in usage with the present language way. -p, -v (*ப், வ்*) indicates the future tense marker.

4. VERB IDENTIFICATION USING MORPHOPHONEMIC RULES

The proposed algorithm Verb Identification using Morphophonemic Rules (VIMR) gets an intermediate text file which has gone through different modules in the processing of the file. The input file at the initial stage consists of text file which was created in UTF-8 format. Fig.8 shows the different modules of the implementation. There are some pre-available modules before input file comes for verb identification module.

These files go through the first module which is the pre-processing step. This pre-processing step involves tokenization and removal of stop words from the text file. The removal of stop word is done using the algorithm Dictionary Based Stop Word Removal Algorithm (DBSWRA) which was proposed earlier. This algorithm is based on dictionary based method where stop words are defined from basic understanding of the language. It also includes words which have been defined in various corpora.

The next step involves two stages. The first stage is identification of Nouns for the given input file. The algorithm Noun Identification for Tamil Language using Morphophonemic rule (NIMR) [14].

The first stage of this module identifies Nouns for the text file using morphophonemic rules. The pronouns are identified using predefined tagset. This tagset consists of pronoun words which are defined in Tamil language grammar. The output of this module is a text file which contains list of words which are identified as Noun and other words which are tagged as unidentified word category.

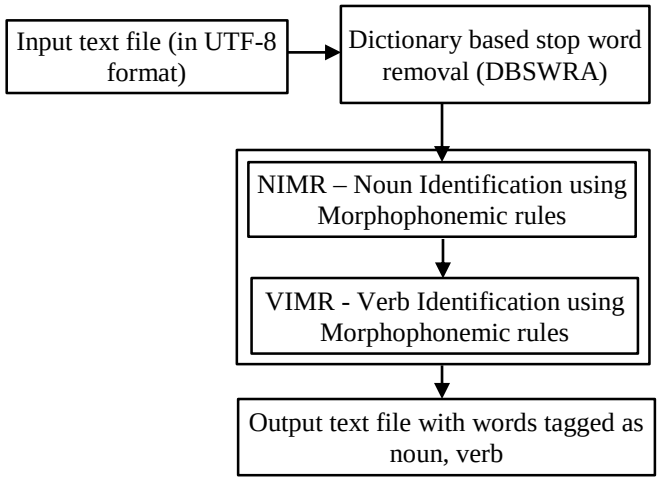

Fig.8. Categorization of Words using Morphophonemic Rules

The next module in this stage is the VIMR which gets as input the list of words which have not been tagged in the previous stage i.e. NIMR. The set of words are analyzed for the rules and if identified are tagged as Verb. The output after this stage consists of a text file which consists of words which are tagged as Nouns and Verbs. It will also have some words which have been missed in these two stages.

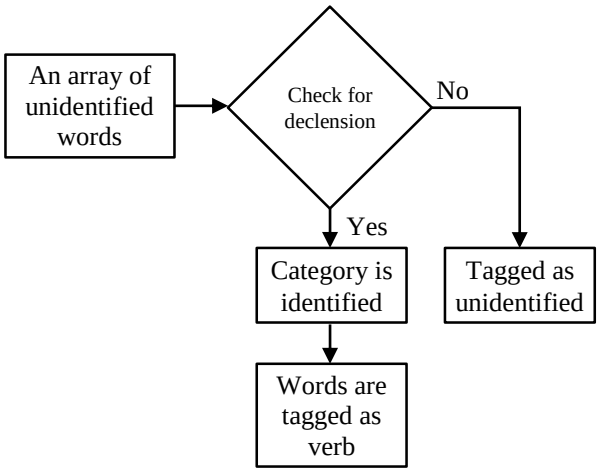

Fig.9. Flow Diagram for Verb Identification in Tamil Language using Morphophonemic Rules (VIMR)

The Fig.9 shows the flow of algorithm. The array of words is checked for declension. If available, then they are identified for the corresponding category and then tagged as VERB. If not available then the words are tagged as UID. The declension check is based on the suffixes of the words being analyzed. If the suffix pattern matches that of verb suffix definition, then reverse

splitting method is adopted to find the root word. The word which falls under these categories is tagged as VERB.

The algorithm takes an array of words as input. These are the words which were tagged as undefined words (UID) in the previous module i.e. the Noun Identification module (NIMR). This module is the third level of implementation in tagging of words. Initially start with a text file which is saved in UTF-8 format. These files are generally Tamil stories which are being processed and tagged as nouns and verbs.

Initially the text file is pre-processed. Then the output which is received as an array of words are checked for nouns and tagged as three categories - noun, pronoun and UID. The UID are words which have not fallen into the other two categories.

