International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Malayalam Noun and Verb Morphological Analyzer: A Simple Approach

Nimal J Valath[1], Narsheedha Beegum[2]

[1, 2] Vidya Academy of Science and Technology, Kerala, India

Abstract: ***This paper discusses the methods involved in the development of a Simple Malayalam Verb and Noun Morphological Analyzer. Since in Malayalam, words can be derived from a root word, a purely dictionary based approach for Morphological analysis is not practical. Hence, a 'Rule-cum-Dictionary' based approach is followed along with the Suffix Stripping concept. The grammatical behavior of the language, the formation of words with multiple suffixes and the preparation of the language are dealt with here, with examples of noun and verb forms in detail.***

Keywords: Morphological Analyzer, Malayalam, Suffix stripping, Transliteration, Retransliteration, Verb and Noun, Sandhi rules, Word Formation, Noun Cases, Algorithm**.**

1. Introduction

Morphological Analysis is the process of studying the structure and formation of words. It gives basic insight to the natural language by studying how to distinguish and generate grammatical forms of words. This involves considering a set of tags to describe the grammatical categories of word form concerned. Morphological Analysis is a part of language analysis.

Consider a word, which is a combination of base and suffixes.
Word = stem + affixes
E.g. *kuTTikaL = kuTTi (noun) + kaL (pl)*

A morphological analyzer split the word into its constituent morphemes. In the morphological point-of-view languages behaves in rather a different ways. Malayalam is highly agglutinative and inflecting language that makes it difficult in processing.

The following sections in this paper will briefly discuss about Malayalam language, resources used, methods adopted, experiments performed with Malayalam language as input and output obtained.

2. Malayalam Language

Malayalam is the mother language the state Kerala. It is one among the 22 official languages of India. Total number of speakers of the language is more than 35 million spreading along the regions Kerala, Lakshadweep, and Pondichery.

Malayalam is a language of the Dravidian family and is one of the four major languages of this family with a rich literary tradition. It is very close to Tamil, one of the major languages of the same family. This was due to the extensive cultural synthesis that took place between the speakers of the two languages. The origin of Malayalam as a distinct language may be traced to the last quarter of 9th Century A.D. Throughout its gradual evolution Malayalam has been influenced by the various circumstances prevailed on different periods.

Mainly Malayalam was influenced by Sanskrit and Prakrit brought into Kerala by Brahmins. After the 11th century a unique mixture of the native languages of Kerala and Sanskrit known as Manipravalam served as the medium of literary expression. Malayalam absorbed a lot from Sanskrit, not only in the lexical level, but also in the phonemic, morphemic and grammatical levels of language. There are different spoken forms in Malayalam even though the literary dialect throughout Kerala is almost uniform.

2.1. Word Formation of Noun

Noun is the utterance typically denoting a person, place, thing, animal or idea. It can occur in isolation or can take gender markers, plural markers, case suffixes, postpositions, clitics etc. It takes the form,

W = noun root [Gender] [plural- suffix] [casesuffix] [postpositions] [clitics]+-,

Where W is any word having the properties of a noun.
Gender Suffixes: -an (limited)
Plural Suffixes: -nGaL,-kaL,-ar,-maar
Post positions: -thanne,-pinne,-aano,-pati etc.

2.1.1. Noun Cases

Malayalam follows the system of marking grammatical relations and semantic roles through a set of case suffixes, a feature common to the Dravidian languages. As roles and relations are conveyed through suffixes, word order changes do not normally alter sentence meaning in Malayalam. The case system of Malayalam includes six cases; nominative, accusative, dative sociative, instrumental and locative. The suffixes for each are listed below.

Table 1: Showing Noun cases of Malayalam.

Cases	Morpheme
Nominative	Null
Accusative	-e,-ine
Dative	-kku,-inu
Sociative	-ooT
Instrumental	-aal,-konT
Locative	-il,-kal,-uuTe,iluLLa,-ile

The case suffixes are capable of conveying different shades of meaning over and above the basic grammatical meaning. Nominative, accusative, dative and sociative cases link the nouns to the basic structure of the sentence. When these cases are removed the sentence becomes ungrammatical or semantically defective. Instrumental and locative nouns can be removed from the sentence without affecting the grammaticality of the sentence. Nominative, accusative, dative and sociative can be treated as core cases and the remaining two as peripheral cases.

