

Rule Based Morphological Analyzer for Malayalam Nouns: Computational Analysis of Malayalam Linguistics

Jancy Joseph, Dr. Babu Anto

School of Information Science and Technology, Mangattuparamba Campus, Kannur University, Kerala, India

ABSTRACT: The morphological analysis deals with the study of internal structure of words of a language based on its grammatical category. The morphological analyzer system is developed for plural markers, case markers, post positions and clitics (gathi) markers for Malayalam nouns. This work focuses on segmenting a morphologically inflected word into its root word and its associated morphological components along with the features specifying the morphological structure. The outputted words in this system are categorized into different classes of noun, which is implemented using Malayalam Unicode Standard. Malayalam Morph Analyzer would help in automatic spelling and grammar checking, natural language understanding, machine translation, speech recognition, speech synthesis, part of speech tagging, and parsing applications. The common man can also get in-depth information about the Malayalam nouns from the software.

KEYWORDS: Morphological Analysis, Malayalam Noun Morphology, Root word, Suffix

I. INTRODUCTION

The field of Natural language processing (NLP) is primarily concerned with getting computers to perform useful and interesting tasks with natural languages. NLP is a field of computer science, integrating artificial intelligence and linguistics, which is concerned with the interactions between computers and human (natural) languages. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation. NLP is the process of computer analysis in which input provided in a natural language is analyzed and converted into an output in a useful form of representation. The field of NLP is secondarily concerned with helping us come to a better understanding of natural language.

II. MORPHOLOGY

Morphology is the study of the structure and formation of words. It deals with the ways that words are built up from smaller meaningful units called morphemes. Morphemes can usefully be divided into two classes- Stems and Affixes.

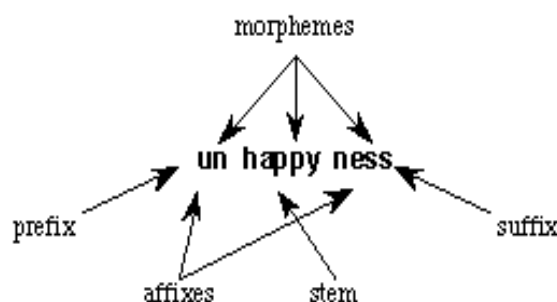


Fig.1 Stem And Affixes



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

III. MORPHOLOGICAL ANALYSIS

This phase separate words into individual morphemes and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the morphology of the language being considered. In Languages like Malayalam, each dictionary entry has thousands of possible word forms, as Malayalam is one of the highly agglutinated Indian Language. So building morphological analyzer for Malayalam is a complex task. Generally, Rule Based Approaches are used for morphological analysis which are based on a set of rules and dictionary that contains root words and morphemes. In Rule Based Approach, a particular word is given as an input to the morphological analyzer and if the corresponding morpheme or root word is missing in the dictionary, then the rule based system may fail. Here, each rule depends on the previous rule. So if one rule fails, it affects the entire set of rules which follows.

IV. MORPHOLOGICAL ANALYZER

A Morphological Analyzer split the word into its constituent morphemes. This Morphological Analyzer for Malayalam words can be further used in a machine translation system. Malayalam is an agglutinative language. Morphological Analyzer will help to identify the inflection of a word and it segment the word into stem and affixes. This affixes can be gender (feminine, masculine or neutral), person (1st, 2nd or 3rd), or number (Singular or Plural) information in the case of nouns, and tense, aspects and modality of the word in the case of verbs. Morphological analyzer for Noun returns the root of the word along with its gender, number and case. For verb it will return the root form along with its tense, modality, and aspects.

Eg: മരങ്ങൾ=മരം(N)+കൾ(PL)

V. CHALLENGES IN BUILDING MORPH ANALYZER

Many changes take place at the boundaries of morphs and words. Identifying the rules that govern these morpho-phonemic changes is a challenge because dissimilar changes take place in similar contexts. In such cases it is necessary to look into the morphological as well as phonological factors which make such changes.

VI. VARIATIONS OF MORPHOLOGY

Morphological Variations for a word occurs due to inflections, derivations, clitization, word compounding etc. Word morphology can further divided into three broad classes.

a. Inflectional Morphology

Inflection is the process of changing the form of a word so that it expresses information such as number, person, case, gender, tense, mood and aspect, but the syntactic category of the word remains unchanged. Inflectional morphology concerns with the combination of stems and affixes where the resulting word has the same word class as the original and it serves a grammatical/semantic purpose that is different from the original, but is nevertheless transparently related to the original.