Algorithm: Verb Identification using Morphophonemic Rules: (VIMR)

Input: An array of words which were not tagged in the previous step

Output: An array of words identified as verbs.

Step 1: The input contains a list of words which weren't tagged in the previous step (NIMR)

Step 2: This module identifies verb in its different declension form. This procedure includes the following steps:

a. For each word in the list of words

b. Rules are checked for the word suffix condition

c. If satisfied; Tag Word as Verb

Step 3: The output is generated as a list of words identified as verbs.

The algorithm takes an array of words as input. These are the words which were tagged as undefined words (UID) in the previous module i.e. the Noun Identification module (NIMR) [19]. Each word is analyzed with reverse splitting technique where patterns are checked from the last character. If the pattern matches with any of the rules defined for verbs, then they are tagged as Verb, otherwise they are tagged as undefined word. In this implementation the traditional way of identifying verbs using the grammatical rules of the Tamil language is followed. The words are checked for their suffixes, particularly to verbs which have suffixes satisfying the grammar rules for vi<u>n</u>aimu<u>rr</u>u (வினைமுற்று). Those words are tagged as Verbs in this implementation.

5. RESULTS AND DISCUSSIONS

For the implementation of VIMR algorithm, text files in UTF-8 format were considered. First the text was pre-processed using DBSWRA. The output generated was a list of words. This output was passed through NIMR for identification of nouns. The Table.2 gives a review of files considered for implementation, the percentage of words identified correctly by VIMR compared to the manual method.

The Fig.10 shows the graphical representation of the words identified as verb by VIMR. The algorithm has done the identification using the grammar rules for Tamil Language. VIMR algorithm has considered the tense based rules of morphology for identification. During this process of identification, some words which are not a verb have been identified as verb by VIMR. This is because of the transition they have in their suffix. But the percentages of these words are very minimal.

The performance for the VIMR module proposed is compared between identified values and the actual values. A better way to evaluate the performance of the module is by defining confusion matrix and evaluating different metrics. It is a table representation which describes the performance of the model. The basic terms which are defined by Confusion matrix are True Positives (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These terms are general in the form of whole number. Based on these terms, the metrics which can be calculated include Accuracy, Precision, Prevalence, and F-score and so on. For the module considered here, the confusion matrix is given in Fig.11. Here the total data includes the words which are considered for identification module.

Table.2. VIMR implementation for Verb identification

File Name	Total Number of words excluding Nouns	Verbs Identified Manually	Verbs correctly Identified by VIMR	Non-Verbs identified as Verb by VIMR
File 1	268	23	16	2
File 2	219	18	15	4
File 3	257	39	37	5
File 4	276	26	21	3
File 5	165	16	9	1

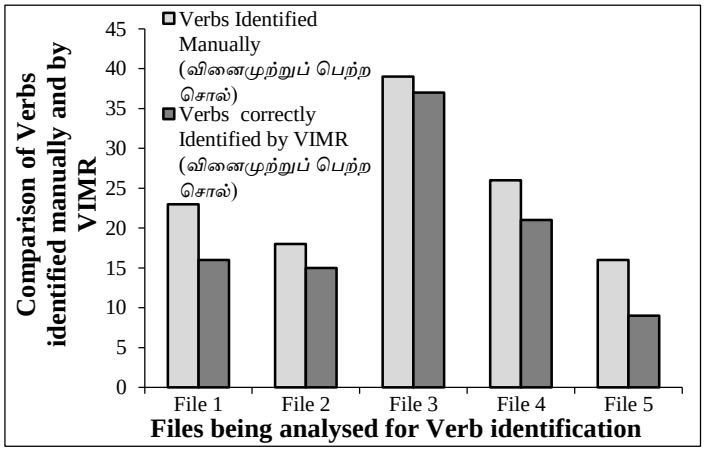

Fig.10. Graphical Representation of Verb Identification in Tamil Language using Morphophonemic Rules (VIMR)

Table.2. Definition of Confusion Matrix terms for the proposed model

N = Number of data considered	Predicted : NO (0)	Predicted : YES (1)
Actual : NO (0)	Non-verbs - not identified as verb (TN)	Wrongly identified as Verb (FP)
Actual : YES (1)	Verbs - missed to be identified (FN)	Verbs - correctly identified (TP)

Table.3. Tabulation of different metrics based on Confusion Matrix terms for the Input considered

