DESIGN AND DEVELOPMENT OF A CORPUS BASED LEXICON IN MALAYALAM

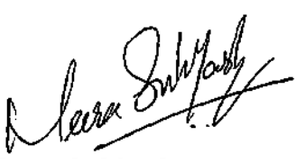
Thesis submitted to the University of Kerala
for the award of the
Degree of Doctor of Philosophy in
Computational Linguistics

MEERA SUBHASH

Department of Computer Science,
Department of Linguistics
University of Kerala
Kariavattom
Thiruvananthapuram
Kerala India
2012

DECLARATION

I hereby declare that the Ph.D thesis entitled **"DESIGN AND DEVELOPMENT OF A CORPUS BASED LEXICON IN MALAYALAM"** *is an independent work carried out by me and it has not been submitted any where else for any other degree, diploma or title.*



27/7/2012 **Meera Subhash**

UNIVERSITY OF KERALA

CERTIFICATE

This is to certify that the work embodied in the thesis entitled **"DESIGN AND DEVELOPMENT OF A CORPUS BASED LEXICON IN MALAYALAM"** *has been carried out by* **Mrs. MEERA SUBHASH** *under our supervision and guidance.*

Dr. M. Wilscy
(Guide)
Professor and Head
Department of Computer Science
University of Kerala
Thiruvananthapuram

Dr. S.A Shanavas
(Co-guide)
Assistant Professor
Department of Linguistics
University of Kerala
Thiruvananthapuram

ACKNOWLEDGEMENT

*By the grace of **God**, I have finished this work.*

*I am very much grateful to my guide **Dr. M. Wilscy.** She accepted my best efforts and always encouraged and supported me throughout the work. Her constant supervision and persuasion inspired me to complete my work in time. I express my sincere gratitude to her for the continuous support for my research, her patience, motivation, enthusiasm and helping mentality.*

*I am deeply indebted to **Dr. S. A. Shanavas** for his valuable support and active involvement throughout the course of this investigation. His enthusiasm, imagination, deep intuition and quest for motivating me for a higher result encouraged me to work hard.*

*I take this opportunity to express my sincere gratitude to **Dr. Shoba L**, Member Research Staff, AUKBC Research Centre, M.I.T campus of Anna University, Chennai, for her invaluable help on tracking my work when I was immersed in the chaos of designing a lexicon. She has enlightened me through her wide knowledge of Artificial Intelligence and her deep intuitions about the work. She sincerely helped me as her own research student by constantly stimulating me to learn more and grow.*

*I am indebted to Late **Dr. A.P Andrewskutty**, former Professor and Head of the Department of Linguistics who paved me the way for doing research in this interdisciplinary field. But I missed him to share this happiness. Thanks to the great teachers, **Dr. N. Rajendran, Dr. S. Kunjamma, Dr. Rosemary A, Dr. A. Marykutty** and **Dr. Aji. S** who supported me whole heartily. Supports from staffs of office and lab technicians from both the departments are appreciable.*

A special recognition needs to be given to the facilities provided by the Librarians in the Department of Computer Science, Department of Linguistics, Campus Library, Kariavatom, University Library and the Library in International School of Dravidian Linguistics, Thiruvananthapuram, without which the compilation of this thesis would have been difficult.

I thank the members in Centre for Development of Advanced Computing (CDAC), *State Institute of Languages and Centre for Development of Image Technology* (CDIT), *Thiruvananthapuram, who invited me to become a member in various discussion groups*

related with my field which enhanced my confidence. I also thank the participants who heard my presentations in seminars and conferences for giving comments and suggestions for improving my abilities. I extent my gratitude to social networking group, Swathantra Malayalam Computing for updating my knowledge in this field. Special thanks to the authors mentioned in the reference who inspired me a lot.

The encouragement and support given by my parents, in-laws, husband and child are invaluable. Most especially to my mother, words alone cannot express what I owe her for her encouragement and support to complete this work. I would like to thank my child who co-operated with me when I was busy with my work. Thanks to my father and my in-laws, who encouraged me to complete the work within the stipulated period.

It was a luck and pleasure to share doctoral studies with good friends and colleagues in a green house like environment in the campus of University of Kerala, Kariavattom.

27/7/2012 **Meera Subhash**

CONTENTS

ABSTRACT

Words are tools of life which is omnipresent in every language. They have their own unique functions and meanings. The syntactic and semantic knowledge about individual words can be encapsulated in a highly structured repository known as computational lexicon which is very essential for many applications including Machine Translation. Many Natural Language Processing tools require computational lexicon for identifying the features of words in a language. In this work, a computational lexicon for Malayalam Language is developed by applying corpus based approach.

The software system developed identifies the lexical items in Malayalam documents with their linguistic details and provides information such as syllabic structure, etymology, root form of the word, grammatical category of root word, inflected forms of the root word, the suffixes agglutinated with the word, name of each suffix, words obtained after removing each suffix and their grammatical category.

For designing a computational lexicon, the first and foremost task is to identify the head words or root words in the language with its grammatical properties from the corpus. A web based corpus for Malayalam is automatically created using a web crawler and it is used for developing computational grammar. Root Word Identifier is implemented using rule based approach which will automatically remove the inflected parts of a word and derive its root form using morphophonemic rules of Malayalam grammar.

Another important information provided by a computational lexicon is the etymology of words. In this work, foreign words written in Malayalam orthography are identified as English words assuming that they are of English origin. Sanskrit words are treated as Malayalam words as most of the Malayalam words are of Sanskrit origin. Statistical language modelling using n-gram is used to decide whether a word is of native origin or foreign origin.

The work is implemented using Practical Extraction and Report Language (PERL) in Linux environment using the 10.04 LTS version of Ubandu Operating System. Unicode supportive font *Meera* is used for displaying Malayalam scripts. Statistical evaluation metrics like Precision, Recall and F-measure are used to evaluate the performance of the Root Word Identifier and foreign word identification. These measures obtained are above 90% for Root Word Identification system and 70% for Foreign Word Identification.

The corpus based, rule based and statistical approaches are incorporated in this work. Since a large amount of automatically retrieved documents are used, the generated lexicon is far better than that generated using other methods. This automatic system can incrementally update the lexicon using richer corpora in future. The system can be successfully integrated with high entity tasks like Machine Translation, coining technical terminology, developing bilingual dictionary, lexical database, speech synthesis etc.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RWI	Root Word Identifier
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
OCR	Optical Character Reader
S	Singular
P	Plural
H	Honorofic
I	Inclusive
E	Exclusive
M	Masculine
F	Feminine
N	Neuter
w	Web address

Chapter I

INTRODUCTION

With the vast development of Information Communication and Technology (ICT), world has become a global village and one of the major challenges faced by the global society is the communication gap, termed as Language Barrier. This is caused by the usage of different languages with different dialects and styles. This impairs the progress in every field. For example, researchers are now referring only books/web sites/journals available mainly in English or that in their first or second language. Some of the linguistic factors like lack of knowledge in script, vocabulary, interpretation of facts and meaning, which depend upon the regional culture are holding them back from referring materials from alien languages. This results in information accessibility gap between students and their subject. To resolve such linguistic barriers, Machine Translation is a solution. Machine Translation is a means used to translate the source text to the required language so that any document written in any language can be utilised without a human assistance.

Today, the language engineering is highly equipped to develop a full fledged Machine Translation system with the advent of free open source operating/software systems like Linux, PERL, PHP etc. and with translation support system like WordNet and with the language tools like analysers, generators, computational lexicon etc. using rule based approaches and or statistical approaches. A lexicon [Zgusta 1971] plays an important role in translation since it is the place where all the information about the vocabulary of a language is recorded for proper usage. Detailed information about lexical items such as grammatical category, its usage etc. are required for automatic translation. For the above purpose, a computational lexicon is the most suitable tool for understanding and retrieving the features of a language. Computational lexicon, the backbone for language computing, is a highly structured repository of the rich syntactic and semantic knowledge about individual words in Natural Language Processing system [w1]. The definition postulated by Makkai in 1980 for Associate Lexicon [Makkai 1980] can be used as a suitable definition for a computational lexicon. It states that it is an information retrieval system that represents in visual and audible form, the knowledge native speakers' posses about the Lexis of their language. It is capable of associatively group the lexemes to form natural semantic nets around concretely observable and non-observable entries. Multidimensional information

accessibility helps to extract the frequency of usage, typologically based frequency, exact range of dialectical habitat, the speakers' sociological status, etc. It offers access of words not only alphabetically but also according to various linguistic features of classification as well as the word's associative or semantic interconnections with other words [Ooi 1998].

With the integration of computational lexicon Machine Translation will be accurate. For example, in order to automatically translate the word 'Indian indigo' in English to its equivalent in Malayalam, the description of this is verified in an English lexicon and it was found that the presence of this word is in the domain of medical plants. Equivalent word has to be searched in the Malayalam lexicon (in the medical plants category) where the English name 'Indian indigo' for *niila amari* (നീല അമരി) is mentioned. This will help to retrieve more details about the language than from a traditional lexicon.

Computational lexicon and other related tools are being developed for most of the advanced languages for efficient processing of the language by computers. Usually computational linguists make use of manually typed limited words as lexical database for language processing applications. Such an approach is time consuming, limited in number of samples, having an intuitive nature and often non-representative.

Many dictionaries are developed / developing / being developed by different groups. But they are designed for specific purposes or as a tool for computer assisted language learning/teaching, for spell/grammar checking, machine/human aided translation, etc. In this age of information explosion, retrieving the information should be easy, compatible, vivid, accurate and must be customisable. So the need for developing a well represented computational lexicon is a must for understanding the language to its depth and for the application in language engineering. Many attempts for developing computational lexicons for Arabic, Hebrew, Maltese are in progress. WordNet [w2], the widely used lexical database for different languages can be developed to a full-fledged computational lexicon.

1.1 The Proposed Work

In this research an attempt has been made to

1. design and develop a computational lexicon for Malayalam from a corpus generated from Malayalam digital documents available in world wide web and
2. automatically identify the foreign words

The following seven major tasks are addressed as part of this aim

1. Creation of a corpus for Malayalam language

Corpus is a huge collection of digital text documents. Here a corpus, the "Malayalam corpus" is developed using a web crawler. This corpus represents the current Malayalam language in use. The corpus is used in the work for collecting words for analysis, developing computational grammar, designing rules, studying the phonemic pattern of words, training and testing different programs etc.

2. Designing of a computational grammar

Computational grammar is designed to understand the suffixes added with a word, grammatical features, morphophonemic combination of root words with suffixes etc. The list of suffixes are manually identified and classified according to the function they possess, whether they come with noun (noun suffix), with verb (verb suffix) or with both noun and verb (dual functional suffix).

The computational grammar is implemented to obtain:

a) All the words in the corpus
b) Root form of the words
c) Syllabic structure of words
d) Grammatical category of each word
e) Grammatical category of root word
f) Morphophonemic variations of root form with suffixes
g) Sandhi rules for root word identification etc.

The inferences and exceptions of this rule based approach are also discussed.

3. Pre-processing of corpus

Word segmentation and syllabification are carried out in order to prepare the words to generate their root form. This root word is considered as the lexical entry in the computational lexicon.

4. Root word identification

Root Word Identifier (RWI) is a program designed to find out the root words in the corpus and to extract them by removing the agglutinated suffixes by using morphophonemic rules. Noun, verb, dual functional suffixes and exceptional words are the inputs to the RWI. The output

is designed in such a way as to obtain a lexicon containing the root words with corresponding grammatical properties.

5. Back end processing for studying the phonotactic pattern of words

The back end processing module is a part of the whole system which consists of a training corpus for modelling the phonemic combination of English and Malayalam words.

6. Statistical Language modelling of words

Language modelling of words is done by using n-gram for analysing the phonotactic pattern of words.

7. Automatic identification of foreign words

Foreign words or *out-of-vocabulary words* or *loan words* are those words which are borrowed from other languages, written in native script. English and Sanskrit words are the most frequently occurring foreign words in Malayalam language. Here the English words written in Malayalam orthography (Eg. മൊബൈൽ (mobile), നാനോ (nano) etc.) are termed as foreign words assuming that they are of English origin. A module is implemented for identification of borrowed words from English language alone assuming the Sanskrit words as Malayalam words.

The root words obtained from RWI is compared with the modelled words obtained from back end processing thereby deciding whether the word is of native origin or of a foreign origin.

The architecture of the proposed system for generating the computational lexicon is diagrammatically shown in Figure 1.1. In this pictorial representation, rectangles represent the input/output to the system and ovals represent the process. Input and output to the system are simple files.