2.2. Word Formation of Verbs

Verb is the utterance denoting action. It is associated with the action, namely their form, character, time, manner etc.
Verb takes the form,
W = Base [tense][aspect][mood]+_, where W is the word.

2.2.1 Tenses

There are morphologically distinct tenses in the language, and these are labeled as past, present and future. The combination of the three tenses with different aspects and moods are used for a given time specification. Tense is the last feature marked on the verb form and follows causative and aspect suffixes.

Past tense is marked by -i added to the verb root or derived stem, or by -u preceded by one or another of a range of consonants or consonant sequences.

The selection of the appropriate past tense suffix depends on a combination of both morphological and phonological conditioning. Present tense is marked by -unnu suffixed to the verb root or derived stem. The future tense is marked by -uM (occasionally uu) suffixed to the verb root or derived stem. The use of -uu is restricted to sentences in which one element carries the emphatic particle -ee.

Table 2: Examples showing Malayalam verb forms

Base	Past	Present	Future
iLakuka- to move	iLaki	iLakunnu	iLakum
minnuka-to glitter	minni	minnunnu	minnum
vilasuka-to shine	vilasi	vilasunnu	vilasum

Table 3 : showing suffixes associated with the Malayalam verb.

Past	*Future*	*Present*	*Mood*	*Aspect*
-i	-um	-unnu	-aavu	-ka
-ru			-aalum	-uka
-tu			-atte	-ave
-ttu			-aatte	-kil
-ccu			-in	-ukil
-nnu			-aam	-enGil
-ntu			-aNam	-
-nJu			-	-

3. Literature Survey

Morphological Analysis is the first step in any natural language processing systems. Researchers have been done in this area to make an efficient Morphological Analyzer. For example Stanford Morphological Analyzer for English is a complete tool for English language [17]. In India, there are multiple languages spoken throughout the country. For processing of those languages, we need Morphological Analyzers for each language [9]. Morphological Analyzer for certain Indian languages [9], such as Kannada, Hindi etc. are already available. But, for some language like Malayalam, researches are still going on to develop a complete and efficient Morphological Analyzer [11].

Research of Morphological Analyzer can be tracked from the development of Root-Word identifier for Malayalam. Later, Suffix separating approaches came into existence. Suffix were collected and compared with the given Malayalam word to separate them from the root word [6]. Thus the morphemes are identified. Rule Based suffix stripping method was used to strip the correct morphemes from the complete word. Malayalam is an agglutinative language [13, 6]. Therefore, ambiguous nature of Malayalam made Morphological Analysis a difficult process.

Methods involving Probabilistic models were introduced to do Morphological Analysis of Malayalam. Probability of existence of suffix after a root and suffix after another suffix is used to predict the analysis of Malayalam. System is first trained with the pre-calculated values of probabilities. With the help of a pre-tagged Corpus [4]. Finite State Automation is another approach in developing Morphological Analyzer for Malayalam. Finite state Transducers model the morphotactics are used to split the words into constituent morphemes [2]. Success of the system depended on well-defined paradigm system. Morphological Analyzer using hybrid approach is proposed with the help of LTTool box, an important module in the Apertium package [1]. The method used recursive suffix stripping to achieve an 83% average accuracy.

4. Problem Definition

If we had an exhaustive lexicon, which listed all the word forms of all the roots, and along with each word form it listed its feature values, then clearly we do not need a morphological analyzer. Given a word, all we need to do is to look it up in the lexicon and retrieve its feature values.

But this method has several problems. It is definitely the wastage of memory space. Every form of the word is listed which contributes to the large number of entries in such a lexicon. Even when two roots follow the same rule, the present system stores the same information redundantly.

Second, it does not show relationship among different roots that have similar word forms. Thus it fails to represent a linguistic generalization. Linguistic generalization is necessary if the system is to have the capability of understanding (even guessing) an unknown word. In the generation process, the linguistic knowledge can be used if the system needs to coin a new word.

