*International Journal of Latest Trends in Engineering and Technology*
*Special Issue SACAIM 2016, pp. 1-6*
*e-ISSN:2278-621X*

# 'BHASHAMAKUDAM' KNOWLEDGE DICTIONARY FOR MALAYALAM WITH KNOWNET

Raji Sukumar[1] A, Dr. Babu Anto P[2]

Abstract- Bhashamakudam is a lexical database for the Malayalam language. The number of words in a language is much greater than those in good dictionaries. Bhashamakudam groups Malayalam words into sets of synonyms called synsets, which records the various semantic relations between these synonym sets. The purpose is to fold to produce a combination of dictionary and thesaurus that is more naturally usable to support automatic text analysis and AI applications. Discovering word meaning from their usage involves grouping the usages based on a concept, so that those in the same group are semantically close to each other under Ontology. Whereas those in different groups are distant, each such group represents a sense of the target. Identifying idioms and specific idiomatic usages of multiword expressions involves determining whether a usage (or a set of usages) of the expression is semantically distant from the usages of its components. As measures of semantic distance between concepts can be extended to calculate the distance between larger units of language, such as phrases and documents, understanding and improving these measures will have a significant and wide-ranging impact. This works designed a KnowNet for Malayalam which creates all inflected form of a word based on Keralapanineeyamgrammer rules and creates a knowledge dictionary.

Keywords –AI, KB, NLP.

## I. INTRODUCTION

Pieces of text that can be used more or less interchangeably and can be identified by their property being semantically close. Knowledge rich measures of concept-distance rely on the structure of a Knowledge Basis, such as WordNet is used to determine the distance between two concepts. Distributional measures of word-distance rely simply on text and can give the distance between any two words that occur at least a few times. WordNet-based measures can capitalize on the manual encoding of lexical semantic relations, while distributional approaches are widely applicable because they need only raw text. This work focuses mainly on the knowledge required to process NL in its semantic level, which bridges the gap between linguistic and common sense knowledge. Knowledge graphs, as a kind of representation which points out a new way for NL describing and modeling, and semantic understanding. With linguistic knowledge it is necessary to have semantic knowledge for the understanding the system implemented with Ontology. The Malayalam vocabulary binds with knowledge graphs, with a network of knowledge. It is represented in both verbs and noun's knowledge graphs. A knowledge dictionary 'Bhashamakudam' is designed as a lexical database for the Malayalam language. It groups Malayalam words into its linguistic and semantic level knowledge. Synonym sets provide synsets, general definitions, records the various semantic families between these synonym sets. 'Bhashamakudam' organizes the lexical information in terms of word meanings which can be labeled as a lexicon based on psycholinguistic principles. The design of the 'Bhashamakudam' includes a ML-KnowNet is inspired by the famous English WordNet, thesaurus, SynSet. The purpose is to fold a combination of dictionary and thesaurus that is more intuitively usable and to maintain automatic text analysis in semantic level AI applications.

[1] *Department of Information technology Kannur University, Kannur, Kerala, India*
[2] *Department of Information technology Kannur University, Kannur, Kerala, India*

## II. BACKGROUND STUDY

There are many works carried out in the field of Malayalam Morphological Analyzer [1], but no complete system is available for common people. [2] reports some important works related to Malayalam MA. Two common approaches identified towards the development of MA are suffix stripping and paradigm approaches[3] mentioned about a hybrid approach for developing the Malayalam MA and its comparison with the above mentioned approaches [5].The main part in the development of a MA is by creating the morphological dictionary [4], is needed in creating the dictionary. The paradigm facility helps a lot to handle the inflections of the words. In order to cover all cases we created noun paradigms, verb paradigms [5], adjective paradigms [5]. Intelligent systems are KB which is developed as a result of knowledge extraction in the field of AI [6]. Text manipulation for knowledge extraction is possible by producing text in a desired format, has been recognized as an important area of research in NLP [7]. Knowledge Acquisition (KA) is a broad field that encompasses the processes of extracting, creating, and structuring knowledge from experts and heterogeneous resources [8]. A NL text processing system may begin with morphological analyses [9]. A project reported by, viz Aristotle, which aims to build an automatic medical data system that is capable of producing a semantic representation of the text in a canonical form [10].

## III.GRAMMATICAL KNOWLEDGE

ammar is a body of rules imposed on a given language for speaking and writing. There is no limit to the number of complex words or sentences in a language. Digital Infinity is achieved by rearranging discrete elements in particular orders and combinations, how are words connected to form sentences not by varying some signal along a continuum like the mercury in a thermometer. Each level of knowledge and correspondingly, each section contribute to subsequent levels. The Structural requirement of grammatical structures, in particular, adheres with three main principles which dictate how they can be built and joined. The following grammatical properties are in the dimension shown in the given table.

