

Project Documentation

Morphological Analyzer for Malayalam

Jincy Baby

International Centre for Free and Open Source Software (ICFOSS)

Project Overview

Morphological analyser is a program which compiles and analyses words belonging to a natural language. The pre-processing stage before the analysis, which splits the input word into morphemes, is the major part in an analyser. Morpheme generator is utilized for this part.

Introduction

Morphological analysis for Indian Languages is defined as the analysis of a word in terms of its lemma, gender, number, person, case, vibhakti, tense, aspect and modality. A major way in which morphologists investigate words, their internal structure, and how they are formed is through the identification and study of morphemes, often defined as the smallest linguistic pieces with a grammatical function.

System Description

Morpheme generator is utilized as a tool for generating the root and corresponding suffixes adjoined. An addition module is included for tagging the stem and suffixes after splitting the input into its appropriate stem and suffix list. The module make use of a tag database included specifically for tagging the suffixes. The stem is tagged into noun and verb based on the suffix list obtained. The additional functions introduced are as follows

NV



The function assigns appropriate tags to the root word taking into account the list of suffixes obtained from the finallist. The function considers two dictionaries verbsuffix which contains suffixes which can be attached with a verb and a noun dictionary noundict which contains suffixes that can be attached with a noun.

Algorithm(*NV*)

Input: root, clist=list of suffixes obtained from the finallist

Output: [morpheme,tag]

Steps:

1. if root in verb dictionary, return [morpheme, Verb]
2. if clist==[] or clist[0] in verbsuffix dictionary do
3. replace “” by “” and if the obtained word in dictionary, return [word,Verb]
4. if clist==[] or clist[0] in noundict do
5. return [word,Noun]

tag

tag assigns appropriate tags to the input suffix. tags for the suffixes are obtained from the database introduced for tags. Each suffix is searched in the database and upon finding return the list with the suffix and tag.

Morphological analyser

The main body of the morph analyser uses the root function to derive the list of stem and suffixes to be analysed further. The obtained stem is first searched to identify it as Noun or Verb utilizing the function *NV*. If the stem is verb, then the suffix list is checked in a database if the combination of suffixes are present. If not the suffixes are tagged one by one using the tag function. Algorithm(*Morph Analyser*)


Input: Word

Output: Analysis result

Steps:

1. Import *morph_gen*
2. Derive list of morphemes using *morph*
3. Use *NV* to determine if the stem word is Noun or Verb
4. If verb do
5. if the list of suffixes is in the database, derive the tag of suffixes and give the output.
6. break
7. else, use *tag* function and return the result

Results



The screenshot shows the web interface of the 'Morphological Analyser for Malayalam'. At the top, the title 'Morphological Analyser for Malayalam' is displayed in a large, bold, brown font. Below the title, there is a prompt 'Enter the malayalam word to split:' followed by a text input field containing the placeholder 'Enter the word'. A blue 'SUBMIT' button is positioned below the input field. The output area displays the analysis of the word 'കൊച്ചിയിലെത്തിയ' (Kochiyilathathiya). The analysis is shown in three lines: 'കൊച്ചി ഇൽ എത്തിയ' (Kochi ilathathiya), 'കൊച്ചി Noun' (Kochi Noun), and 'ഇൽ LOC എത്തുക Verb ഇയ RELATIVE' (ilathathiya LOC ilathathiya Verb ilathathiya RELATIVE). The ICFOSS logo is visible in the bottom right corner of the interface.



The screenshot shows a web application titled "Morphological Analyser for Malayalam". It features a text input field with the placeholder "Enter the word" and a blue "SUBMIT" button. Below the input field, the Malayalam word "പറഞ്ഞുകൊണ്ടിരിക്കുക" is displayed. The analysis results are shown as a list of tokens: "പറയുക", "ഞ്ഞു", "കൊണ്ട്", "ഇരിക്കുക", and "ASP_ITER". The word "പറയുക" is identified as a "Verb". The ICFOSS logo is visible in the bottom right corner of the interface.

Morphological Analyser for Malayalam

Enter the malayalam word to split:

Enter the word

SUBMIT

പറഞ്ഞുകൊണ്ടിരിക്കുക

പറയുക ന്നു കൊണ്ട് ഇരിക്കുക

പറയുക Verb

ഞ്ഞു കൊണ്ട് ഇരിക്കുക ASP_ITER

ICFOSS

The tests produced an average 90% accuracy for the analyser. Tests were carried out with testing corpus from various sources. The updation of dictionary and tag list seems to increase the overall accuracy rate of the system.