File name	Verbs - correctly identified TP	Non-verbs - not identified as verb TN	Wrongly identified as Verb FP	Verbs - missed to be identified FN	Predicted Yes	Accuracy	Precision	Recall	F-Score
File 1	16	241	2	9	18	95.90%	88.89%	64.00%	74.42%
File 2	15	193	4	7	19	94.98%	78.95%	68.18%	73.17%
File 3	37	208	5	7	42	95.33%	88.10%	84.09%	86.05%
File 4	21	244	3	8	24	96.01%	87.50%	72.41%	79.25%
File 5	9	147	1	8	10	94.55%	90.00%	52.94%	66.67%

True Positive gives the whole term of words which were correctly tagged as verb by the VIMR module. False Positive gives the number of words which were wrongly tagged as VERB by the module. These two totally sum up to predict YES count. The True Negative term defines the non-verbs which were correctly tagged as unidentified. False Negative gives the count of words which were missed by the VIMR module to be tagged as verb. The Table.3 gives the values for the different terms in the Confusion Matrix for the input files considered for training the module. Based on these terms, other metrics like Accuracy, Precision, Recall and F-Score are calculated. The F-score gives the weighted average of true positive rate and precision.

6. CONCLUSION

VIMR algorithm has been designed with modules pertaining to different grammatical rules available for Tamil Language. The rule specifications have been taken from the book of Nannool. For obtaining the root word, the suffixes attached with them are removed. The grammatical rules considered for removing these suffixes include the rule defined according to morphophonemic change taken by a word when it combines with adjacent morphemes. During the implementation, every word is analyzed for its suffixes. If it falls under a category of morphophonemic rule, the suffix is stripped and the word is tagged as verb. The input considered for this implementation includes story files in Tamil Language. They have been downloaded and saved in UTF-8 format. The output consists of list of words tagged as Verb. This output is compared with manually computed list of words for the corresponding files. The output shows that half the percent of verb words are identified correctly. Some words were prejudicially identified as verbs because of the transition undergone by them and some verb were left unclassified. As future work, the definition for the grammatical rules has to be enhanced for identifying the verb which is in form of Adverb. In this module, the verbs identified and stemmed to the root form have to be stored in a data file which can be a resource of identifying verbs. This can enhance the identification process in future.

REFERENCES

[1] K. Sarveswaran, G. Dias and M. Butt, "ThamizhiFST: A Morphological Analyser and Generator for Tamil Verbs", *Proceedings of 3rd International Conference on Information Technology Research*, pp. 1-7, 2019.

[2] Sasi Raja Sekhar, Suresh Varma Penumathsa, and Somayajulu G. Sripada. "Verb Morphological Generator for Telugu", *Indian Journal of Science and Technology*, Vol. 10, No. 13, pp. 37-45, 2019.

[3] Lushanthan Sivaneasharajah, Ruvan Weerasinghe and Dulip Herath, "Morphological Analyzer and Generator for Tamil Language", *Proceedings of International Conference on Advances in ICT for Emerging* Regions, pp. 10-13, 2014.

[4] Nimal J Valath and Narsheedha Begum, "Malayalam Noun and Verb Morphological Analyzer: A Simple Approach", *International Journal of Software and Hardware Research in Engineering*, Vol. 2, No. 8, pp. 41-48, 2014.

[5] R. Kumar, K. Sulochana and V. Jayan, "Computational Aspect of Verb Classification in Malayalam", *Proceedings of International Conference on Information Systems for Indian Languages*, pp. 15-22, 2011.

[6] General Grammar of Tamil Words, Available at: http://www.tamilvu.org/courses/diploma/c021/c0212/html/c02121ea.htm

[7] Nannul Available at: https://en.wikipedia.org/wiki/Nann%C5%ABl

[8] Sanford B.Steever, " *The Tamil Auxiliary Verb System*", Routledge Press, 2005.

[9] Thomas Lehman, "*A Grammar of Modern Tamil, Pondicherry, India*", Pondicherry Institute of Linguistics and Culture Publication, 1989.

[10] S. Agesthialingom, "A Note on Tamil Verbs" Anthropological Linguistics", *Anthropological Linguistics*, Vol. 13, No. 4, pp. 121–125, 1971.

[11] L. Lisker, "Tamil Verb Classification", *Journal of the American Oriental Society*, Vol. 71, No. 2, pp. 111-114. 1951.

[12] Charles Theophilus Ewald Rhenius, "*A Grammar of the Tamil Language: with an Appendix*", Church Mission Press, 1836.

[13] S.A. Pillai, "Tamil Nouns", *Anthropological Linguistics*, Vol. 6, No. 1, pp. 7-12, 1964.

[14] M. Mercy Evangeline and K. Shyamala, "Noun Identification for Tamil Language using Morphophonemic Rules", *International Journal of Recent Technology and Engineering*, Vol. 8, No. 4, pp. 1-13, 2019.