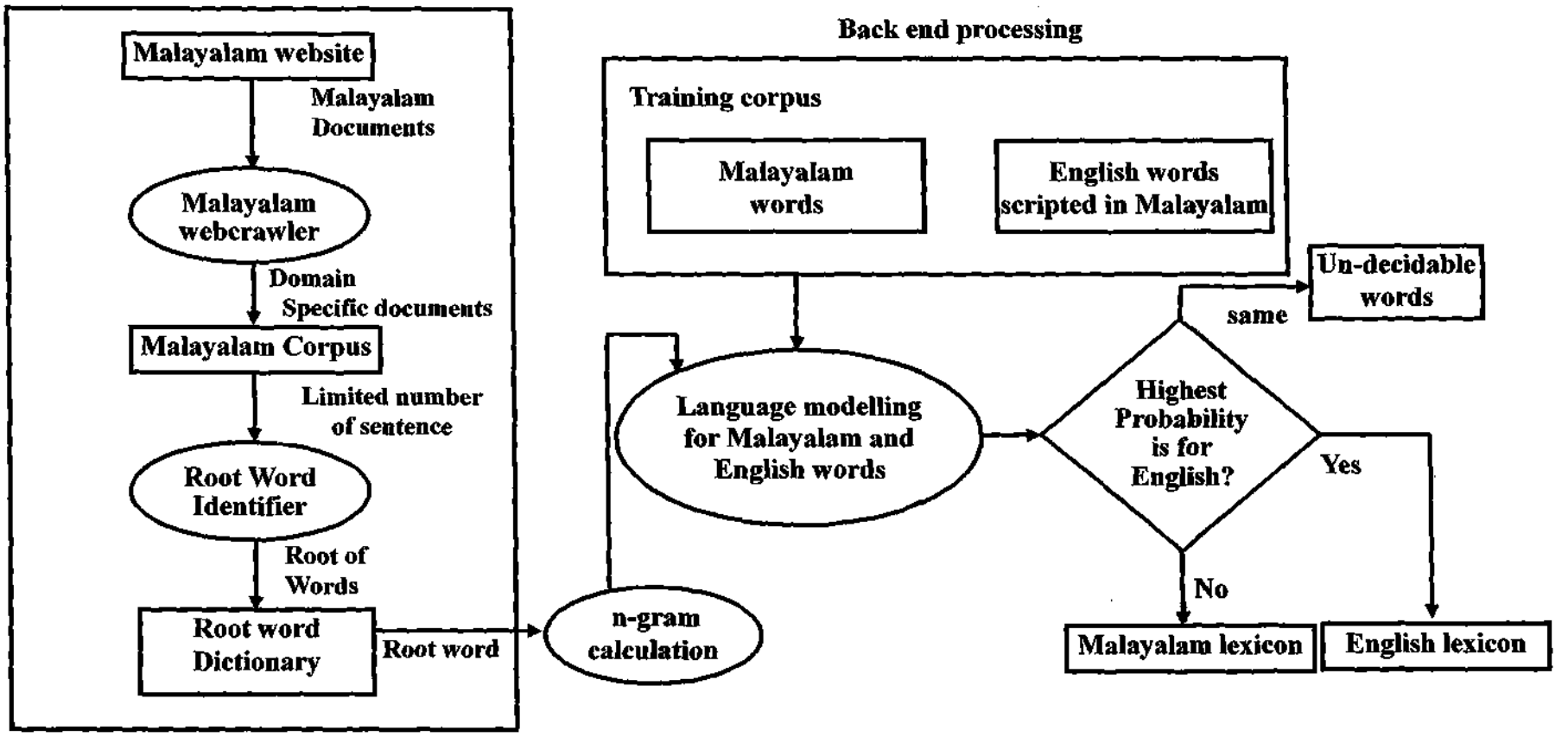


Figure 1.1: Full architecture of the proposed system

1.2 The Task and Goal

The work can be implemented using Practical Extraction and Report Language (PERL) [Wall et al. 2000] in Linux environment using the latest version of Ubandu Operating System (Ubandu 10.04 LTS version). Unicode supportive font *Meera* [w3] can be used for the display of Malayalam scripts. Free open source platform Apertium [w4] or transliteration using WX notation [Gupta et al. 2010] which are commonly used for developing similar systems are not used in the proposed work.

The system, if implemented, will be used for extracting lexical information from the Malayalam documents such as:

a) Words
b) Root form of words
c) Noun words
d) Pronouns
e) Postposition
f) Verb words
g) Clitics
h) Syllabic structure of word
i) Morphemic structure of word
j) Suffixes agglutinated with noun, verb, noun and verb
k) Inflected form of root words
l) Name of each suffixes
m) Native words
n) Foreign words
o) Morphophonemic rules involved in each word and
p) Phonotactics of native and foreign words.

In addition to the computational lexicon, suffix dictionaries, noun dictionary, verb dictionary, exceptional dictionary, root dictionary are also obtained from the system. Suffix dictionaries contain noun suffixes, verb suffixes and dual functional suffixes.

A total of 132 suffix categories with 24 noun suffixes, 42 pronouns, 34 postpositional suffixes, 51 verb suffixes and 23 dual functional suffixes are manually identified. By applying the morphophonemic rules, these suffixes can be removed and root form of the word is derived.

Thus the proposed computational lexicon extracts around 21,500 lexical entries correctly with its grammatical category from a sample corpus of around 23000 words. From around 19000 root words, etymology of 3000 words is identified as English. Identifying the meaning of these words is not addressed in this research.

Statistical evaluation metrics like Precision and Recall [Manning 1999] is to be used to evaluate the performance of the RWI system and for the identification of foreign words. In order to measure the overall performance, F-measure is also to be evaluated. The values for these measures showed by Root Word Identifier system are to be noticed. For the identification of foreign words the measurements are also be calculated.

Some of the important inferences to be obtained from the results are (1) The suffix attached with the root form of the word can decide the grammatical category of the root word, (2) The phonemic combination of words helps to trace out the etymology of words, (3) Statistical techniques like n-gram can be implemented to explore the phonemic co-occurrence pattern, useful for Malayalam computing. (4) In the process of designing and developing a corpus based computational lexicon, it seems to be sufficient to identify whether the head word or root word is a noun or a verb. Suffixes which agglutinate with them can be used further to find its other grammatical categories like adjective, adverb etc. using the proposed lexicon.

1.3 Scope and Applications

The system developed can be integrated with many language processing tools and software. The lexicon developed from the corpus can be successfully applied to high entity tasks like Machine Translation, lexicography, coining of technical terminologies, speech synthesis, developing bilingual dictionary, pronunciation dictionary etc.

The output obtained from this research can also be used in applications like parts-of-speech tagging, chunking, spell/grammar checkers, named entity recognition, identifying lexeme for building lexicon, detection of compound words, developing a transliterated Malayalam and English dictionary, for improving the accuracy rate of morphological analyser / generator, Malayalam to English parallel sentence generation, syntactic analysis, semantic and anaphoric annotation, cross-lingual information retrieval system etc.

The huge list of native and non-native root words have applications in phonological analysis, automatic tagging of foreign words (proper names of person, places, objects etc.), for

studying foreign influence in a language and are useful for language learners, students, computational linguist, computational lexicographers, grammarians and the researchers concentrating on corpus based approaches.

1.4 Organization of the Thesis

The thesis is organized into eight chapters. Chapter I give the introduction about the work. Chapter II discuss about the concepts of dictionary, lexicon, lexicography, Machine Readable Dictionaries and computational lexicon. Details about the present scenario of lexicography in different languages are also discussed. Related works discussed about morphological analysis, parts-of-speech tagger and foreign word identification in different languages. Chapter III (lexical structure and computational grammar of Malayalam) shows how the words are analysed to develop a computational grammar. The morphophonemic changes are studied and morphophonemic rules are developed for finding the root form of words. All these are implemented in chapter IV. In order to identify the foreign words, statistical language modelling is carried out and implemented in chapter V. These chapters also discuss about the unprocessed words with sufficient recommendation and further scope. The chapter VI will explain and evaluate the whole system. The last part gives the application and different areas in which the proposed system can be used. The conclusion is in chapter VII which also provide the scope of further research in this area.

Chapter II

CORPUS BASED LEXICOGRAPHY

This chapter reviews the general background of the two important repositories required for developing a computational lexicon, the corpus and the lexical database. The chapter also discusses the concept of lexicon, computational lexicon, its advantages over traditional lexicon, information that can be computationally extracted from a lexicon, foreign word identification etc. A short view of related works is also presented in this chapter.

Natural Language Processing (NLP) or Language Technology or Language Engineering is a sub field of Artificial intelligence. It deals with the analysis and processing of natural language by computers. It studies the problem of automated understanding and generation of human languages. Different stages of linguistic analysis such as phonological, morphological, syntactic, semantic, discourse, pragmatic etc., have to be carried out for the automatic processing of natural language. Some of the applications in this field are Automatic Summerization (AS), Information Extraction (IE), Information Retrieval (IR), Machine Translation (MT), Named Entity Recognition (NER), Natural Language Generation (NLG), Natural Language Understanding (NLU), Optical Character Recogniser (OCR), Anaphora Resolution (AR), Question Answering (QA), Speech Recognition (SR), Speech Synthesis (SS), Spoken Dialogue System (SDS), Text Simplification (TS), Word Sense Disambiguation (WSD) and Foreign Language reading and writing aid. There are different approaches such as rule based, symbolic, corpus based, corpus driven, statistical, connectionist and hybrid methods for developing these systems.

For all these applications a language tool known as computational lexicon which provides almost all linguistically relevant information at the lexical level is required. For developing a computational lexicon, a lexicographer prefers to obtain the data which mirrors the current status of the language in use. Malayalam documents having large collection of words are required as the resource for developing lexicon.

Corpus is a real time authentic resource useful for language investigations and for developing language tools. This widely accepted electronic text can be used to understand almost all the features of a language. An attempt is made to automatically extract Malayalam web contents to develop a corpus for Malayalam language, named here as "Malayalam Corpus". This corpus is used in the present work to analyse the properties of the language, develop a computational grammar, for statistical language modelling and to test the proposed system.

2.1 Lexicon and Lexicography

Lexicon is a place where facts about the words are recorded aimed to make use of by scholars, students and people from all walks of life [Litkowski 2005]. It can act as a general reference tool, a language standardiser, an aid for studying foreign language, a vocabulary builder, and for browsing books.

Dictionary is a type of reference work in which the words of a language, language variety, speaker or text are listed and explained, either in alphabetical or in thematic order [Zgusta 1971]. There are different types of dictionaries [NairS 1982] like monolingual, bilingual, concise, learners etc. and lexical sources such as glossary, thesaurus, morphological /paradigm /spelling/ concordance dictionaries etc. These are related to lexicon which provides limited lexical information. The entries in a lexicon will be exhaustive that it will include almost all the old and new words in a language. For example, all words used by O.Chandu Menon in his work *induleekha* (1889) [w5] may not be found in a dictionary but it will be definitely included in a lexicon.

The art or practice of writing dictionaries or the science of methods of compiling dictionaries is a part of Applied Linguistics and is known as lexicography [Jackson 2002]. The term Lexicography which is derived from the Greek word *lexico* means word and *graph* means writing [Zgusta 1980]. Lexicology is the scientific study of the words and aims to discover the principles underlying its behaviour and use. It concerns with the description of a range of observable phenomena which it defines by drawing on a set of linguistic principles [Ooi 1998].

2.1.1 Classical Methodology of Compiling Lexicon

Classical methodology for lexicography starts with the planning of a dictionary. First decide the type of the dictionary (reference/dialect/ learners/academic). Another point to decide is, whether it is necessary to incorporate earlier literature in order to obtain idioms, proverbs etc. The social and stylistic variations of the language are also to be considered. Compiling a lexicon is a long, time-consuming task which takes several years. The different phases include [Hartmann 1983] [Pillai 1965].

Phase I: Preparation (planning of a dictionary; collection of data and selection of entries)

Phase II: Editing (setting of entry; fixation of headwords; supply of pronunciation, grammatical features and definitions, synonyms, usage, and citation of head words)

Phase III: Preparation of press copy (arrangement of entries, use of notations and preparing an introduction for the dictionary)

It is the responsibility of a lexicographer to decide whether the dictionary aims at presenting all the professional registers, all the slangs and vulgarisms (or decide how much of them are to be given in the dictionary). The collection of data for a dictionary is done by the method of 'extraction'. A single lexical unit is extracted on one 'card' with its full context which is adequate enough to express the meaning of the lexical unit clearly and unambiguously. In the editing phase, the lexical unit is related with the grammatical and semantic features. Arrangement of lexical entries with details are carried out either in alphabetical, semantic or in general.

In all the above phases, lexicographer faces different problems [Halliday et al. 2004]. All the decisions taken by him/her should be recorded so that when new hands join the project there is no difficulty in continuing the line of work. The blue print may contain details like (1) Collection of materials (sources) (2) Preparation and filling up of cards (3) Compilation of word list (4) structure of entry (5) Description & definition of meaning (6) Labels (7) Phraseology (8) Illustrations (9) Grammatical characteristics (10) Script (11) Pronunciation etc.

2.1.2 Dictionary of the Age: Electronic Dictionaries and Machine Readable Dictionaries

The concept of the lexicography is now in a new generation with the advent of computers having huge storage facilities, techniques for easy retrieval, automatic analysis, sophisticated database design and exchange of data. In the early 1990's of the last millennium, computer technology made it possible to release dictionaries in floppy diskettes or in CD. The electronic edition of lexicon [Schryver et al. 2003] is a heterogeneous system, in which information can be made available in hard copy on magnetic tapes and discs, multi volume reference package (Eg. Microsoft bookshelf CD-ROM), video discs and also from online sources such as internet and mobile. Such a lexicon facilitates informal, independent and individual leaning. There are a large number of on line dictionaries for Malayalam (Eg. [w6a]-[w6e]), Machine Readable Dictionaries (MRD's) and online vocabulary builders which is freely available in the World Wide Web. Malayalam dictionaries which can be used in mobile phones are also freely downloadable.

2.1.3 Computational Lexicon and Computational Lexicography

Bennett defined computational lexicography as the term which refers to the automation of lexicographic task [Bennett et al. 1986]. A computational lexicon is an electronic version of dictionaries which can be accessed and manipulated using computers. It is open ended and the size can be increased to any extent. The information on a particular lexical item can include anything from all inflectional and derivational forms and their related information, synonyms or

antonyms, to required encyclopaedic or pragmatic knowledge, apart from the conventional information [Shanavas 1996]. Structure of a lexicon depends upon its use and its size depends on its application.

The first proposal for the design of a computational lexicon was made by the Soviet linguist, Igor Melcuk, and his team who proposed the concept of the explanatory combinational dictionary [Marcus 1986]. COBUILD, LDOCE, COMPLEX, CELEX, COMLEX and FrameNet are the leading computational lexical databases available for different languages developed by Sinclair (1984), Cowie (1986), Klavans (1988), Piepenbrock (1993), Grishman (1994), and Fillmore (2001) [Sinclair 1984] [Cowie 1989] [Klavans 1988] [Piepenbrock et al. 1995] [Grishman et al. 1994] [Fillmore et al. 2001] respectively.

Dogru and Slagle (1999) proposed a model of lexicon which involves automatic acquisition of the words as well as representation of the semantic content of individual lexical entries [Dogru et al. 1999]. Kazakov et al. (1999) reported research on word segmentation based on automatically generated annotated lexicon of word tag pairs [Kazakov et al. 1999]. Lovis proposed the design of a lexicon for use in the NLP of medical texts [Lovis et al. 1998]. Pedersen and Bruce (1998) proposed a corpus based approach to word-sense disambiguation that only requires information that can be automatically extracted from untagged text [Pedersen et al. 1998]. Development of computational lexicon for contemporary Hebrew language [Itai et al. 2006], Thai language [Noikongka et al. 2007], Maltese [Rosner et al. 1998], Arabic language [Al-Shalabi et al. 2004] [Gangemi et al. 2006] [Al-Yahya et al. 2010], Korean language [Lim et al. 2006] are on progress.