Table 1. : Various grammatical properties

| | |
|---|---|
| Aspect | Conveys time-related details like duration, completeness, habitualness, continuousness, progressiveness, etc.; e.g., Ram drove to the work versus Ram is driving to work. |
| Case | Conveys syntactic cues like subject, direct and indirect object, possession, instrument, etc.; e.g., John SUBJECT gave Mary INDIRECT-OBJECT [the book]DIRECT-OBJECT. |
| Gender | Conveys either the natural gender for most living things or often an arbitrary gender for nonliving things. See Noun Gender on page 15 for more information. |
| Mood | Conveys factuality, possibility, uncertainty, likelihood, etc.; e.g., If today were Saturday, we would not be working |
| Number | Conveys the quantity of a noun or nouns. |
| Person | Conveys the speaker, addressee, third party, etc.; e.g., I am running for President, as spoken by the candidate. |
| Tense | Conveys the time when something happened or when it was reported. |
| Voice | Conveys what acted on what; e.g., Ram stole the car versus the car was stolen by Ram versus the car was stolen. |

The hierarchical nature of language builds its larger structures from smaller ones. Phrases constitute the major building blocks in this process by systematically combining isolated words into semi-meaningful fragments. Phrases structures are the result of applying phrase rules to individual words to build the major components of a sentence. When phrases are joined properly, they form larger grammatical units called clauses, which, in turn, can be combined in various ways to build sentences. The different phrase structure and class are given below. This issue is better addressed within the domain of psycholinguistics

and philosophy, as it relates closely with the notions of thought and intent. Only the grammatical aspects of sentences are considered here.

## IV. SEMANTIC RELATIONS

Most people are familiar with a variety of word relations, as they are commonly found in language resources like dictionaries or thesauruses [Jannedy&Poletto, 1994].

*Synonymy*: The set of words that are relatively equivalent in meaning; e.g., couch and sofa, start and begin, in Malayalam '*peru, nama'm*are synonyms of Name. Synonyms can usually substitute for each other with slight different connotation; one may be a more appropriate choice.

*Homonymy*: The set of words that are relatively equivalent in form in one of two says; e.g., homophones like night/knight or flour/flower, in Malayalam '*KuzhaKuNu/KuzhyKuNu'*. Such words lead to lexical ambiguity, which complicates the semantics of a sentence by providing multiple interpretations.

*Antonym*: The set of words that are relatively opposite in meaning in different ways:

Gradable antonyms are an open group of words with a range of scalar opposites; e.g., hot and cold versus hot and cool, big and small versus big and tiny, In Malayalam a*kale*X *aduThu*, *doore* X *aduThu*. The degree can be qualified; e.g., very hot, marginally cold.

*Hyponymy*: is also called as entailment is the set of words that are related taxonomically in meaning; X is a kind.

## V. KNOWNET (BHASHAMAKUDAM)

The semantic of language chunk has many different dimensions, this may obtained from the linguistic property of the words or its knowledge variations based on its context. It is a tedious task to bind the vocabulary of a language under a dictionary. In this semantic model, the investigator is proposing an ER model of all possible words in Malayalam language. This model shows how the various categories of lexicons incorporate with knowledge with in the linguistic features and develops the knowledge dictionary *Bhashamakudam*. This knowledge dictionary will be able to give both linguistic and semantic meaning of a word. This is an ontology that will provide all information like a WordNet and detailed knowledge information about the meaning distribution. These synsets are linked with each other by means of lexical and semantic relations. These relations include synonymy, hypernymy/hyponymy, meronymy/holonymy, antonymy, etc.*Bhashamakudam* distinguishes between adjectives, nouns, adverbs and verbs, because they follow different grammatical rules it also include prepositions, determiners, quotable thoughts, proverb, riddles, commonly used multi words etc.

Bhashamakudam is a lexical database for the Malayalam language. It groups Malayalam words into sets of synonyms called synsets to provide records the various semantic relations between these synonym sets. The purpose is to fold: to produce a combination of dictionary and thesaurus that is more naturally usable to support automatic text analysis and AI applications. Both nouns and verbs are organized into hierarchies, separate by hypernym or IS A relations. *Bhashamakudam* also provides the polysemy count of a word: the number of synsets that contain the word. If a word participates in several synsets (i.e. has several senses) then typically some senses are much more common than others. Bhashamakudam quantifies this by the frequency score: in which several sample texts have all words semantically tagged with the corresponding synset, and then a count provided indicating how often a word appears in a specific sense. The morphology functions of the software distributed with the database try to deduce the lemma or root form of a word from the user's input; only the root form is stored in the database unless it has irregular inflected forms. The complete relational model designed for the Bashamakudam is given in figure 1.