Eg: കുട്ടി/കുട്ടികൾ

b. Derivational Morphology

Derivation changes the syntactic category of a word. Derivational morphology is includes irregular meaning change and changes of word class. Eg: ഭംഗി/ഭംഗിയുള്ള

c. Clitization

A clitic is an element that behaves like an affix and a word. However, they are quite complicated in that they are also part of word formation.

Eg: Cat/Cat's

VII. MORPHOTACTICS

Morphotactics is the order in which the morphemes are arranged. It is also a kind of restriction on morphemes. The order in which the morphemes appear in a word must be described any computational model of the morphology. It is a fact of any language that one can usually stack up morphemes in some orders but not in others.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

There are many conditions on morpheme ordering that may follow straightly from the kind of constraints like affixation, infixation, pre-fixation and suffixation (Rajeev,2008).

VIII. NEED AND SIGNIFICANCE OF MORPHOLOGICAL ANALYSIS

Morphological Analyzer can either be an important stand-alone component of many applications like spelling correction, information retrieval, machine translation etc. It can simply be a link in a chain of further linguistic analysis. Any Natural Language Processing (NLP) application for any language starts with the development of Morphological Analyzer or Word Analyzer, which analyzes the inflected word and provides information such as root word or stem and its constituent morphemes with which the original word was constructed. Building morph analyzers for highly inflectional languages is difficult but crucial for applications such as Machine Translation (MT) and Dialog Based Natural Language Understanding Systems.

The development of NLP applications is challenging because computers traditionally require humans to communicate to them in a programming language that is precise, unambiguous and highly structured or, perhaps through a limited number of clearly-enunciated voice commands. Human speech, however, is not always precise, it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

A morphological analyzer or generator supplies information concerning morph syntactic properties of the words it analyses or constructs. The design and implementation of morphological analyzer and generator for Malayalam is a promising research for various applications in NLP.

IX. OBJECTIVES OF THE STUDY

The central goal and core of the study is to develop a Morphological Analyzer (MA) for Malayalam Nouns using Suffix-Stripping Method in a Rule-Cum-Dictionary Based Approach by integrating different modules and various computational linguistic tools.

X. REVIEW OF RELATED LITERATURE

A Suffix Stripping based Morph Analyzer for Malayalam Language, by Rajeev R.et.al. (2008) adopted the suffix stripping method as it contains suffixes, which are very close to the affix stripping method in its approach. In Recognizer, a transducer takes a word as input, and accept outputs if the given word is in the language and rejects it if it is not present. Two-level morphology represents a simple splitting of the morphemes from the word so as to get the root/stem. Sumam Mary Idikkula(2007) & team developed a morphological processor for Malayalam language which aimed at building a morphological processor for language, with two main components: a morphological generator and a morphological analyzer. Jisha P. Jayan et.al(2006) developed a Morphological Analyzer and Morphological Generator as part of their thesis on Malayalam - Tamil Machine Translation using a bilingual dictionary. They make use of suffix stripping method for morphological analysis. Jisha. P J et. al ((2009) mentioned about the two common approach towards the morphological analysis, paradigm approach and suffix stripping approach in their study titled Morphological Analyzer for Malayalam- A Comparison of Different Approaches. It also discussed comparison with the hybrid approach. Vinod P M, et.al (2001), makes use of Lttoolbox for morphological analysis, generation, lexical processing etc. The program compares each inflected form. A Morphological analyzer for Malayalam using machine learning was developed by V.P Abeera,S et.al.(2009). Morphological analysis for Malayalam verbs using a hybrid approach (paradigm and suffix stripping method) is attempted for morphological generation of verbs. Saranya S K(2008) developed a Morphological Analyzer for Malayalam Verbs using a hybrid approach of Paradigm method and Suffix Stripping method. Nimal J Valath, et.al(2012) developed a morphological analyzer for nouns and verbs using combined approach of paradigm and suffix stripping method. Initially they transliterated Malayalam to English, which help to find the occurrence of affixes easily.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

XI. MALAYALAM NOUN MORPHOLOGY

A *noun* is a word that functions as the name of some specific thing or set of things, such as living creatures, objects, places, actions, qualities, states of existence, or ideas. Linguistically, a noun is a member of a large, open part of speech whose members can occur as the main word in the subject of a clause, the object of a verb, or the object of a preposition. The nouns in Malayalam are marked for number and gender.