WordNet [Fellbaum 1998] is one of the widely used lexicon based on a semantic network. It was originally conceived and developed as a lexical database for English on the basis of Psycholinguistic properties. The major lexical categories like Noun, Verb, Adjective and Adverb are organized in terms of sets of synonyms (synsets) each one representing a lexical concept. WordNet has been restructured according to the principles of formal ontology in the OntoWordNet project [Gangemi et al. 2010] and has been represented using the W3C standard Web Ontology language (OWL) [Scheffczyk et al. 2006]. The different structural models which are adopted for representing word semantics include semantic network models, frame-based models and ontology-based models [Al-Yahya et al. 2010]. VerbNet [Schuler 2009] is an online verb lexicon for English language.

A National workshop on WordNet was organized by the Indian Institute of Technology, Bombay and Amrita University, Coimbatore on June 2009 to integrate the developments of WordNets in Indian Languages. A number of Indo-Aryan language WordNets are being developed at the Centre for Indian Language Technology (CFIL) at Department of Computer Science and Engineering, IIT Bombay. An IndoWordNet [Bhattacharyya et al. 2010] was developed to link the WordNets of 17 Indian Languages using Suggested Upper Merged Ontology (SUMO) [Bhatt et al. 2011]. IIT Bombay is having a Hindi WordNet [Chakrabarti et al. 2004]. The first version of Tamil WordNet has been prepared by AUKBC and Tamil University [Thiyagarajan et al. 2002]. Telugu WordNet [Selvaraj 2010] is being constructed expanding from Hindi WordNet. *SanskritNet* has been developed for Sanskrit WordNet [Kulkarni et al. 2010]. There is also Oriya WordNet [Mohanty et al. 2002]. LexTool is a tool used to develop Nepali lexicon [Bista et al. 2007]. All these WordNets do not contain the full range of information that is found in an ordinary dictionary. A Unified Computational Lexicon for Hindi-English Code-Switching [Achla et al. 2004], Bengla lexicon [Pavel et al. 2006], and Corpus based Urdu lexicon [Ijaz et al. 2007] are notable works. Amrita University, Coimbatore is engaged in developing Malayalam WordNet [W11].

Developing an electronic lexicon cannot be completely achieved through the traditional methodology of dictionary making. The five stages of development of a computational lexicon are

1) Lexicon extraction and building,
2) Lexicon based language modelling
3) Computational storage of the lexicon
4) The development and enrichment of richer lexicons and
5) Defining standards for lexicon exchange and reusability [Ooi 1998].

2.1.4 Lexical Database of a Corpus Based Lexicon

Words are tools of the basic communication which is omnipresent in every language. All words are unique having their own meaning and functions. An ideal lexicon should be able to provide all the features of every word used in a particular language. A corpus based lexicon [Granger et al. 2003] incorporating statistical methods and/or rule based approaches can be made to avail some of the following details [Hartmann et al. 1998] about the lexical items in a language:- (1) Head words (2) Compound words (3) Technical words (4) Foreign words (5) Empty words

(6) Function words (7) Proper names (8) Affixes and suffixes (9) Syllables (10) Set expressions and multi word units (9) Proverbs (10) Clitics (11) Synonyms (12) Antonyms (13) Hypernyms (14) Hyponyms (15) Acronyms (16) Phrases and idioms (17) Abbreviations (18) Spelling (19) Pronunciation (20) Etymology (21) Grammatical information (22) Semantic information (23) Thesaurus (24) Definitional meaning (25) Description and definition (26) Equation (27) Illustrative examples (28) Quotation (29) Citation (30) Illustrative pictures (31) Labels (32) Phonemic / syllabic structure (33) Word / lexical analysis (34) Word / lexical formation (35) Word structure (36) Lexical classification (37) Word generation (38) Archaism (39) Agnonym (40) Ghost words (41) Domain (42) Alphanumeric combinations (43) Aptronyms (44) Archilexeme (45) Assimilated words (46) Blending words (47) Catch phrases (48) Deprecated terms (49) Mnemonic device (50) Palindromes (51) Tongue twisters etc.

2.2 Malayalam Lexicography

Malayalam is the official language of the state of Kerala situated in the southern half of west coast of India. The state which is literally known as *God's own country* is bounded by Arabian Sea in the West and Western Ghats in the East. Malayalam language is one among the 22 official languages of India and one among the four major languages of the Dravidian family. A speaker of Malayalam is called a *malayaaLi* (മലയാളി). The influence of other languages like Sanskrit, Tamil, Telugu, Tulu, Toda, Kota, Kodagu and Badaga is seen in phonemic, morphemic and grammatical levels of this language. Malayalam is a morphologically rich and agglutinative language [Caldwell 1956]. It is relatively of free order [Varma 1999].

India has a long and ancient lexicographic tradition with Yaksha's *Nirukta*, the earliest etymological work of Sanskrit, *Amarkosh* of Amarasimha, a traditional verse dictionary, and with many *nighaNtus* [Rajashekaran 2008]. Malayalam lexicography started in 1746 with Arnose Padiri's dictionary, *'Dictionarium Malabaricum'*. The dictionary of Richard Collins (1865), *MalayaalaM~ nighaNtu* is the first monolingual dictionary in Malayalam. Other notable works are Benjamin Bailey's *Dictionary of High and Colloquial Malayalam and English* (1846), Dr. H. Gundert's *Malayalam and English dictionary* (1872), the *Sabdathaaraavali* by Sreekanthesvaram Padmanabha Pillai (1932) [Pillai 2011] etc. A pedagogical dictionary in Malayalam (Sabda Surabhi, 2 volumes, 2005) compiled by former Malayalam Lexicon chief editor B.C. Balakrishnan, contains all the Malayalam words available in the Malayalam books upto 12th standard with details and meaning. This is a good lexical resource for researchers. The *Malayalam lexicon* [Pillai 1965] of the Department of Lexicon, University of Kerala is a

Malayalam-Malayalam-English lexicon, is useful for both native and foreign learners. The compilation is still to complete and volume 9, 10 and 11 are in its final stage.

Many efforts are going on in different parts of the country for the development of lexical resources such as thesaurus development, ontology development, lexical modelling etc. [Murthy 2004]. The ongoing projects are corpus development and lexical resources by CIIL Mysore, Malayalam-Hindi bilingual lexicon for *Anusarak* project [Bharathi et al. 1997], English-Malayalam bilingual lexicon for *Anglabharati* project etc. Technology and Resource Centre for Malayalam Language (TRCML), University of Kerala is also engaged in the preparation of Malayalam corpus and lexical resources for Malayalam using modern technologies [w8]. The two projects, "Preparation of generative lexicon using MRD's" and systems for "Indian Language to Indian Languages" aimed to develop Machine translation systems. The by-products of these projects are a bilingual mapping dictionary of Tamil-Malayalam and Malayalam-Tamil dictionaries [Rajendran 2010]. AUKBC Research Centre, Chennai has undertaken project on Tamil-Hindi transfer lexicon, Named Entity Recognizer, Anaphora Resolution [Sobha 1999] which are useful for the creation of lexical resources for Indian Languages. The CALTS of Hyderabad, is engaged in preparation of lexical resources for Indian Languages. Anna University of Chennai developed an online pedagogical dictionary (English-Tamil) which is a notable contribution to the Tamil lexical resources under the Technology Development of Indian Languages (TDIL). Tamil-Malayalam translator's dictionary [Rajashekaran 2008], Telugu-Malayalam *Nighantuvu* [Rao et al. 2010], etc. are in progress. Centre for Development for Imaging Technology (C-DIT), Centre for Development of Advanced Computing (CDAC), Language Technologies Research Centre (LTRC), IIIT Hyderabad, Cochin University of Science and Technology (CUSAT), Swathantra Malayalam Computing Group, etc. are also engaged in the development of different language tools for Malayalam computing [w9a]-[w9e].

2.3 Corpus

The term *corpus* is derived from the Latin word 'corpus', means 'body' [Dash 2005]. In the domain of law, it represents a set or a collection of legal documents and their sources. The corpus of a trust is defined as the sum of money or property that is set aside to produce income for a named beneficiary [w10]. One can trace many definitions for corpus [Dash 2005]. Within the domain of general linguistics, it refers to "*A collection of linguistic data, either written text or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypothesis about a language.*" [Crystal 2003]. Another definition for

corpus is that *it is a large collection of written and/or spoken text samples available in machine-readable form, collected in a scientific way to represent the use of a language* [Dash et al. 2003]. It is not simply a random collection of texts, but represents a natural language or some part of a language used in the real world. It can be considered as a 'language research sample' for doing any type of research with that language. The language corpus is also defined as *Capable Of Representing Potentially Unlimited Selections of texts which is Compatible to computers, Operational in research and applications, Representative of a language or variety, Processable by man and machine, Unlimited in amount of data, and Systematic in formation and representation* [Dash 2005]. The validity of a corpus as a source for lexical description depends on how representative or well-balanced it is for the language it purports to represent [Ooi 1998].

Thus, corpus (plural *corpora*) is a large structured set of text documents which is electronically stored and processed. It may contain texts in single language (*monolingual corpus*), two languages (bilingual corpus) or in multiple languages (*multilingual corpus*) [Biber et al. 1998]. Corpus can also be classified as *written corpus* and *speech corpus*. The written corpus contains documents collected from books, newspapers, encyclopaedias, journals or an advertisement paper of varying length. The speech corpus contains speech samples obtained from natural dialogues, formal or informal talks or similar speech sequences. A written corpus can be developed by manually inputing the textual data to a computer thereby storing them in electronic format which can be easily processed and retrieved. Scanning the text using optical character reader (OCR), downloading the texts available on the internet, copying the text from CD ROM are some of the other techniques for collecting text samples.

The first major corpus of English, designed for computer analysis, was the Brown Corpus developed at Brown University by Henry Kucera and Nelson Francis in the mid- 1960's. It was Nelson Francis [Teubert et al. 2007], who first applied the term corpus to his electronic collection of text and John Sinclair [Teubert et al. 2007] was the first to use a corpus specifically for lexical investigation. The first corpus in Indian language is the *Kolhapur Corpus of Indian English (KCIE)* developed by Prof. S.V Shastri (1988) and his colleagues at Shivaji University, Kolhapur. It contains approximately one million words of Indian English [Dash et al. 2003]. Some of the English corpus in use are British National Corpus (BNC), European Corpus Initiative Multilingual Corpus I (ECI/MCI), EMILLE/CIIL, Lancaster-Oslo-Bergen (LOB) etc. [Teubert 2000]. The British National Corpus is a 100 million word collection of samples of written (90%) and spoken (10%) language from a wide range of sources, designed to represent a wide cross-

section of British English. Now there are monolingual corpora for fourteen south Asian languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telugu and Urdu [McEnery et al. 2006]. The corpus developed by Central Institute of Indian Languages (CIIL) Mysore for Malayalam language is a well known corpus which is freely downloadable and can be used for language research. The current status of entire corpus is having 52,18,043 words [w11].

Corpus can be from any domain and one can utilise a corpus for knowing more about their field of interest. Another advantage is that a corpus can provide huge amount of information in a single stretch without browsing the net or referring the library for hours. Even journalists, script writers, poets, politicians or a priest can utilise a corpus for improving their writings and speeches which enrich their vocabulary. Teachers, students, and researchers from the domains of languages can use corpora for making teaching materials, classroom exercises, dissertations, etc. The software developers, language engineers and programmers use corpora primarily to develop language reference tools and Natural Language Processing applications. Corpora allow researchers not only to verify the traditional approaches to language but also to observe new phenomena that have not been noticed before. The scientific study of linguistic phenomenon and analysis of data obtained through large collections of machine-readable texts (corpora) is known as Corpus Linguistics [Mitkove 2001].

The linguistic information implicitly present in the text can be made explicit through the process of concrete annotation [Dash 2005] like parts-of-speech (POS). It is the process of annotating parts of speech (noun, adjective, verb, etc.) to the corpus using linguistic and probabilistic rules thereby identifying the grammatical categories of the word. Multifunctionality, reusability, efficient data extractions are some of the advantages of annotated corpus.

A word in a language can occur at one or more domains or registers [Biber et al. 1998]. *Registers* are defined as varieties of language that is used in different situations or domains. A lexicon should specify in which domain / domains or context the word appears. So the documents seen in the web sites should be stored according to the domain.

Corpus plays an important role in linguistic analysis of a language and in the field of language technology. There are many language processing tools like word processor, morphological analyser and generator, spell checker and grammar checker, sentence parser etc. to analyse the language computationally.

For developing language tools, many researchers and institutions are creating their own corpora, solving problems in the corpus without studying on the issue, and producing project specific systems, which cannot easily be re-used. Thus the process of creating a corpus is an area, which lacks standardization and appropriate tools. The corpus for Malayalam developed so far, are not large enough and neither widely representative nor balanced. There has never been any attempt to enlarge this corpus with regular inclusion of new text samples from different text types or sources as in Bank of English [w12] or in American National Corpus [w13]. Even though many institutions are having corpora, they are not yet released to the public domain. As part of the proposed work, an attempt is being made to develop a corpus for Malayalam language. This corpus is termed as "Malayalam Corpus".

2.4 Malayalam Corpus

For the proposed work, a corpus for Malayalam having a large collection of sentences is required to analyse the language and to design the lexicon. A full fledged lexicon is one which can provide all the words available in the vocabulary of that particular language. For extracting maximum Malayalam words, a monolingual general corpora for the language is required. A general corpus contains texts belonging to almost all subject domains obtained from different sources. So here, the type of documents or sources written in Malayalam text is collected. They are listed in Appendix I. For developing corpus, only the documents from web source are considered in this work. A general monolingual text corpus is developed using a web crawler. A web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terminologies for Web crawlers are ants, automatic indexers, bots, Web spiders, Web robots, or Web scatters [w14].