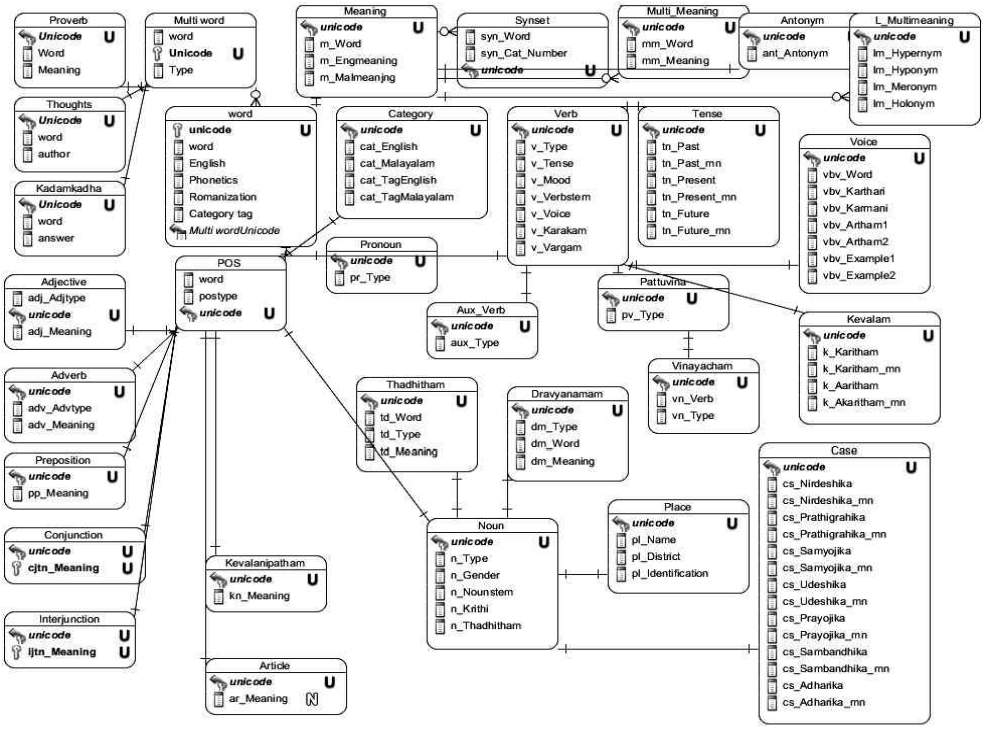


**Figure 1: Class diagram of the *Bhashamakudam* a relational model for all inflected words in Malayalam words**

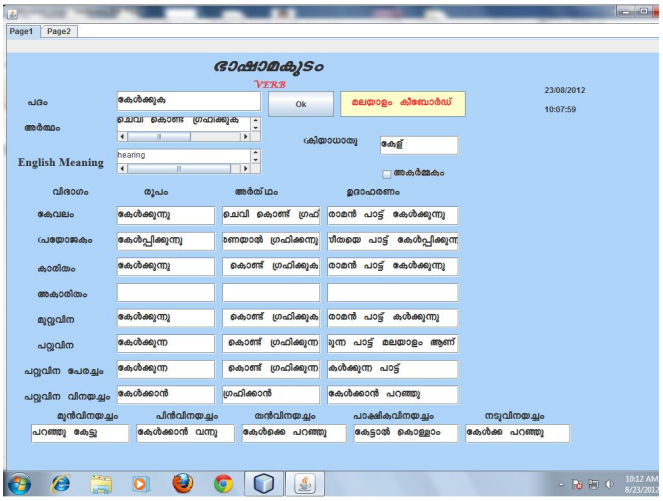


**Figure 2 Morphological variations of Malayalam verb**

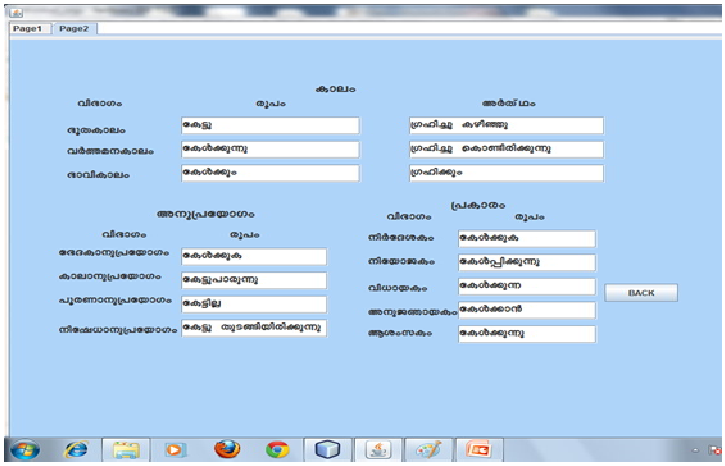


**Figure 3 Morphological variations of Malayalam verb**

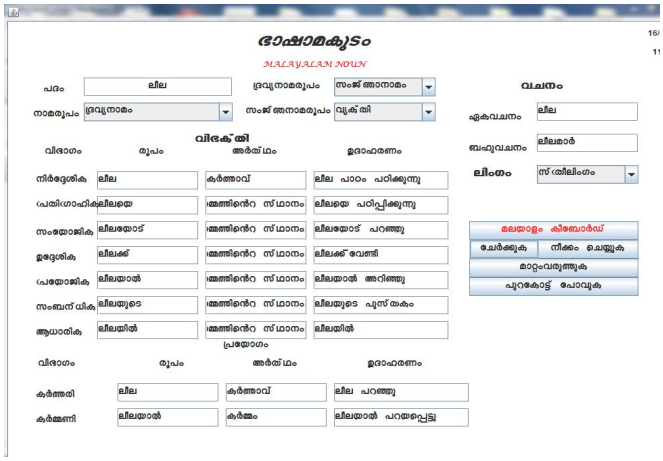


**Figure 4 Morphological variations of Malayalam noun**

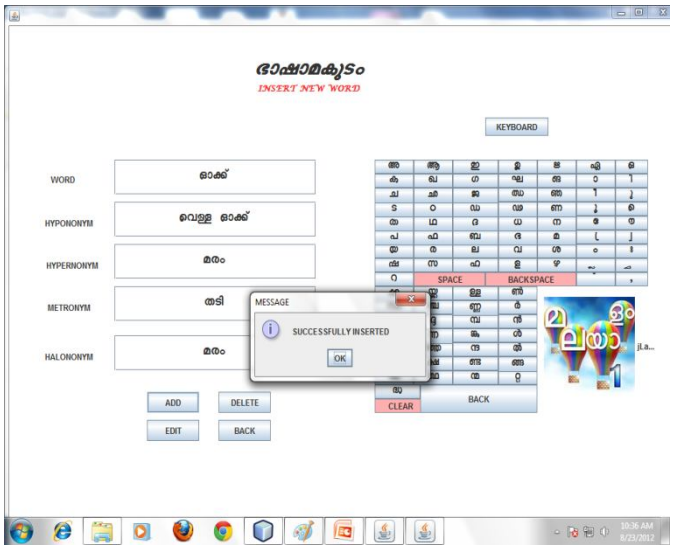


**Figure 5 Knowledge variations of Malayalam word**

## VI. CONCLUSION

The knowledge dictionary include following tasks: Unicode, Romanization, Phonetics Generation for each Word, Stemming. Inflected form generation of all types of words. Data integrity can be maintained by enforcing a standard for the purpose of knowledge extraction. This system allow a user to add root form of a verb or a noun and in the case of a verb 25 inflected forms will be generated automatically and the user can add meaning or related information in this system. If it is a noun 9 inflected forms will be generated automatically corresponding details can be added. This KnowNet contains 21172 word forms. And also have facility to add quotes or riddles. This is an innovative and efficient system for knowledge extraction compared to other WordNets.