Third, some languages like Malayalam have a rich and productive morphology. The number of word forms might well be indefinite in such a case. Clearly, the above method cannot be used for such languages.

Thus, objective is to perform morphological analysis of Malayalam words using a rule based-pattern matching hybrid approach with the help of a pre-tagged language lexicon.

5. Methodology

The aim is to develop a Morphological Analyzer for Malayalam words representing Nouns and Verbs. For designing the Analyzer, a Combined Approach of Paradigm method and Suffix Stripping method, which are the two promising approaches for building an Analyzer, with Linguistic Rules of Malayalam is used. All that needed is a Dictionary of Stems and a list of all possible inflections in Malayalam. Using programming the morphophonemic changes are handled. The Stem and the Morphemes, which play different morphological functions, can be identified.

E.g. of Malayalam word.
അമ്മയ്യോടൊപ്പം
Transliteration in English.
ammayootoppam

The method consists of following:
Transliteration, dictionary lookup, suffix extraction, application of Sandhi rules, root word identification and retrasnliteration

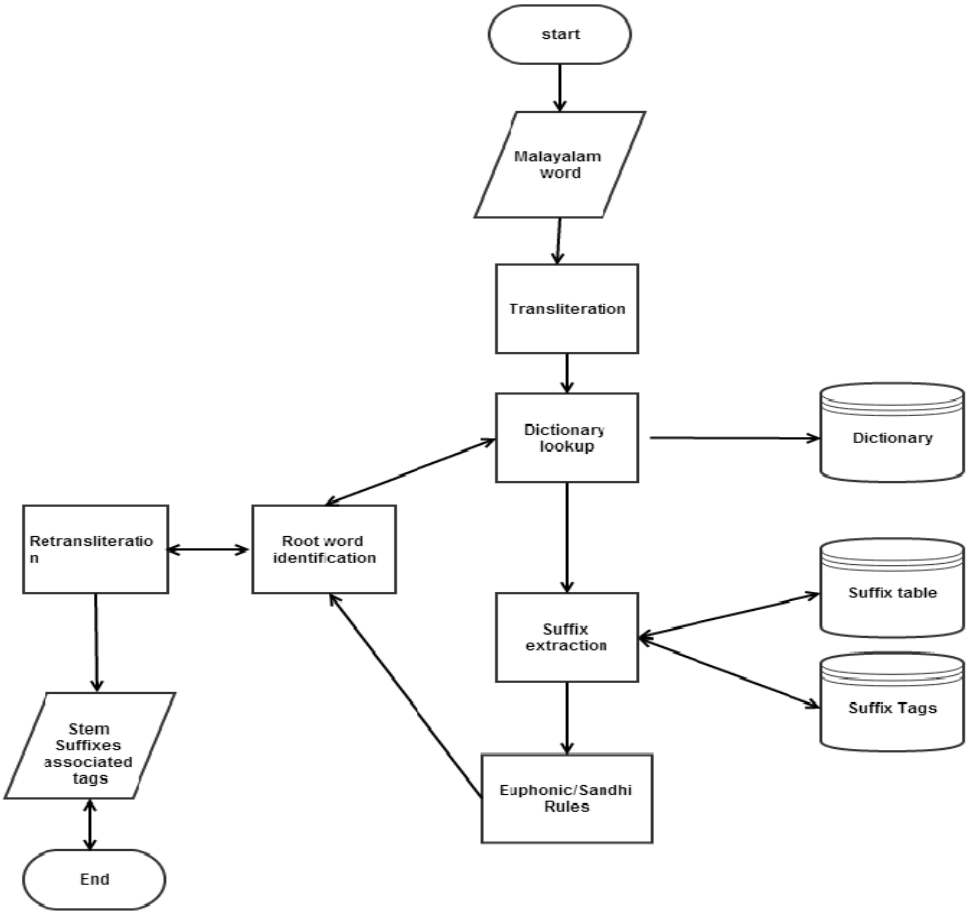

Figure 1: showing the data flow diagram of morphological analyzer for Malayalam word and verb

5.1 Transliteration

Transliteration is the process of conversion of text from one script to another. Transliteration is not concerned with representing the phonemics of the original. It only strives to represent the characters accurately. Transliterator is used to convert Malayalam to English for the ease of processing. The converted Malayalam into English characters is used to find out the occurrence of affixes, here only word with suffixes are taken into consideration.