In the present work, certain keywords which are commonly seen in each domain are listed and the web crawler will crawl the documents according to these keywords. They are stored in the corresponding registers as shown in Appendix II. Using this Malayalam web crawler, huge amount of documents can be stored in word processing files in the Unicode format [w15]. Meta data, the details about the documents, are manually tagged with the text. This will help the language researcher or a computational linguist to trace the context of the documents for further investigations. Meta data usually contain extra linguistic variables like site addresses, page addresses, sub classes, keywords, institution, author, genre, style, setting, font, font type, script language, topic, date of publication and time, age and intended audience of readership, remarks etc.

Biber (1990) pointed out that small corpora can be used for investingating the more common features of language such as grammatical features [Kobayashi et al. 2000]. So a focused crawler [Chakrabarti et al. 1999] is used to crawl the documents from internet having a limited number of documents using perl [w16]. Using this web crawler, a small corpus containing 24,055 words is created by selecting documents from the domains of health, film, technology, law, religion and lyrics of songs. This corpus is used for the proposed work. The advantages of developing a corpus from web as source are discussed in the following section.

2.4.1 Web Based Malayalam Corpus

There are many advantages in developing a diachronic corpus from web even though there are some disadvantages in selecting only the web contents [Kilgarriff et al. 2003]. Online texts in machine-readable form are free and instantly available. Transferring of this digital text to corpus, with high potential content, saves considerable amount of time, manpower and money.

Online archives, blogs, personal sites, web based diaries etc. are over whelming with short stories, songs, jokes, and word puzzles. Large amount of documents in the form of PDF, HTML, XML, PPT, e-text are now available in web which is free of copyright permissions. Similarly, Wikipedia for Malayalam language provides a large collection of information sorted in alphabetic order and in domain specific categories. Spiritual collection of online books, inscriptions from Ramayana, Gita, Bible, Quran, naaraayaNiiyaM~ etc. and lyrics of film songs, folk music, poems etc. are also abundantly available in digital form. Almost all leading dailies in Malayalam language have their online editions and about hundreds of journals or weeklies are published through web. It can also be noticed that the dailies set up to six or seven editions, updating many times in a day which is accessible freely. These rich resources can provide high quantum of words, sentences, paragraphs and different constructions encapsulated in the language. Words required for discourse categories such as apologies, hedges, greetings, humour, politeness, emphasis, etc. can be obtained from a corpus. Words from sastras like jyothisa, ayurveda, tantra, silpa etc. can also be collected from a corpus. Aggregator [Horacek 2002] helps to search the web contents easily and can be adopted for this work. Search engines like Google, Bing, www.malayalamsearch.com, www.tapioca.in, www.anweshanam.com are some of them which are used to search the Malayalam contents in the web.

All aspects of language such as the pronunciation, vocabulary, structure and its use are subjected to change. The variation of language according to the domain, period, regions, speakers, writers and context can be obtained from the social networking sites like web 2.0,

orkut, facebook, twitter etc. Many new words which are not available in books or dictionaries can be seen in web documents. Taboo words, dialects, technical terms, words and expressions used by namboothiris, brahmins, fisher-folks, blacksmiths, artists etc. can also be collected from different sites. Since the old documents and new documents are available in web, words used at different points in history can be included in a lexicon. Some sites will provide the description and definition of certain lexical units which can be directly cited to a lexicon. Besides these advantages, web can be used for analysing a number of discourses other than political such as legal, business, academic, media and medical. Similarly the stylistic analysis of poems, technical vocabulary, formal and informal documents can be linguistically analysed to understand and identity them with respect to their domain. Another advantage is that the possibility of skipping the commonly used vocabularies is low.

A corpus should be error free since it has to become the test bed for many linguistic research and technological developments with language tools. Malayalam corpus developed contains different types of errors. There are spelling errors, grammar errors, errors occurred due to the non-standardisation of fonts and errors due to absence of certain characters / words / sentences in the corpus. Developing systems for correcting these errors, developing handwriting recognizes and optical character readers help to enhance the corpus creation.

2.5 Associated Stages in Computational lexicon

There are different associated stages which add more features and facilities in the development of a computational lexicon. Among them, morphological analysis of the language is an important stage. Morphological analysis has to be carried out to obtain the grammatical features and semantics of the word. Another important one is the identification of etymology of words. An overview of the works published in literature for these works in different languages are given in the following section.

2.5.1 Morphological Analyser

Morphological analyser / Morphological parser is the process of segmenting a morphologically inflected word into its root word and its associated morphological components along with the features specifying the morphological structure [Menaka et al. 2010]. The morphological analysis deals with the study of internal structure of words of a language based on its grammatical category. A Morphological analyser is a program for analysing the morphology

of a word [Beesley et al. 2003]. It reads the inflected surface form of each word in a text and provides its lexical information. Various NLP research groups have developed different methods and algorithm for morphological analysis. Some of the algorithms are language dependent and some of them are language independent [Saranya 2008].

Various methods involved in morphological analysis includes

1) **Finite State Automata (FSA):** This is used to accept or reject a string in a given language. It uses regular expressions. When the automaton is switched on, it will be in the initial stage and starts working. In the final state it will accept or reject the given string. In between the initial state and finite state there are transitions, a process of switching over to another state. Finite state approaches to morphological analyser have been used for Bantu, Persian languages etc. [Elwell 2006]. The use of finite state automaton to morphological analysers allows corrections, modifications and extensions to the lexicon. In designing a Noun Phrase Chunker for Tamil language, partial chunking of the text is done by a rule-based approach where the rules are embedded in a finite state automaton. It recognizes the chunks at a high accuracy rate and speed with good performance evaluation of the system having a recall of 93.7% and precision of 94.9% using Finite State Automata [Sundar et al. 2010]. A hybrid approach using Artificial Immunity System (AIS) and Rule Based Method is used for Phrase chunking in Malayalam [Bindu et al. 2011]. 96% precision and 93% recall is claimed for the system.

2) **Two Level Morphology:** Kimmo Koskenniemi's (1993) model is "Two-level" in the sense that a word is represented as a direct, letter-for-letter correspondence between its lexical or underlying form and its surface form. The KIMMO system [Karttunen et al. 1983] known as a PC-KIMMO had two analytical components- the rules component and the lexical component, or lexicon. The rule component consisted of two-level rules that accounted for regular phonological or orthographic alternations. The lexicon listed all morphemes (stems and affixes) in their lexical form with specified constraints [Antworth 1990].

3) **Finite State Transducers (FST):** This is an advanced version of FSA which represents a two tape automaton. One can combine lexicon, orthographic rules and spelling variations in the FST to build a morphological analyser. Kannada morphological analyser and generator utilize this principle [Veerappan et al. 2011]. There are two approaches used to build the Morphological Analyzers at LDC-IL, the Word and Paradigm Approach [Bharati

et al. 1995] and the Rule Based Affix Stripping Approach. Many attempts are there for developing morphological analysers for languages like Marathi [Vaidhya et al. 2009], Telugu, Tamil [Parameswari 2011] and Malayalam [Vinod et al. 2011] using the Apertium – Lttoolbox [w17].

4) **Stemmer Algorithm:** Stemmer is used for stripping of affixes. It uses a set of rules containing list of stems and replacement rules. The most prominent ones are those introduced by Lovins (1968), Paice/Husk(1977), Dawson (1980), Porter (1980), Krovetz (1993) and Xu and Croft (1998) [Jivani 2011]. The Porter Stemmer is based on the idea that the suffixes in the English language (approximately 1200) are mostly made up of a combination of smaller and simpler suffixes. It has five steps, and within each step, rules are applied until one of them passes the conditions. If a rule is accepted, the suffix is removed accordingly, and the next step is performed. The resultant stem at the end of the fifth step is returned [Sukhadeve et al. 2011]. Retrieval experiments on English, French, and Bengali data sets showed that the stemming algorithm is efficient for languages that are primarily suffixing in nature [Majumder et al. 2007]. The experiment also showed that corpus based analysis can be used to enhance the performance of stemming algorithms. Rogati et al. [Rogati et al. 2003] used a machine-learning approach to build an Arabic stemmer. In Indian language context, Ramanathan and Rao [Ramanathan et al. 2003] have developed a stemmer for Hindi, using rule based approach. Another statistical stemmer by Oard did the suffix discovery statistically from a text collection and eliminated them from the word endings to get the stemmed output [Oard et al. 2001]. In their experiment, the end n-grams frequencies of the strings were counted (where n = 1, 2, 3, 4) for the first 500,000 words of the text collection. Each instance of every word was considered to get these frequencies. Then, the frequency of the most common subsuming n-gram suffix was subtracted from the frequency of the corresponding (n-1)-gram. Brute Force stemmers [Kumar et al. 2011] make use of a lookup table, which contains relations between root forms and inflected forms. Many attempts were started recently to develop Malayalam morphological analysers using this algorithm.

5) **Corpus Based Approach:** Corpus based approach also being used in morphological analyser. It takes a raw corpus as input and produces a segmentation of the word forms observed in the text. Such segmentation resembles morphological segmentation [Oostdijk et al. 1994]. A morphological analysis of an Italian lexicon was done with a corpus-based approach [Zanchetta et al. 2005].

6) **Directed Acrylic Word Graph (DAWG):** DAWG [Sgarbas et al. 1995] is a very efficient data structure for lexicon representation having fast string matching, with a great variety of application. This method has been successfully implemented for Greek language by University of Partas, Greece. It can be used for both morphological analysis and generation [Mair et al. 2000]. This approach is language independent and it does not utilize any morphological rules or any other special linguistic information [Sagarbas 2000].

7) **Paradigm Based Approach:** A Paradigm refers to a complete set of related inflectional and derivational word forms of a given category. The linguist is asked to provide different tables of word forms covering the words in a language. Each word form's table covers a set of roots which means that the roots follow the pattern or paradigm. The paradigm based approach is efficient for inflectionally rich languages. The A*nusaraka* research group used a language independent paradigm based morphological compiler program for Indian Languages. Tamil morphological analyser used paradigm approach with FST, for efficient processing of the system [Parameshwari 2011].

8) **Rule based affix stripping method:** In the case of Assamese, Bengali, Bodo and Oriya, the Paradigm Model seems to be unsuitable and inadequate to capture all morphological functions. The reason for this is that these languages are classifier based languages. Even though the classifiers are finite in number, they can occur in various combinations with nouns. This would increase the manual effort of paradigm creation immensely. Moreover, in these languages morpho-phonemics does not play much role. Hence, the suffix stripping approach has been found to be suitable. As the name suggests, this method involves identifying individual suffixes from a series of suffixes attached to a stem/root, using morpheme sequencing rules. This approach is highly efficient in case of agglutinative languages. However, in languages that display tendency for morpho-phonemic changes during affixation (such as Dravidian languages), this method will require an additional component of morpho-phonemic rules besides the morpheme sequencing rules [Parakh et al. 2011].

Even though a full-fledged morphological analyser is not there for Malayalam language, there are a few attempts like the one developing by Indian Institute of Technology (IIT) Hydrabad along with Linguistic Department of Tamil University. This morph analyser is based on paradigm method and they are using *Anusaraka* tool for developing it. The different grammatical categories given by them include noun, verb, adjective, adverb, NSP and postposition and avyaya.

Morphological analysis for Malayalam verbs using a hybrid approach (paradigm and suffix stripping method) is an attempt made to attain morphological generalisation of verbs to an extent [Saranya 2008]. In that work, there is a dictionary for Malayalam, which contains lexical items, grammatical category and paradigm type. Respective words of each type of paradigm are also provided with the list of inflected forms. The program compares each inflected form and come out with dictionary lexeme and suffixes. The sandhi rules occurring when the past tense marker is attached to the stem is studied. The four sandhi rules, Elision, Augmentation, Substitution and Reduplication are also discussed in detail. The verbs are categorized into 28 classes or paradigms based on the past tense marker. They identified around 1100 inflections of verb.

Using the same hybrid approach method, a Malayalam morphological analyser using Apertium Lttoolbox was developed at Language Technology Centre, Centre for Development of Advanced Computing (C-DAC), Thiruvananthapuram [Vinod 2011] as a part of Machine Translation task. Lttoolbox is available with the Apertium toolkit, which is an open source shallow-transfer machine translation system originated with in the project "Open-Source Machine Translation for the Language of Spain". Lttoolbox can be customised to any language by including the required lexical dictionary [w17]. It uses the FST approach for doing lexical processing. A morphological dictionary, also known as morphological analyser specification file is manually created to define and use paradigms which help to share the same inflection pattern. It is in XML format and the alphabets used in the dictionary are in WX notation. The paradigm facility helps to handle the inflections of the words. In order to cover all cases they have identified 24 noun paradigms, 58 verb paradigms, 9 pronoun paradigms, 12 adjective paradigms. Paradigms are not defined for postpositions and adverbs. Using post processing module, suffix stripping module and sandhi rules certain complex words are also processed. The dictionary contains around 52,000 entries. 1000 words are taken for evaluation. They claimed 96% root identification accuracy, 82% POS identification accuracy, 70% suffix segmentation accuracy and 82.67% average accuracy for their morphological analyser.

Some of the other ongoing attempts on Malayalam computing include (1) the Data Driven approach of Malayalam verb morphology [Narendranath et al. 2011], (2) Analysis of Malayalam compound words and Implementation of a compound word splitter tool using Finite State Models [Bindu et al. 2009], (3) Named Entity Identifier for Malayalam Using Linguistic Principles Employing Statistical Methods [Bindu et al. 2011] etc.

2.5.2 Parts-Of-Speech (POS) Tagging

As a part of morphological analyser, POS tagging (Parts-Of-Speech tagging) is an important pre-processing task which assigns each word in a sentence with its parts-of-speech [Dash 2005]. This helps in doing deep parsing of text and in developing Information Extraction systems, semantic processing etc. POS tagging for different natural language texts have been developed using linguistic rule, stochastic models [Kumar et al. 2010] and a combination of both as shown in Figure 2.1.