## REFERENCES

[1] Jisha P Jayan, Rajeev R R, S Rajendran. "Morphological Analyser and Mophological Generator for Malayalam-Tamil machine translation." Morphological Analyser and Mophological Generator for Malayalam-Tamil machine translation, Published by Foundation of Computer Science (2011): 15–18.

[2] Rajeev R R, Elizabeth Sherly. "Morph Analyser for Malayalam Language: A suffix stripping approach." Proceedings of 20th Kerala Science Congress, Thiruvananthapuram (2008).

[3] Jisha P Jayan, Rajeev R R, S.Rajendran. "Morphological Analyzer for Malayalam-A comparison of Different Approaches." IJCSIT (2009): 155-160.

[4] Suranad Kunjan Pillai. "Malayalam Lexicon." The University of Kerala (2000).

[5] Sunil R, Nimtha Manohar, V Jayan, K G Sulochana. "Morphological Analysis and Synthesis of Verbs in Malayalam." ICTAM (2012).

[6] Nirenburg S, Carbonell, J, Tomita, M, Goodman, K. "Machine Translation: A Knowledge-Based Approach, Morgan Kaufman." San Mateo, CA (1992).

[7] Mani I, E Bloedorn. "Machine Learning of Generic and User-Focused Summariza¬tion." In Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI/IAAI, pages 821-826. AAAI Press / The MIT Press, 1998. (1998): 821-826.

[8] Payne, Thomas Edward. "Describing morphosyntax: a guide for field linguists." Cambridge University Press (1997): 238–241.

[9] Smeaton A F. "Using NLP or NLP Resources for Information Retrieval Tasks. In: T. Strzalkowski ." Natural Language Information Retrieval, Kluwer Academic Publishers (1999).

[10] Roux M, Ledoray V. "Understanding of medico-technical reports." . Artificial Intelligence in Medicine (2000): 149-72 .