5.2 Dictionary Lookup

The available input is checked to present in the dictionary of Malayalam words. A pre-release version of Datuk Dictionary corpus is used as the dictionary of Malayalam words. Datuk is a pre tagged corpus of Malayalam words.

5.3 Suffix Extraction

Suffix extraction is the process of extracting suffixes from the given word. In highly agglutinative languages such as Malayalam, a word is formed by adding suffixes to the root or stem. Absolutely no prefixes and circumfixes are there in Malayalam. But morphologically highly complex words exist in such languages, which are formed by continuously adding suffixes to the stem. Suffix Stripping method make use of this property of the language, i.e., having complex suffixes attached to the stem. Once the suffix is identified, the stem of the whole word can be obtained by removing that suffix and applying proper Sandhi/Euphonic rules. In this method we use multiple suffix stripping technique. For stripping the suffixes, first step is the collection of available suffixes present in Malayalam. Secondly, create a dictionary for the Suffixes and finally, create another dictionary to store tags associated with the collected Suffixes.

5.4 Application of Sandhi rules

On word creation, various Sandhi rules are defined in Malayalam for joining two words to form a new one. While applying these rules, the original appearance of the words taking part in this process is altered. Rules are applied by observing the sounds of the end syllable of the first word and the start syllable of the second word. As Malayalam is morphologically rich and agglutinative with complex structures; and there are so many morphophonemic changes in the word formation process, root driven and Brute force methods are not sufficient to generate the words and its forms.

For Morphological analyzer for Malayalam words, Suffix stripping method, which is very close to the affix stripping method in its approach, with Sandhi rules can be used. Since Malayalam requires many morphophonemic changes in the word formation, one has to deal with Sandhi rules in each and every place where two morphemes combine to form a suffix and also where a suffix combines with the stem. Separating the suffixes from its base form is a reverse process of Sandhi where the essence of Sandhi rule is applied in the reverse direction.

5.4.1 Malayalam Sandhi Rules

On the basis of sounds involved, Sandhi can be grouped into Vowel Sandhi, Vowel-Consonant Sandhi, Consonant - Vowel Sandhi and Consonant- Consonant Sandhi. In Malayalam grammar, a classification of Sandhi rules is done based on whether a word ends with a vowel (swaram) or a consonant (vyanjanam).

Figure 2: showing the types of Sandhi in Malayalam

Category	Type of sandhi	Rule	Example
I	Swarasandhi	swaram + swaram	മഴ + ഉണ്ട് = മഴയുണ്ട്
II	Swaravyajana sandhi	swaram + vyanjanam	താമര + കുളം = താമരക്കുളം
III	Vyanjanaswara sandhi	vyanjanam + swaram	തേൻ + ഇല്ല = തേനില്ല
IV	Vyanjana sandhi	vyanjanam + vyanjanam	നെൽ + മണി = നെന്മണി

Sandhi can be classified into following
(a) Elision (Lopa Sandhi)
When followed by any Vowel, unrounded u undergoes elision.
Eg:
taNuppu (chillness)+uNTu (is) = taNuppuNTu (there is chillness)
kaaRRu (wind)+ aTikkunnu (blows) = kaaRRaTikkunnu (wind blows)

(b) Augmentation (Agama Sandhi)
As the vowels have independent pronunciation, it is inconvenient to pronounce two vowels together. So it has to be avoided
E.g.
tii(fire)+ aaTTu (dance) = tiiyaaTTu(fire dance)
kai(hand)+ uNTu (has) = kaiyuNTu(has hand)