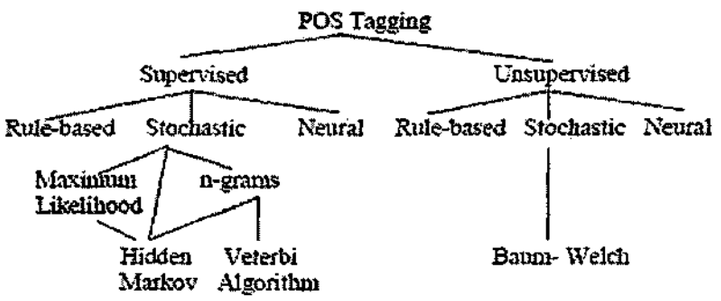


Figure 2.1: POS tagging schemes [Kumar et al. 2010]

In Supervised tagging method, a pre-tagged corpora is used to learn rules for tagging. Unsupervised tagging method do not require pre-tagged corpus and they use advanced computational techniques like the Baum-Welch algorithm [Baum 1972] to automatically induce tag sets, transformation rules etc. Based on this information, they either calculate the probabilistic information needed by the stochastic taggers or induce the contextual rules needed by rule based systems or transformation based systems [Allen 1995] [Jurafsky et al. 2000].

Hidden Markov Model (HMM) and Maximum Entropy (ME) based stochastic taggers were proposed for Bengali language [Dandapat et al. 2007]. There are different tagging approaches for Hindi language such as Morph Driven Tagger, Maximum Entropy Based Tagger, HMM based Tagger, CSF Based Tagger [Kumar et al. 2010]. A rule based part-of-speech tagging approach is used for Punjabi, which is further used in grammar checking system for Punjabi [Mandeep et al. 2008]. For Telugu, three POS taggers have been proposed by using different POS tagging approaches viz., (1) Rule-based approach, (2) using Transformation based learning (TBL) approach of Erich Brill and (3) using a machine learning technique known as Maximum Entropy Model [w18][w19].

A stochastic Hidden Markov Model (HMM) based part-of-speech tagger has been proposed for Malayalam [Manju et al. 2009]. The morphological analyzer accepts the input text and transliterated to an intermediate representation and is stored as a file. This representation is used while traversing the Finite State Automata (FSA). Each sentence is given to the *tokenizer*. The token is checked with the dictionary to check if it is a valid word. If not, then the word (token) is given to the Splitter where the word is separated into root and affix based on the orthographic rules. After Identifying the Root, the analyzer searches the affix based on the morphotactics of the category of the root word. By using the Morph Analyzer the tagged corpus is generated. The authors claimed an accuracy of about 90% for this POS Tagger.

Another tagger for Malayalam was proposed [Antony et al. 2010] which is based on machine learning approach with Support Vector Machine (SVM) [Jes´us Gim´enez 2006]. The objective was to identify the ambiguities in Malayalam lexical items and to develop an efficient and accurate POS Tagger appropriate for Malayalam. Their proposed tagset for Malayalam language has 29 tags where there are 5 tags for nouns, 1 tag for pronoun, 7 tags for verbs, 3 for punctuations, two for number, and 1 for each adjective, adverb, conjunction, echo, reduplication, intensifier, postposition, emphasize, determiners, complimentizer and question word [Manju et al. 2009]. The POS tagging architecture consists of different modules which perform different functionalities to achieve better accuracy of POS tagger. They used SVM tool for tokenization and the desired input in column format was given to this tool. Blank space is used as a column separator. The output of tokenize module is a corpus of untagged tokens. So the corpus is manually tagged using the proposed tagset. In the initial phase, 20,000 words are tagged manually. The manually tagged corpus is trained using SVM tool. This output of the tool is a dictionary with merged model and its lexicon. The remained pre-edited corpus is given to the SVM (SVMTagger, component of SVM tool) for tagging in step by step. After tagging, the displayed output is checked manually and the tags are corrected properly. The proposed POS tagger has a tagged Malayalam corpus with size of 1,80,000 tagged words. The performance of the POS tagger system in terms of accuracy is evaluated using SVMTeval. Initially, when the size of the lexicon is small the tagger achieves low accuracy. The authors claim that when the size of the lexicon is 1,80,000 the POS tagger achieves 94% accuracy.

2.5.3 Identification of Foreign Words from other languages

Etymological classification plays an important role in information retrieval and cross-lingual information retrieval system. Finding equivalents between a source word and its various

target language realizations improves indexing of search terms and subsequently document recall [Kang et al. 2002].

Oh and Choi formulates a problem of automatically identifying and extracting English words from Korean text [Oh et al. 1999]. They treated it as a syllable-tagging problem in which each syllable is identified as foreign or Korean, and each sequence of foreign tagged syllables is extracted as an English word. They modelled it with a Hidden Markov Model where states represent a binary indication of whether a syllable is Korean or not. Hand tagged corpus of one lakh words are used for finding the probability of a syllable being Korean or foreign.

Kang and Choi (2002) employed a similar advanced approach which delivered promising results. They used a sample corpus of about 1,900 documents taken from KTSET 1.0 corpus in which each syllable was manually tagged as being either Korean or foreign. As in the case of Malayalam language, word segmentation problem is an issue in Korean since some of the words in a sentence are written without space. The noun extraction task in Korean consists of detaching functional word sequences and compound noun segmentation [Jeong et al. 1997]. The compound noun is segmented into simple nouns and foreign words are identified. Statistical methods utilizing the differences in syllabic unigram or syllabic bigram patterns between pure Korean word and foreign word have been developed [Oh et al. 2002]. They used HMM based foreign word recognition method to recognize the words. The results proved that after compound noun segmentation and functional word detachment, both Precision and Recall are significantly increased when compared with those obtained only from syllable tagging. They used smoothing techniques to increase the performance by 20% thereby obtaining a Precision/Recall rate of 84.33%/92.02% [Kang et al. 2002].

For getting more accurate results for the identification of languages more training data is required. Baker and Brew (2008) addresses the issue of obtaining sufficient labelled data for training the system. As a solution to this problem a small set of phonologically-based transliteration rules were developed to generate potentially unlimited amount of pseudo training data at low cost [Baker et al. 2008]. This approach can be further used to train a classifier to distinguish etymological classes of actual words.

Goldberg and Elhadad (2008) presented a loosely supervised method for context free identification of transliterated foreign names and borrowed words in Hebrew text. The method is purely statistical and does not require the use of any lexicon or linguistic analysis tool for the source language, Hebrew. They used the approach of identifying transliteration by comparing the

letter structure of words with models trained in a way that captures the sound structure of the language-one in Hebrew and one in English, as written in Hebrew writing system. The model for transliterated English is trained on data automatically generated from an English corpus [Goldberg et al. 2008]. This noisy data performs much better than using just a small amount of hand-made clean data. Corpus taken for the experiment is 28MB of text containing prose, poetry and essay of 26 authors. For foreign models, they used text over-generated from 6MB portion of the Brown corpus distributed with NLTK tool kit. Total of 9618 words (4044 unique words) are tested. Prefixes are manually removed. Output obtained are 3608 Hebrew, 368 foreign words of which 251 are foreign proper names and 68 words are ambiguous words. 5 fold cross validation [Kohavi 1995] is used to evaluate the performance of the developed system. From their analysis, it was found that majority of foreign words in Hebrew text are proper names. By adding a lexicon, the precision increased and recall decreased. The decrease in recall is due to very common borrowed words which are included in the lexicon. The work is conducted without removing the prepositions appended as prefix to Hebrew words. Such unsegmented data which are wrongly tagged are left for future research.

Other notable works in the direction of language identification include

(1) The capturing of Out-of-Vocabulary words in Arabic Text [Abdusalam et al. 2006],
(2) Extracting loanwords from Mongolian corpora and producing a Japanese-Mongolian bilingual dictionary [Khaltar et al. 2006],
(3) Statistically-Enhanced New Word Identification in a Rule-Based Chinese System [Wu et al. 2006],
(4) Automatic Language Identification of Chinese [Mariani et al. 2008],
(5) Language Identification of Short Text Segments with n-gram models [Vatanen et al. 2010] etc.

This chapter discussed about the resources for developing a computational lexicon. The chapter also discussed different approaches attempted in Malayalam and other languages. In the proposed work, a different approach, but similar to existing ones, with minimum monolingual resource was attempted to identify the words, root words with their grammatical category. The system developed will automatically generate a lexical database with linguistic details. Next chapters discuss about the analysis of words in the Malayalam corpus to design the computationnal lexicon.

Chapter III

LEXICAL STRUCTURE AND COMPUTATIONAL GRAMMAR OF MALAYALAM

In this chapter, a structural study of Malayalam words in its surface form is carried out using linguistic theories and principles, to derive the computational grammar of Malayalam. Due to the morphologically complex and agglutinative nature of Malayalam, the morphemic structure of words in the corpus has to be analysed. The process of extracting the grammatical information encapsulated in the language is also presented. Morphophonemic changes occurring in the boundaries of root word and suffixes are studied. Manually developed rules are applied to identify the root form of a word. The analysis is concentrated only on its lexical entity level, which is context free. With the aim of generating computational rules for RWI, suffixes attached to the words are studied and classified. Different suffixes with their morpho-phonological variations during the process of suffixation are explained in detail with examples and the exceptions are identified. Lastly the morphophonemic changes (sandhi changes) occurring when the initial phonemes of the suffixes are added to the root word's end phoneme is also systematically listed.

3.1 Structural Analysis of Words

Words, the fosters of language, are the fundamental compositional units of a sentence. In the context of computational modelling, a language is viewed as a set of well-formed sentences and sequence of words framed in a finite description known as grammars [Krishnamurti 1985]. In order to understand the underlying structure of a language, systematic analysis of surface structure of word is required. Extracting the grammatical properties and understanding the word formation rules are the basic tasks in Natural Language Processing.

A corpus for Malayalam is created as discussed in section 2.4 This Malayalam corpus will contain words in its root form, inflected form, derived form, compound form and in reduplicated form. The linguistic definition of *root word* is that *'it is the form of a word after all affixes are removed* [w20]. Another definition given is that *'a root word is a real word that can make new words from root words by adding prefixes and suffixes. They are the most basic form of a word that is able to convey a particular description, thought or meaning. Prefixes and suffixes are grammatical and lingual affixes'* [w21]. Prefixes are affixed before, and suffixes after a base word or word stem to add information. Inflected words are formed by the affixation of grammatical features such a case, number, tense, aspect, mood etc. to the root word. The process

of separating the affixes from an inflected word can provide the root of the word and its grammatical information.

Common grammatical categories (parts-of-speech) for Malayalam language are noun, pronoun, verb, adverb, adjective, postpositions, clitics [Andrewskutty 1971] etc. Here all forms which are affixed immediately after the root is considered as suffix. The addition of suffixes can change the part-of-speech of the word. It is also observed that this suffix part can decide the grammatical features of the root. In this attempt of designing a "Corpus Based Computational Malayalam Lexicon", the root word identified from the corpus is considered as one of the lexical entries of the lexicon.

From the computational point of view, the internal structure of words can be studied as explained by Dash N.S (2005). Let W be a string of n characters having a form $s_1,s_2,s_3,s_4,.....s_n$ where s stands for a single character. The total number of characters in word W is denoted as |W|. If |W|=0, it is a null string. The concatenation of two strings, W1= $s_1, s_2,s_3,s_4,.....s_n$ and W2= $t_1,t_2,t_3,t_4,.....t_n$ is denoted by W1 $\oplus$W2 = $s_1, s_2, s_3, s_4,.....s_n, t_1, t_2, t_3, t_4, ...t_n$.

According to the morphology of the language, if W is a valid word, then it is possible that either W is a root word or $W = W_r \oplus W_s$, where W_r is the front substring representing the root part and W_s is the back substring representing the suffix part. They combine in such a way that W_r and W_s grammatically agree with each other. So segmentation of the suffix W_s from a word W, can provide the root form of the word, W_r.

The following examples show how a Malayalam word is integrated to W_r and W_s. Words in Malayalam script, phonetic notation, its meaning and suffix name is given with them.

Examples:

(1)	കുട്ടികൾ	-	കുട്ടി	+	കൾ		
	kuttikaL~	-	kutti	+	kaL~		
	children	-	child	+	plural suffix		
	word	-	noun root	+	noun suffix		
(2)	കുട്ടികളുടെ	-	കുട്ടി	+	കൾ	+	ഉടെ
	kuttikaLute	-	kutti	+	kaL~	+	ute
	children's	-	child	+	plural suffix	+	genitive case suffix
	word	-	noun root	+	noun suffix	+	noun suffix

(3)	കുട്ടികളുടെകൂടെ	-	കുട്ടി	+	കൾ	+	ഉടെ + കൂടെ	
	kuttikaLutekuute	-	kutti	+	kaL~	+	ute + kuute	
	with the children	-	child	+	plural suffix	+	genitive case suffix + suffix showing interior movement	
	word	-	noun root	+	noun suffix	+	noun suffix + noun suffix	
(4)	രക്ഷിക്കണേ	-	രക്ഷ	+	ഇക്ക്	+	അണേ	
	raxikkaNee	-	rax	+	ikk	+	aNee	
	please protect	-	protection	+	causative suffix	+	debitive emphatic suffix	
	word	-	verb root	+	verb suffix	+	verb suffix	

(5) വന്നുകൊണ്ടിരിക്കുകയാണോ?-

വ + ഉന്നു + കൊണ്ട് + ഇരിക്ക് + ഉക + ആണ് + ഓ ?

vannukoNtirikkukayaaNoo?- va + unnu + koNt + irikk + uka + aaN + oo

Have (you) been coming?- verb root + present tense marker + cause expressing / instrumental suffix + perfect aspects marker + infinitive suffix + equation / finite verb suffix + interrogative / or coordination suffix.