The present work has considered four main categories of inflections for nouns. They are: Gender, Plural (Number) Markers, Case Markers and Clitics(gathi/postpositions).In Malayalam grammar, a classification of sandhi rules is done based on whether a word ends with a vowel (swaram) or a consonant (vyanjanam).സ്വരസന്ധി, സ്വരവ്യഞ്ജനസന്ധി, വ്യഞ്ജനസ്വരസന്ധി and വ്യഞ്ജനസന്ധി.

Sandhi can also be categorize into four on the basis of the changes occurring.They are ലോപസന്ധി, ആഗമസന്ധി,ദിത്വസന്ധി,and ആദേശസന്ധി.

XII. CATEGORIZATION OF NOUNS

Noun is a word that can be used to refer to a person, place, thing, quality, or action.It is the word class that can serve as the subject or object of a verb, the object of a preposition, or in postposition. The noun can be classified as Concrete Noun, Quality Noun, Verbal Noun, and Pronoun. Concrete noun further classified as Proper Noun, Common Noun, Material Noun, and Collective Noun

XIII. METHODOLOGY AND IMPLEMENTATION

Malayalam morph analyzer for nouns is done using suffix stripping method with the reverse application of Sandhi Rules. This rule based system uses a predefined set of dictionary of root words developed for the purpose. When a word input is made, it checks with the dictionary to identify the stem word given and return the noun as output. If it doesn't match with the stem words in the dictionary, it checks for suffixes and strip it off to provide the output as stem word and suffixes. The novelty of this work lies in the medium of input and output that is done in *Malayalam Lipi* without transliteration and retransliteration process.

Since Malayalam is an agglutinative language, a noun can have a number of inflections by adding different suffixes to it. There is no hard and fast morphotactic rule or order in adding suffixes to words in Malayalam. The major challenge associated with suffix stripping method while using *Malayalam Lipi* is the replacement of added *Varnam*(syllables) which is resulted due to the application of *Sandhi Rules*.

A hand built dictionary is used in this work. The word in the dictionary was derived from performing a corpus analysis of few Malayalam books. All the unique words in this corpus including different categories of nouns were included in the dictionary.

The Malayalam Unicode utf-8 standard is used in this work. Thus it is possible to store Malayalam noun corpus in its own Lipi and could use Malayalam Lipi within the Coding too. This works deals seven different classes of noun (such as Qualitative, Verbal, Proper, Common, Material, Collective and Pronoun) in addition to the Demonstrative and Interrogative Pronoun.

XIV. ALGORITHMS

1. Input the word to be analyzed.
2. Check whether the given word is found in the Root Dictionary.
3. If the word is found in the dictionary, then go to step 8; else
4. Separate any suffix from the right hand side
5. If any suffix is present in the word, then remove the suffix and then re-initialize the word without the identified suffix and go to Step 2.
6. Classify the suffix to its correct class.
7. Repeat this process until the Dictionary finds the root/stem word.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

8. Find the resulting word belongs to which noun category.
9. Output the stem word along with its associating suffixes and stop.
10. Exit.

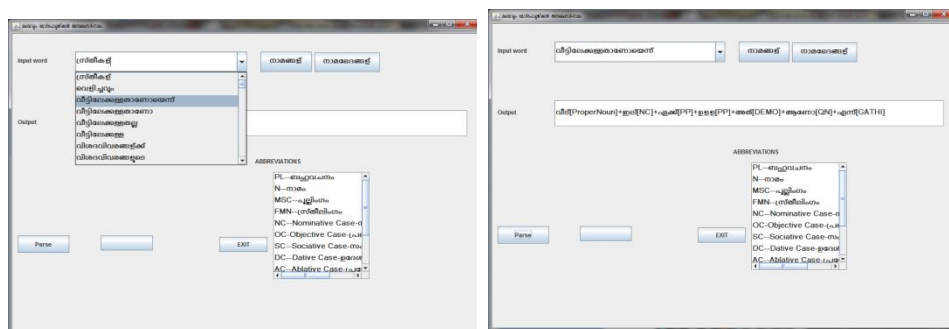
Output:

അമ്മയോടൊപ്പമുള്ള=അമ്മ[CommonNoun]+ഓട്[SC]+ഒപ്പം[SPL_ഗതി]+ഉള്ള[PP]
 വീട്ടിലേക്കുള്ളതാണോയെന്ന്=വീട്[ProperNoun]+ഇൽ[NC]+എന്ന്[PP]+ഉള്ള[PP]+
 അത്[DEMO]+ആണോ[QN]+എന്ന്[GATHI]

XV. RESULTS

The system provides the feature components of inflected nouns along with its morphological category. A testing corpus of 500 nouns has been taken. The testing corpus of morph analyzer will be the inflected words. The nouns include all types of Malayalam nouns and its suffixes. The inflections considered are 40 for nouns. Therefore the testing corpus of morph analyzer consists of 500 nouns that are inflected into 2000 inflections. The system ensures 89.5 % accuracy for tested data. The morphological analyzer system is developed for plural markers, case markers, post positions and clitics (gathi) markers for Malayalam nouns. The system accuracy could be improved by increasing the number of root words in the database, increasing the number of rules and by considering more inflections of Malayalam morphology.

This work focuses on segmenting a morphologically inflected word into its root word and its associated morphological components along with the features specifying the morphological structure. The outputted words in this system are categorized into different classes of noun. The dictionary based approach used here results in a more accurate morphological system.



XVI. CONCLUSION

It is essential to build a morphological analyzer for natural language for any further Language processing works. Malayalam is a morphologically rich language and each root word of noun can take hundreds of inflected word-forms. The corpus of inflected nouns and stem words are defined for better performance of the system. *Rule based Morphological analyzer for Malayalam Nouns: Computational Analysis of Malayalam Linguistics* is implemented with Suffix Stripping Method and application of Sandhi rules of Malayalam in reverse. It proved to be an efficient method to identify the morphological categories of a given noun.

REFERENCES

1. Abeera V. P., S. Aparna, R. U. Rekha, M. Anand Kumar, V. Dhanalakshmi, K. P. Soman, S. Rajendran, Morphological analyzer for Malayalam using Machine Learning, *ICDEM10 Proceedings of the Second international conference on Data Engineering and Management*, Pages 252-254.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

2. Antony P J, Dr K P Soman(2012),Computational Morphology and Natural Language Parsing for Indian Languages: A Literature Survey,*International Journal of Computer Science & Engineering Technology (IJCSSET)*,3(4) April 2012
3. Aswani Shaji ,Sindhu L(2014) ,Morphological Analyzer for Malayalam: A Literature Survey,*International Journal of Computer Applications (0975 – 8887) Volume 107 – No 14, December 2014 24*
4. Jisha P. Jayan, Rajeev R R, S Rajendran(2009), Morphological Analyser for Malayalam- A Comparison of Different Approaches, *IJCSIT, Vol. 2(2)*, pp 155-160,December 2009.
5. Jisha P. Jayan, Rajeev R R, S Rajendran(2006), Morphological Analyzer and Morphological Generator for Malayalam - Tamil Machine Translation, *IJCSIT, Vol. 2(2)*, December 2006.
6. Jurafsky Daniel and Martin James H (2005), “An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”.*Prentice Hall, ISBN: 0130950696*, contributing writers: Andrew Kehler, Keith Vander Linden, and Nigel Ward.
7. Rinju O.R., Rajeev R. R., Reghu Raj P.C. , Elizabeth Sherly (2013).Morphological Analyzer for Malayalam: Probabilistic Method Vs Rule Based Method, *International Journal of Computational Linguistics and Natural Language Processing Vol 2 (10) October 2013 ISSN 2279 – 0756*
8. Raja Raja Varma. A.R, (2000) “Keralapanineeyam”, *D.C Books, Kottayam-12. India.*
9. Rajeev R. R., Rajendran N., Elizabeth Sherly(2008), A Suffix Stripping based Morph Analyser for Malayalam Language, *Science Congress.*
10. RaviSankar S Nair ,(2012). A Grammar of Malayalam, Language in India, *ISSN 1930-2940, 12:11 November 2012*. Retrieved from <http://www.languageinindia.com>
11. Saranya S.K, (2008) “Morphological analyser for Malayalam verbs”, *M.Tech Thesis*, Amrita School of Engineering, Coimbatore.
12. Sumam Mary Idicula, Peter S David (2007).A Morphological processor for Malayalam Language, *South Asia Research Journal Vol1 27(2), h*, SAGE Publications U.K.