(c) Reduplication (Ditva Sandhi)
So the rules of germination are restricted to consonants. In Malayalam, germination is more in tense consonants and less in lax consonants. When two words combine in which the first is the qualifier and the qualified, the tense consonants initial to the second word geminates.
Eg:
pooyi (went)+ paRaunjnju(said) = pooyippaRanjnju(went and said)
manassal (by ill)+koTuttu(gave) = manassaalkkoTuttu(gave by will)

(d)Substitution (Adesha Sandhi)
One letter goes and another letter comes in that position.
Eg:
vil (to sell) + tu(did)= vit+tu=viRRu(sold)
keeL (to hear)+tu(did) = keet+tu=keeTTu(heard)
pin (back) + paaTTu (song) = pilppaaTTu (background song)
pon (gold) + kuTaM (pot) = polkkuTaM (golden pot)

5.5 Root word Identification

After each turn of suffix stripping and Sandhi rules. The modified input is checked with the available dictionary corpus for the stem word. If there is a hit in the table then the Root word is identified, else the whole step through suffix extraction is continued till the last character of transliterated input.

5.6 Retransliteration

If the root word is identified then re-transliteration module will convert the modified input back to Malayalam. Retransliteration module also performs the generation of output consisting of stem and suffixes along with its tags.

5.7 Algorithm

1. Get an input word, w.
2. Check whether the word is in DR. Call function check-dict.
3. If the word is not in DR go to step 4. Else step 11.
4. Find out the suffix of the word by using the Suffix Table. Call function Check-suffix.
5. If a valid suffix is found, strip the suffix and get the remaining (RS) as the stem. Call function strip-root.
6. Check whether RS is a valid stem or not by using DR. Call function checkdict.
   _ If the stem is found in DR, go to step 11.
   _ Else, go to step 7.
7. Check whether the suffix is any tense marker.
   _ If yes, check whether any Sandhi change occurred or not and get the stem after applying sandhi rules. Call function tense-sandhi.
   _ Else, go to step 9.
8. Check whether the new stem is in DR.
   _ if yes stop.
   _ Else stem is not in the Dictionary. go to step 11.
9. Check for other type of morphophonemic change and get a new stem after applying sandhi. Call function sandhi-others.
10. Check for the new stem in DR.
   _ If found go to step 11.
   _ Else the word is not in Dictionary.
11. Exit.

Function check-suffix.
1. Compare the word w, in reverse order, with each of the suffix in ST.
   - If a match is found, keep the suffix and again check for some other match until the word becomes null. Return the suffix.
   - Else, the suffix is not in suffix table.
2. End.

Function strip-root
1. Remove the part of word where the suffix occurs, and get the remaining portion of the word as the stem.
2. Return the stem.
3. End

Function check-dict
1. Compare the stem with each of the entries of the Dictionary.
2. If a match is found, return the stem and its paradigm number.
3. End.

6. Results & Discussion

Entered word is
1. അടിക്കുകയല്ല-aTikkukayilla

Output obtained is
അടിക്കുക-(aTikkuka)-(verb)
അല്ല=(illa)-(NEG)

2. അമ്മമാരോളമുള്ള
ammamaarootuLLa

Output:
അമ്മ-amma-(noun)
ഉള്ള–uLLa-(LOC)
ഓളം–ooLam-(POST)
മാർ-maar-(PL)

3. മകനെയല്ലെങ്കിലെന്ന്-makaneyallenne

Output:
മകൻ-makan-(noun)
എന്ന് –enn-(AVYA)
എങ്കിൽ-enGil-(POST)
അല്ല-alla-(NEG)
എ-e–(ACC)

4. ഇരുന്നിരിക്കുകയാകും
irunnirikkukayaakum

Output:
ഇരിക്കുക-irikkuka-(verb)
ഉം-um-(FUT_TENSE/CON)
ആകുക-aakuka–(VERB)
ഇരിക്കുക-irikkuka-(VERB)
ഉന്നു-unnu-(PRNT_TENSE)

7. Conclusion

Morphological analyzers can be integrated to the language processing systems for a variety of applications in the Natural Language processing sector. They are essential for any type of Natural Language processing works. A full-fledged Morph analyzer is not available in Malayalam. Malayalam is a verb final, relatively free-word order and morphologically rich language. Computationally, each root word of can take a few thousand inflected word-forms, out of which only a few hundred will exist in a typical corpus. Morphological analyzer for Malayalam implemented a hybrid approach of extended Paradigm and Suffix Stripping Method. Thus it is proved to be an efficient method to identify the morphological categories of a given Malayalam verb and Malayalam noun.