The analysis provides the manually identified suffixes and the euphonic change occurring when suffixes are concatenated with root form. Structurally,

Noun	=	root	+	noun suffix
Verb	=	root	+	verb suffix
Noun/Verb	=	word/root word	+	dual functional suffix

Examples:

(6)	കവിയെ	-	കവി	+	എ
	kaviye	-	kavi	+	e
	about poet	-	poet	+	accusative case suffix
	word	-	noun root	+	noun suffix
(7)	ചാടരുത്	-	ചാട്	+	അരുത്
	caataruth	-	caat	+	aruth
	don't jump	-	jump	+	negative imperative marker
	word	-	verb root	+	verb suffix

(8)	എഴുതിയിട്ടുള്ള	-	എഴുതി	+	ഇട്ട് + ഉള്ള
	ezhuthiyittuLLa	-	ezhuthi	+	itt + uLLa
	has been written	-	wrote	+	perfective suffix
				+	adjectival suffix / be-relative participle suffix
	word	-	verb root	+	verb suffix + dual functional suffix

3.1.1 Grammatical Suffixation

Analysis of words showed that words can be classified as root words and words having nominal suffixes, verbal suffixes and post positional suffixes. Some words are agglutinated with other words also. Here the process of suffixation is carried out in two major heads- grammatical and dual functional suffixation. Grammatical suffixation include the suffixes concatenated with noun or verb such as case, number, postpositions, tense etc. and dual functional suffixation includes suffixes or words which are agglutinated to the root form of the word. Dual functional suffixes can be a word or a phrase which gets agglutinated to the root. A total of 132 suffix categories with 24 noun suffixes, 34 postpositional suffixes, 51 verb suffixes, 23 dual functional suffixes and 42 pronouns are manually identified. It is tried to cover almost all common occurring suffixes. The analysis has been done giving importance to computational aspect, without altering the basic Linguistic theories and principles. The classification system proposed by Asher et al. (1997) is used to identify and name the suffixes. Classification of word in terms of their suffixation can be diagrammatically represented as in Figure 3.1

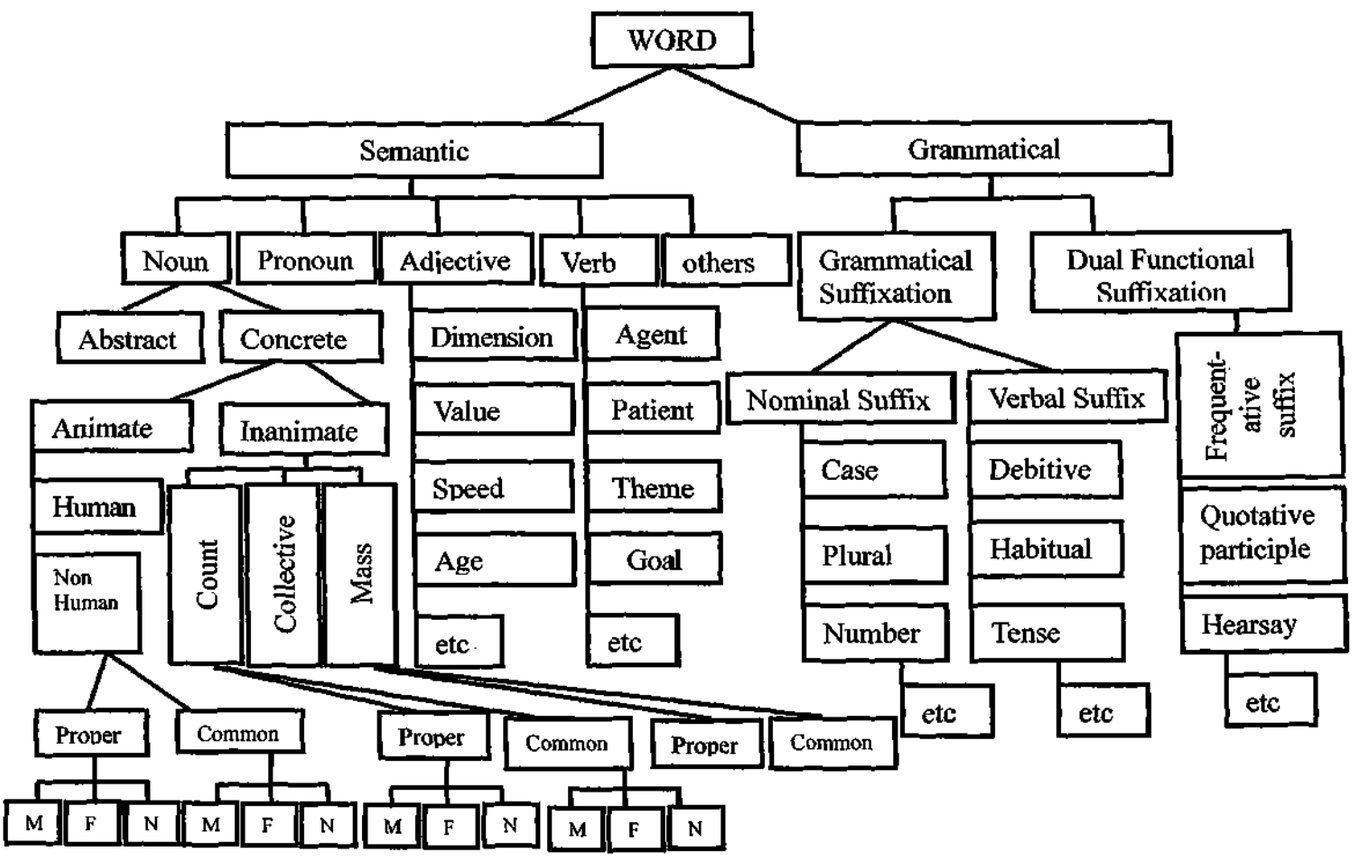


Figure 3.1: Classification of word

3.1.1.1 Noun Morphology

Nouns can occur in isolation or can take gender markers, plural markers, case suffixes, postpositions, clitics etc. They can also make compounds. Noun morphology is less complex than verb morphology. Grammatical information required for nouns are mainly case, number and gender. It can be represented as:

W = noun root ± [plural suffix] ± [case suffix] ± [postpositions] ± [clitics] ± .....

The following section gives an analysis of the suffixes and how the morphophonemic changes occur when they are suffixed to the root word. The examples and concepts cited are mainly depends upon Asher et al. (1997)

3.1.1.2 Noun Suffixation

The Encyclopaedia of Language and Linguistics defined ***Cases*** (*vibhakthi*) as *a system of marking dependent nouns for the type of relationship they bear to their heads.....typically, case marks the relationship of a noun to the verb at the clause level or of a noun to a preposition, postposition, or another noun at the phrase level* [Crystal 2003].

The unique property of case marking for nouns helps to identify the noun root. Addition of case markers may cause morphophonemic change with the final syllable of root word and with the case marker. Different cases and markers with examples are shown in Table 3.1.

Table 3.1: Case markers and their suffixes

Case	Marker	പ്രത്യയം	Example
Nominative	Nil	ϕ	makan~ മകൻ
Accusative	-e	ഏ	makane മകനെ
Dative	-kk/-U	ക്ക് / ്	makan മകന്
Sociative	-oot	ഓട്	makanoot മകനോട്
Locative	-il~	ഇൽ	makanil~ മകനിൽ
Instrumental	-aal~	ആൽ	makanaal~ മകനാൽ
Genitive	-ute/-nte	ഉടെ / ന്റെ	makante മകന്റെ

It can be found that different combinations are possible for each suffixes agglutinated with different root words. Following section discusses about it in detail. The first column in examples shows how the root word concatenates with a suffix to form a derived word. During the process of concatenation, the morphophonemic change occurring in the boundaries is shown in the second column. The analysis helps to derive the morphophonemic rules, that are the rules for identifying the root of a word. [c] denotes the consonant ending and [cc] denotes the consonant cluster ending. The phonetic notation is given in Appendix III.

NS1. Nominative case suffix

Nominative (*niR~ddeeSika*) case is unmarked for suffix. So it is the same root word. Usually denoted by ϕ.

Eg: (9)	ദേവി	+	ϕ	>	ദേവി
	deevi	+	ϕ	>	deevi (a female name)

NS2. Accusative case suffix

The accusative (*prathigraahika*) case marker is /-e/ (/-ഏ/). If the root word is suffixed with /-e/, that noun or pronoun will act as an object in a sentence. A glide /y/ and /v/ may occur when two vowels coming together at morpheme junctures. /y/ is inserted after the front vowels and /v/ is inserted after back vowels.

Eg: (10)

രാമൻ	+	ഏ	>	രാമനെ	ൻ	+	ഏ	>	നെ
raaman~	+	e	>	raamane	n~	+	e	>	ne

NS3. Dative case suffix

The dative (*uddeeSika*) case markers are /-kk/ (ക്ക്) and /-U/ (/ു/), giving the meaning 'pertaining to'. Here /-U/ is named as *candrakala* (ചന്ദ്രകല).

Eg: (11)

മനുഷ്യൻ	+	/ു/	>	മനുഷ്യന്	ൻ	+	/ു/	>	ന്
manushyan~	+	/-U/	>	manushyan	n~	+	/-U/	>	n
ജനത	+	ക്ക്	>	ജനതയ്ക്ക്	അ	+	ക്ക്	>	യ്ക്ക്
janatha	+	kk	>	janathaykk	a	+	kk	>	ykk

NS4. Sociative case suffix

Sociative (*saaM~yoojika*) case markers are /-oot/ (ഓട്) and /-ot/ (ഒട്).

Eg: (12)

പ്രേക്ഷകൻ	+	ഓട്	>	പ്രേക്ഷകനോട്	ൻ	+	ഓ	>	നോ
preexkan~	+	oot	>	preexkanoot	n~	+	oo	>	noo

NS5. Locative case suffix

The marker for locative (*aadhaarika*) case is /-il~/ (ഇൽ) and /kal~/ (കൽ). This marker denotes location. Occurrence of words like varil~ (വരിൽ), varukil~ (വരുകിൽ) etc. where '-il~' added to the root of the verb is rare in modern language.

Eg: (13)

അവൻ	+	ഇൽ	>	അവനിൽ	ൻ	+	ഇ	>	നി
avan~	+	il~	>	avanil~	n~	+	i	>	ni

NS6. Genitive case suffix

Genitive (*saM~bandhika*) case suffix are /-nte/ (ന്റെ), /-ute/ (ഉടെ) which shows possession of something.

Eg: (14)

അരുൺ	+	ന്റെ	>	അരുണിന്റെ	ൺ	+	ന്റെ	>	ണിന്റെ
aruN~	+	nte	>	aruNinte	N~	+	nte	>	Ninte
കുട്ടി	+	ഉടെ	>	കുട്ടിയുടെ	ഇ	+	ഉ	>	യു
kutti	+	ute	>	kuttiyute	i	+	u	>	yu

Finally the case which is known as **vocative** is phonologically predictable. Word-final long vowel can be taken as a rule for identifying the vocative cases.

Eg: (15) acchan (nominative) > acchaa (vocative).

NS7. Plural suffix

Noun can also be identified with number markers. The number in Malayalam can be only singular or plural. The root itself or the root with the gender suffix is used as singular, whereas

the plural is formed by adding the suffix and its allomorphs /-maaR~/ (/-മാർ/), /-ngngaL~/ (/-ങ്ങൾ/) /-kaL~/(/-കൾ/).

Eg: (16)

ആൾ + കൾ > ആളുകൾ	ൾ + ക് > ളു
aaL~ + kaL~ > aaLukaL~	L~ + k > Lu
അമ്മ + മാർ > അമ്മമാർ	അ + മ് > no change
amma + maaR~ > ammamaaR~	a + m > no change

Other suffixes that are inflected with nouns are listed below:

NS8. <u>Allative marker /-ileekk/</u>

The complex case marker /-ileekk/, (/-ഇലേക്ക്/) meaning 'to' is known as allative marker. Allative marker, place locative marker, ablative case suffix, path locative suffix are similar to locative suffix.

Eg: (17)

പട്ടണം + ഇലേക്ക് > പട്ടണത്തിലേക്ക്	ം + ഇ > ത്തി
pattaNaM~ + ileekk > pattaNaththileekk	aM~ + i > ththi

NS9. <u>Place locative suffix /-aththi/</u>

The place locative suffix is /-aththi/ (/-അത്ത്/). Usually seen in place name entities.

Eg: (18)

കൊല്ലം + അത്ത് > കൊല്ലത്ത്	ം + അ > delete both
kollaM~ + aththi > kollaththi	aM~ + a > delete both

NS10. <u>Optative suffix /-aakatte/</u>

/-aakatte/ (/-ആകട്ടെ/) is the optative form of 'aakuka' (be). Colloquially, words like 'avaraavate' (അവരാവട്ടെ), 'dathakkaLavate' (ദാതാക്കളാവട്ടെ) are also seen.

Eg: (19)

പാട്ടുകൾ + ആകട്ടെ > പാട്ടുകളാകട്ടെ	ൾ + ആ > ളാ
paattukaL~ + aakatte > paattukaLaakatte	L~ + aa > Laa

NS11. Relative adjectival suffix /-aaya/

The relative participle of 'aakuka' (ആകുക) is 'aaya' (/-ആയ/) which denotes 'an entity which is'. This is added to the nominal base

Eg: (20)

സഖി + ആയ > സഖിയായ	ഇ + ആ > യാ
sakhi + aaya > sakhiyaaya	i + aa > yaa

NS12. Negative participle suffix /-allaathe/

The negative participle form /-allaathe/ (/-അല്ലാതെ/) gives the meaning of 'besides'.

Eg: (21)

അവൾ + അല്ലാതെ > അവളല്ലാതെ	ൾ + അ > ള
avaL~ + allaathe > avaLallaathe	L~ + a > La

NS13. Path locative suffix /-iluute/

Means of expressing direction and path, include the path locative suffix /-iluute/ (/-ഇലൂടെ/) which gives the meaning 'through it'.

Eg: (22)

വഴി + ഇലൂടെ > വഴിയിലൂടെ	ഇ + ഇ > യി
vazhi + iluute > vazhiyiluute	i + i > yi

NS14. Ablative case suffix /-il~ninn/

Meaning for this suffix /-il~ninn/(/-ഇൽനിന്ന്/), is 'from that'.

Eg: (23)

മരം + ഇൽനിന്ന് > മരത്തിൽനിന്ന്	ം + ഇ > ത്തി
maraM~ + il~ninn > maraththil~ninn	aM~ + i > ththi

NS15. Negative comitative suffix /-illaathe/

Meaning of this suffix /-illaathe/ (/-ഇല്ലാതെ/), is 'without'.

Eg: (24)

നിറം + ഇല്ലാതെ > നിറമില്ലാതെ	ം + ഇ > മി
niraM~ + illaathe > niramillaathe	aM~ + i > mi

NS16. <u>Negative adjectival suffix /-illaaththa/ and /-allaaththa/</u>

Negative quality can be expressed by the use of the adjectival form /-illaaththa/ (/-ഇല്ലാത്ത/) meaning 'not having' or /-allaaththa/ (/-അല്ലാത്ത/) meaning 'other than'.

Eg: (25)

പണം + ഇല്ലാത്ത > പണമില്ലാത്ത	ം + ഇ > മി
paNaM~ + illaaththa > paNamillaaththa	aM~ + i > mi
മിടുക്കൻ + അല്ലാത്ത > മിടുക്കനല്ലാത്ത	ൻ + അ > ന
mitukkan~ + allaaththa > mitukkanallaaththa	n~ + a > na

NS17. <u>Citerior-anterior location suffix /-ethire/ and /-neeR~kk/</u>

Suffix /-ethire/ (/-എതിരെ/) means 'opposite' whereas /-neeR~kkU/ (/-നേർക്ക്/) means 'towards'.

Eg: (26)

സമരത്തിന് + എതിരെ > സമരത്തിനെതിരെ	[c]/[cc] + എ > [c]/[cc]െ
samaraththin + ethire > samaraththinethire	[c]/[cc] + e > [c]/[cc]e
അവന് + നേർക്ക് > അവനുനേർക്ക്	[c]/[cc] + ന് > [c]/[cc]ു
avan + neeR~kk > avanuneeR~kk	[c]/[cc] + n > [c]/[cc]u

NS18. <u>Suffix showing interior movement /-oththu/ /-kuute/ /-kuuti/ /-oppaM/</u>

These suffixes (/-ഒത്ത്/, /-കൂടെ/, /-കൂടി/, /-ഒപ്പം/) possesses relatively the same meaning as 'along with'.

Eg: (27)

അംഗങ്ങളോട് + ഒത്ത് > അംഗങ്ങളോടൊത്ത്	[c]/[cc] + ഒ > [c]/[cc]ൊ
aM~gangngaLoot+oththu> aM~gangngaLootoththu	[c]/[cc] + o > [c]/[cc]o
സീമയുടെ + കൂടെ > സീമയുടെകൂടെ	എ + ക് > no change
siimayute + kuute > siimayutekuute	e + k > no change

ഒത്ത് + കൂടി > ഒത്തുകൂടി	[c]/[cc] + ക് > [c]/[cc]ു
oththu + kuuti > oththukuuti	[c]/[cc] + k > [c]/[cc]u
അവൾക്ക് + ഒപ്പം > അവൾക്കൊപ്പം	[c]/[cc] + ഒ > [c]/[cc]ൊ
avaL~kk + oppaM~ > avaL~kkoppaM~	[c]/[cc] + o > [c]/[cc]o

NS19. Reciprocity suffix /-tammil~/

/-tammil~/ (/-തമ്മിൽ/) can be glossed as 'among themselves'.

Eg: (28)

അവരെ + തമ്മിൽ > അവരെതമ്മിൽ	എ + ത് > no change
avare + thammil~ > avarethammil~	e + th > no change

NS20. Partitive numeral suffix /-peer/ and /-peeR~/

These suffixes (/-പേര്/ and /-പേർ/) are intrinsically plural and it doesn't take a plural suffix. This suffix is usually used for asking/specifying the number of person.

Eg: (29)

എത്ര + പേർ > എത്രപേർ	അ + പ് > no change
ethra + peeR~ > ethrapeeR~	a + p > no change
നാല് + പേര് > നാലുപേര്	[c]/[cc] + പ് > [c]/[cc]ു
naal + peer > naalupeer	[c]/[cc] + p > [c][cc]u

NS21. Sufficient suffix /-mathi-/

/-mathi-/ (/-മതി/) the sufficient suffix means 'enough'.

Eg: (30)

ചായ + മതി > ചായമതി	അ + മ് > no change
caaya + mathi > caayamathi	a + m > no change

NS22. Question_word suffix

Question_word/wh_word suffix consist of /aar/ (ആര് -who), /enth/ (എന്ത് - why/what), /enthin/ (എന്തിന് - for what), /eeth/ (ഏത് - which), /engngane/ (എങ്ങനെ- how), /ethra/ (എത്ര -

how much/how many), /evite/ (എവിടെ - where), /engng/ (എങ്ങ് - where to), /eppooL~/ (എപ്പോൾ/ എപ്പം/എപ്പോ- when), /enthe/ (എന്തേ - why), /-ethramaathraM~/ (എത്രമാത്രം - how much).

Eg: (31)

സാർ	+	ആര്	>	സാറാര്	ർ	+	ആ	>	റാ
saaR~	+	aar	>	saaRaar	R~	+	aa	>	Raa
സുമ	+	എന്ത്	>	സുമയെന്ത്	അ	+	എ	>	യെ
suma	+	enth	>	sumayenth	a	+	e	>	ye
മകൻ	+	എന്തിന്	>	മകനെന്തിന്	ൻ	+	എ	>	നെ
makan~	+	enthin	>	makanenthin	n~	+	e	>	ne
കുട്ടി	+	ഏത്	>	കുട്ടിയേത്	ഇ	+	ഏ	>	യേ
kuutti	+	eeth	>	kuttiyeeth	e	+	ee	>	yee
മീന	+	എങ്ങനെ	>	മീനയെങ്ങനെ	അ	+	എ	>	യെ
miina	+	engngane	>	miinayengngane	a	+	e	>	ye
മണി	+	എത്ര	>	മണിയെത്ര	ഇ	+	എ	>	യെ
maNi	+	ethra	>	maNiyethra	i	+	e	>	ye
കുട്ടി	+	എവിടെ	>	കുട്ടിയെവിടെ	ഇ	+	എ	>	യെ
kutti	+	evite	>	kuttiyevite	i	+	e	>	ye
അവൾ	+	എങ്ങ്	>	അവളെങ്ങ്	ൾ	+	എ	>	ളെ
avaL~	+	engng	>	avaLengng	L~	+	e	>	L
കുട്ടി	+	എപ്പോൾ	>	കുട്ടിയെപ്പോൾ	ഇ	+	എ	>	യെ
kutti	+	eppooL~	>	kuttiyeppooL~	i	+	e	>	ye
അവളും	+	എത്രമാത്രം	>	അവളുമെത്രമാത്രം	ും	+	എ	>	മെ
avaLuM~	+	ethramaathraM~	>	avaLumethramaathraM~	aM~	+	e	>	me

NS23. <u>Emphatic marker</u> <u>/-thanne/</u>

/-thanne/ (/-തന്നെ/) the emphatic marker, can occur after the head noun.

Eg: (32)

കുട്ടി	+	തന്നെ	>	കുട്ടിതന്നെ	ഇ	+	ത്	>	no change
kutti	+	thanne	>	kuttithanne	i	+	th	>	no change

NS24. Numeral suffix

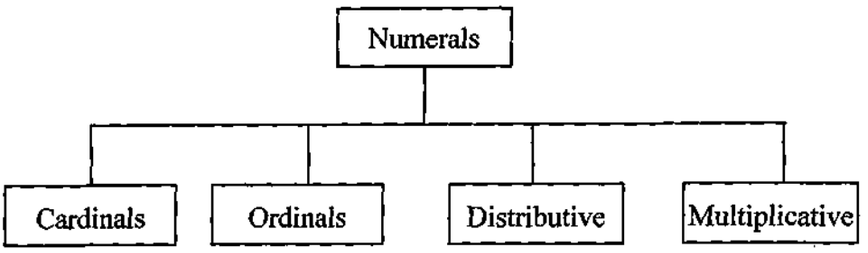


Figure 3.2: Classification of Numerals

In this analysis, cardinals showing the numbers such as onn (ഒന്ന്), raNt (രണ്ട്), muunn (മൂന്ന്) etc. are taken as root words. Ordinals are formed in Malayalam by adding the suffix /-aaM~/ (ആം) or /-aamatte/ (ആമത്തെ) to cardinal numbers. Suffix /-aaM~/ can also be considered as a future model suffix. So here only /-aamatte/ is taken as noun suffix. Distributive numerals are formed by reduplicating the cardinals. They are considered as root word. Multiplicative numerals are formed by adding the word *matangng* (മടങ്ങ് - fold) to the cardinal.

Eg: (33)

ഒന്ന് + ആമത്തെ > ഒന്നാമത്തെ	[c]/[cc] + ആ > [c]/[cc]ാ
onn + aamaththe > onnaamaththe	[c]/[cc] + aa > [c]/[cc]aa
നാല് + മടങ്ങ് > നാലുമടങ്ങ്	[c]/[cc] + മ് > [c]/[cc]ു
naal + matangng > naalumatangng	[c]/[cc] + m > [c]/[cc]u

The analysis also identified certain other suffixes capable of agglutinating with noun words. They are Existential suffix, Negative suffix, Instrumental case suffix, And co-ordination suffix, Or co-ordination suffix, Quotative participle suffix, Adjectival suffix, Equation suffix, Hearsay suffix, Predicted future time suffix, Existential interrogative participle suffix and Ordinal suffix. They have been discussed in section 3.1.2 as dual functional suffixes.

Pronouns are those words which can be used instead of nouns. Since they are in free form they are categorized as root word [Subhash 2006]. List of pronouns commonly seen in the language are listed below in Tables 3.2 to 3.4.

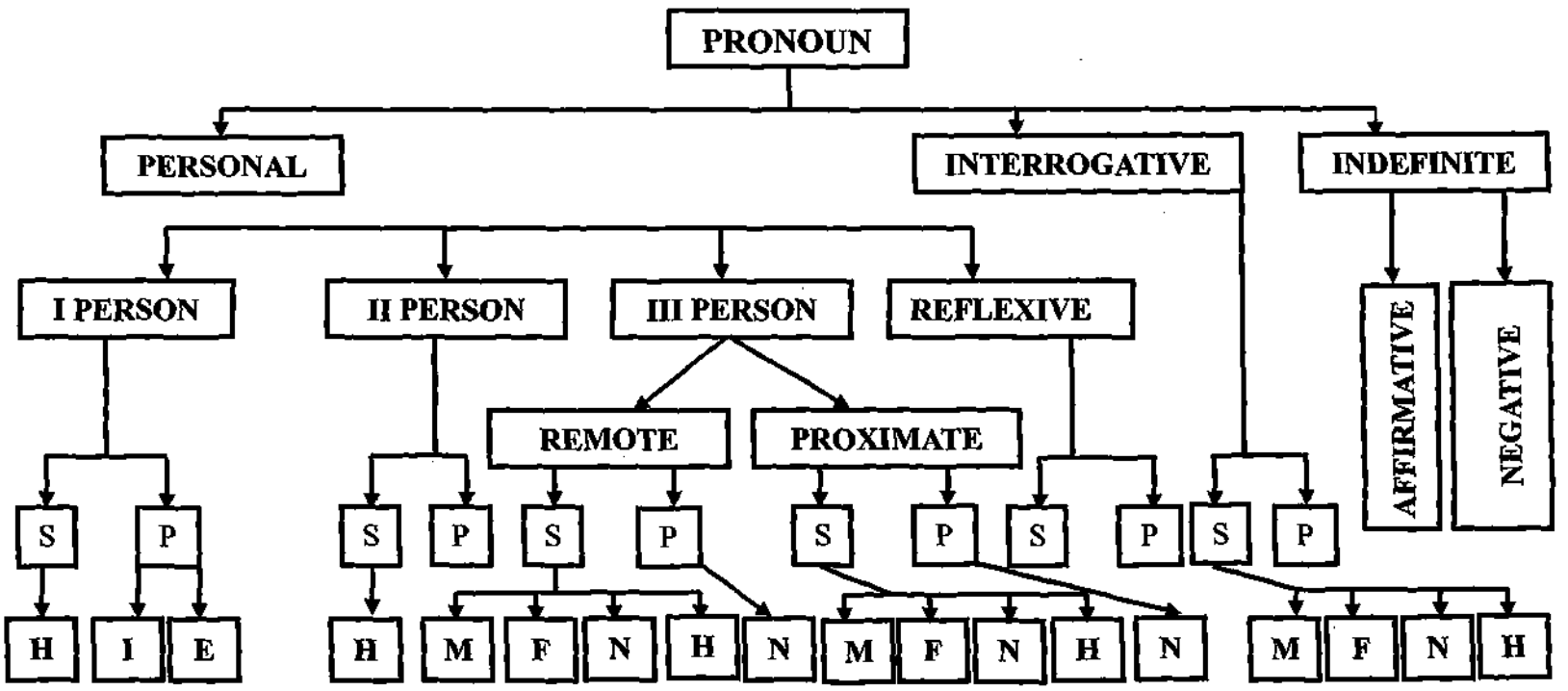


Figure 3.3: Classification of Pronoun

Table 3.2: List of pronouns with description

സർവ്വനാമം	Pronoun	Equalivent English Meaning
ഞാൻ	njaan~	I (personal pronoun, first person, singular)
താൻ	thaan~	One self (reflexive pronoun, remote)
നീ	nii	You (personal pronoun, second person, singular)
നിങ്ങൾ	ningngaL~	You (personal pronoun, second person, plural)
താങ്കൾ	thaankaL~	You (personal pronoun, second person, singular, honorofic)
ഞങ്ങൾ	njangngaL~	We (personal pronoun, first personal pronoun, plural, exclusive)
നമ്മൾ	nammaL~	We (personal pronoun, first person, plural, inclusive)
നാം	naaM~	We (personal pronoun, first person, singular, honorofic)
അവൻ	avan~	He (third person, remote, singular, Masc.)
അവൾ	avaL~	She (third person, remote, singular, Fem.)
അവർ$_1$	avaR~	He/she (third person, singular, Honorofic)
അവർ$_2$	avar	That (third person, remote, plural)
അത്	ath	That (third person, remote, singular, Neuter)
അവ	ava	Those (third person, remote, plural,Neuter)
ഇവൻ	ivan~	He (third person, proximate, singular, Masc.)
ഇവൾ	ivaL~	She (third person, proximate, singular, Fem.)
ഇവർ$_1$	ivaR~	He/she (third person, proximate, singular, honorofic).
ഇവർ$_2$	ivar	They (third person, proximate, plural)
ഇത്	ith	It (Third person, proximate, singular, Neuter)
ഇവ	iva	These (Third person, proximate, plural, Neuter)
തന്നാൽ	thannaal~	One selves (reflexive pronoun, plural)
ഏവൻ	eevan~	Who (interrogative pronoun, Singular, Masc.)
ഏവൾ	eevaL~	Who (interrogative pronoun, Singular, Fem.)
ഏവർ	eevaR~	Who (interrogative pronoun, hon.sing or epicene.pl)
ആരോ	aaroo	Some one(Indefinite pronoun, plural)
ആരും	aaruM~	Some one (Indefinite pronoun, neg)
ഏതോ	eethoo	Some (Indefinite pronoun, non.plural)
ഏതും	eethuM~	Some (Indefinite pronoun, Non. Plural, neg)

Table 3.3: List of third person proximate and remote pronouns

ഈ	ii	This (Third person proximate)
ഇത്ര	ithra	This much (Third person proximate)
ഇങ്ങനെ	ingngne	In this manner (Third person proximate)
ഇപ്പോൾ	ippooL~	Now (Third person proximate)
ഇന്ന്	inn	Today (Third person proximate)
ഇവിടെ	ivite	Here (Third person proximate)
ഇങ്ങ്	ingng	Here (Third person proximate)
ആ	aa	That (Third person remote)
അത്ര	athra	That much (Third person remote)
അങ്ങനെ	angngne	In that manner (Third person remote)
അപ്പോൾ	appooL~	Then (Third person remote)
അന്ന്	ann	That day (Third person remote)
അവിടെ	avite	There (Third person remote)
അങ്ങ്	angng	There (Third person remote)

Table 3.4: List of case suffixation with pronouns.