8. Future Scope

Currently the method identifies nouns and verbs and their inflections only. For noun, the system can identify the suffixes up to postpositions and some of clitics. For verb, system identifies up to future, present and past tenses. Since there are variety of inflections occur with past tense and base

verb, more work has to be done to include the other suffixes with the verb. The method can be also extended to a fully-fledged Morphological analyzer by adding rules for specific markers.

References

[1] Rajeev, R.R. Research Scholar, Dept. of Linguistics, University of Kerala, Dr. S. Rajendran, Tamil University, Thanjavur State of Art of Morphological parser for Malayalam in the Global Scenario.2005.
[2] J. Hankamer, Finite state morphology and left to right phonology, Proceedings of the Fifth West Coast Conference on Formal Linguistics, Stanford, CA, pp 29-34, 1986.
[3] E. Antworth PC-KIMMO: A two-level processor for morphological Analysis, Dallas, TX: Summer Institute of Linguistics, 1990.
[4] Aswathy.P.V, Morphological Analyzer for Malayalam Nouns, M.Tech Project thesis, Amrita University, Coimbatore. July 2007.
[5] Rajendran, S. Morphological Analyzer for Tamil, Language in India, languageinindia. com.2005.
[6] Rajeev R,R, Rajendran N and Elizabeth Sherly, A Suffix Stripping Based Morph Analyser For Malayalam Language, Science Congress 2007.
[7] Ahmed, S.Bapi Raju, Pammi V.S. Chandrasekhar, M.Krishna Prasad, Application Of Multilayer Perceptron Network For Tagging Parts-Of-Speech Language Engineering Conference, University of Hyderabad, India, Dec. 2002.
[8] P. Anandan, Dr. Ranjani Parthasarathi , Dr. T.V.Geetha, Morphological Generator For Tamil Tamil Inayam, 1-2, Malaysia 2001.
[9] Morphology Based Natural Language Processing tools for Indian languages, www.iitb.ac.in.
[10] Jurafsky, Daniel and Martin, James H Speech and Language Processing- An Introduction to Natural Language processing, Computational Linguistics and Speech Recognition, 2002.
[11] Morphological Analyzer for Malayalam: Probabilistic Method Vs Rule Based Method Rinju O.R, Rajeev R. R, Reghu Raj P.C., Elizabeth Sherly Dept. of CSE, Govt. Engg. College Sreekrishnapuram Palakkad, Kerala, India.
[12] A Paradigm-Based Finite State Morphological Analyzer for Marathi,Mugdha Bapat, Harshada Gune, Pushpak Bhattacharyya, Department of Computer Science and Engineering, Indian Institute of Technology Bombay.
[13] D.C Books-Keralapaanineeyam, Rajaraja Varma.
[14] Morphological Analysis of Malayalam Verbs,Saranya S K, Amitha Enginneering College, Coimbatore.
[15] https://code.google.com/p/kaltura-ce-windows-edition/wiki/WampServer.
[16] http://msdn.microsoft.com/en-us/vstudio/aa496123.
[17] people.umass.edu/moiry/morphology.pdf

Author Profile

Nimal J Valath received the B.Tech from Cochin University in 2011 and Continuing M.Tech degrees in Computer Science and Engineering from Vidya Academy of Science and Technology in 2014, respectively. This paper is a part of research done in Linguistics of Malayalam, native language of Kerala state in India.

Narsheedha Beegum received the B.Tech from Calicut University in 2008 and Continuing M.Tech degrees in Computer Science and Engineering from Vidya Academy of Science and Technology in 2014, respectively. This paper is a part of research done in Linguistics of Malayalam, native language of Kerala state in India.

.