എന്നെ	enne	about me (first person, accusative)
നമ്മെ/നമ്മളെ	namme/ nammLe	About us (second person, accusative)
എനിക്ക്	enikk	For me (first person, singular, dative)
എന്നിൽ	ennil~	Within me (first person, singular, locative)
നമ്മിൽ/നമ്മുക്ക്	nammil~/ nammukk	Within us (second person, locative / dative)
നമ്മുടെ	nammute	Our (Second person, genitive)
എന്നാൽ	ennaal~	by me (first person , instrumental)
എന്റെ	ente	My (first person, genitive)

നിന്റെ	ninte	Your's (second person, singular, possessive pronoun)
അയാൾ അങ്ങേര് ഇയാൾ, ഇങ്ങേര്	ayaaL~, angngeer/ iyaaL~, ingngeer	That person/This person (third person,masculine)
അദ്ദേഹം/ ഇദ്ദേഹം	addeehaM~/ iddeehaM~	That person/this person (third person,masculine, honorific)

3.1.1.3 Verb Morphology

The morphological structure of verb is complex. Verbs are action words, which are essential for making sense [Kroeger 2005]. They are capable of taking tense markers. In Malayalam, verbs are not inflected for person, number and gender. All verbal forms in Malayalam, both finite and non-finite, consists of verb stems followed by affixes which express various grammatical categories such as tense, aspect, mood, voice, valency change etc.

Generally **tense** are classified into past, present and future. **Aspect** as perfective, imperfective, progressive. **Mood** as indicative, interrogative, imperative, conditional, optative, debitive, potential. Two **voices** are active voice and passive voice. **Valency change** is classified into causative and passive [NairG 2008].

3.1.1.4 Verb Suffixation

In this analysis, identification of the nouns is a major task since more borrowed words fall into the noun category than the verb category. Verbs from English are usually used with *-ceyy-* (-ചെയ്-). For example, *Draiv ceythu* (ഡ്രൈവ് ചെയ്തു), *TTaipp ceyyunnu* (ടൈപ്പ് ചെയ്യുന്നു) etc. Here *Draiv, TTyipp etc.* are borrowed and used by inflecting it with native suffixes. In order to get the phonemic pattern of English words noun suffixes are removed. Verb suffixation can be used to identify the verbs in the corpus. No attempt is made to produce a complete list of all suffixes which make up the verbs. Also the morphophonemic transformation for verbs is beyond the scope of this study. Manually identified list of suffixes affixed to verb stems are shown in Table 3.5.

Table 3.5: List of suffixes attached with verb, with examples

Suffix code	Suffix name	suffixes	suffixes	Examples
VS1	Debitive suffix	/-അണം/, /-ണം/, /-ഏണ്ട-/, /-ഏണ്ടി-/	/-aNaM~/, /-NaM~/, /-eeNta-/, /-eeNti-/	നടക്കണം, പഠിപ്പിക്കേണം, പറയേണ്ട, അലയേണ്ടി natakkaNaM~, paTippikkeeNaM~, paRayeeNta, alayeeNti
VS2	Debitive emphatic suffix	/-അണേ/	/-aNee/	വരണേ varaNee
VS3	Negative debitive suffix	/-അണ്ട/	/-aNta/	പോകണ്ട pookaNta
VS4	Passive suffix	/-അപ്പെട്-/	/-appet-/	കാണപ്പെടുന്ന kaaNappetunnu
VS5	Negative imperative marker	/-അരുത്/	/-aruth/	ചാടരുത് Caataruth
VS6	Verbal noun Suffix	/-അൽ/, /-അവ്/ /-ത്തം/, /-പ്പ്/	/-al~/, /-av/, /-ththaM~/, /-pp/	വിടൽ, വരവ്, നടത്തം, നടപ്പ് vital~, varav, natathThaM~ natapp
VS7	Nominal tense	/-അവ/	/-ava/	വന്നവ vannava
VS8	Optative in verb-mood category	/-അട്ടെ/	/-atte/	ചെയ്യട്ടെ ceyyatte
VS9	Simultaneous suffix	/-അവേ/, /-അവെ/	/-avee/, /-ave/	ഇരിക്കവേ irikkavee
VS10	Negative adverbial participle suffix	/-ആണ്ട്/, /-ആതെ-/	/-aaNt/, /-aathe/	വരാണ്ട്, കഴിയാതെ varaaNt, kazhiyaathe
VS11	Negative marker	/-ആത്-/	/-aath-/	വരാതിരിക്കില്ല varaathirikkilla
VS12	Purposive infinitive suffix	/-ആൻ/, /-ഉവാൻ/	/-aan~/, /-uvaan~/	വരാൻ, വരുവാൻ varaan~, varuvaan~
VS13	May_be suffix	/-ആയിരിക്കാം/, /-ആകാം/	/-aayirikkaaM~/, /-aakaaM~/	അടിച്ചതായിരിക്കാം, പറഞ്ഞതാകാം aticcthaayirikkaaM~, paRanjnjathaakaaM~
VS14	Adjectival suffix	/-ആവുന്ന/	/-aavunn/	പോകാവുന്ന pookaavunna
VS15	Imminent suffix	/-ആറായി/	/-aaRaayi/	വരാറായി varaaRaayi

VS16	Habitual suffix	/-ആറുണ്ട്-/	/-aaRuNt/	വായിക്കാറുണ്ട് vaayikkaaRuNt
VS17	Old condition Suffix	/-ആകിൽ/	/-aakil~/	അങ്ങനെയാകിൽ angnganeyaakil~
VS18	Old negative relative participle suffix	/-ആഞ്ഞ-/, /-ആത്ത-/	/-aanjnja-/, /-aaththa-/	വരാഞ്ഞ, പറയാത്ത varaanjnja, parayaaththa
VS19	Causative suffix	/-ഇക്ക്-/, /-(ഇ)പ്പിക്ക്-/	/-ikk-/, /-(i)ppikk-/	തിറ്റിക്കുക, പറപ്പിക്ക്, പറയിപ്പിക്ക് thiTTikkuka, paRappikk, paRayippikk
VS20	Perfect aspect marker	/-ഇരിക്ക്-/	/-irikk-/	പോയിരിക്കുന്നു pooyirikkunnu
VS21	Imperfective suffix	/-ഉന്നുണ്ട്/,	/-unnuNt/,	വായിക്കുന്നുണ്ട് vaayikkunnuNt
VS22	Positive imperative Suffix	/-ഉവിൻ/, /-വരിൻ/	/-uvin~/, /-varin~/	പോകുവിൻ നടന്നുവരിൻ pookuvin~, natannuvarin~
VS23	Positive indicative future suffix	/-ഊ/	/-uu/	വരൂ varuu
VS24	Self-beneficiary suffix	/-എടുത്തു/	/-etuththu/	നേടിയെടുത്തു neetiyetuththu
VS25	Nominal quotative suffix	/-എന്നത്/	/-ennath/	വരുമെന്നത് varumennath
VS26	Casual consent suffix	/-ഏക്കാം/, /-ഏക്കും/	/-eekkaaM~/, /-eekkuM~/	വന്നെക്കും, വന്നെക്കാം vannekkuM~, vannekkaaM~
VS27	Permission negation suffix	/-കൂടാ/, /-പറ്റില്ല/	/-kuutaa/, /-paTTilla/	വന്നുകൂടാ, ചെയ്യാൻപറ്റില്ല vannukuuta, ceyyaan~paTTilla
VS28	Benifactive suffix	/-കൊടുക്ക്-/	/-kotukk/	ചൊല്ലികൊടുക്കാൻ collikotukkaan~
VS29	Progressive suffix	/-കൊണ്ടിരി-/ /-കൊണ്ടിരിക്കെ/	/-koNtiri-/, /-koNtirikke/	വായിച്ചുകൊണ്ടിരിക്കയാണ്, പഠിച്ചുകൊണ്ടിരിക്കും vaayiccukoNtirikkayaaN, paTiccukoNtirikkuM~
VS30	Infinitive suffix	/-ഉക/	/-uka-/	ചെല്ലുക celluka
VS31	Time marker	/-പോൾ/ /-പ്പോൾ/ /-മ്പോൾ/	/-pooL~/, /-ppooL~/, /-mpooL~/,	വന്നപ്പോൾ, പറഞ്ഞപ്പോൾ, വരുമ്പോൾ,

		/-മ്പോഴേയ്ക്കും/	/-poozheeykkuM~/	വരുമ്പോഴേയ്ക്കും vannappooL~, paRanjnjappool~, varumpooL~ varumppoozheeykkuM~
VS32	Past tense marker	/-ഞ്ഞു/, /-ന്നു/, /-റ്റു/, /-ത്തു/, /-തു/, /-ിു/, /-ച്ചു/, /-ന്തു/, /-ട്ടു/	/-njnju/, /-nnu/, /-TTu/,/-ththu/, /-thu/, /-i/,/-ccu/,/-Ntu/, /-ttu/	പറഞ്ഞു, ഇറന്നു, പറ്റു, കൊടുത്തു, എഴുതി, വിധിച്ചു, കണ്ടു, കടത്തിവിട്ടു paRanjnju, thuRannu, paTTu, kotuththu, ezhuthi, vidhiccu, kaNtu, kataththivittu
VS33	Present tense marker	/-ുന്നു/	/-unnu/	വരുന്നു varunnu
VS34	Relative participle marker	/-ുന്ന/	/-a/	നടന്ന natanna
VS35	Verify marker	/-നോക്ക്-/	/-nookk-/	തുക്കിനോക്കുമെന്നറപ്പ് thuukkinookkumennuRapp
VS36	Conditional verb suffix	/-/ുമെ/	/-eene/	പറഞ്ഞെനെ paRanjnjeene
VS37	Perfective suffix	/-ുട്ട്/	/-itt/	അടഞ്ഞിട്ട് atanjnjitt
VS38	Contingent suffix	/-ക്കാം/	/- kkaaM~/	നടക്കാം natakkaaM~
VS39	Conditional permissive suffix	/-ുാവു/	/-aavuu/	എഴുതാവു ezhuthaavuu
VS40	Permissive suffix	/ -ൊള്ളു/	/-ooLu/	പറഞ്ഞൊള്ളു paRanjnjooLu
VS41	Non formal imperative suffix	/-ക്കൂ/	/-kkuu/	നടക്കൂ natakuu
VS42	Suffix showing beginning	/-ുടങ്ങി /	/thutangngi/	കണ്ടുടങ്ങി kaNtuthutangngi

3.1.1.5 Postposition Suffixation

A postposition is defined as "a particle or word placed after word to indicate its grammatical or syntactical relationship to the other words in the sentence" [Pei et al. 1954]. Instead of prepositions in English language, Malayalam language has got postpositions. Different types of classifications are given for postposition according to their origin, morphological similarities, such as participles, case indicators, co-ordinator, derived noun etc. Another classification listed by Seshagiriprabhu has been taken here for analysis. He defines postpositions as certain words added to the case affixes to indicate special meaning. Since postpositions indicate case relations, they are functionally the same as case affixes [Seshagiriprabhu 1983]. Postpositions will concatenate with verbs and nouns [Mallassery 1994]. Robert Caldwell (1913) commented that every postposition annexed to a noun constitutes, properly speaking, a new case. On the basis of above definition, postpostions are also considered here as markers. After removing the postpositions, the remaining portion decides whether it is noun or a verb.

Eg: (34):	മേശക്കുള്ളിൽ	- മേശക്ക്	+	ഉള്ളിൽ
	meeSakkuLLil~	- meeSakk	+	uLLil~
		noun	+	postposition

The postpositions which are commonly agglutinated with nouns are discussed below from PS1 to PS26. Remaining postpositions are explained in the dual functional suffixation. Postpositions with their occurrence and morphophonemic changes are explained with relevant examples. The morphophonemic change occurring at the boundaries are shown in second column, which helps to derive the morphophonemic rule discussed in section 4.2 for root word identification.

PS1. <u>Duration/Distance suffix /-ooLaaM~/</u>

The suffix /-ഓളം/ (/-ooLaaM~/) meaning upto / approximately / till / as far as, shows the duration/distance suffix. Commonly occurs with nominative case. This postposition is not taken by instrumental case, suffixed with noun.

Eg: (35)

നിയമം + ഓളം > നിയമത്തോളം	ം + ഓ > ത്തോ
niyamaM~ + ooLaM~ > niyamaththooLaM~	aM~ + oo > ththo o

PS2. Equality suffix /-poole/ and /-maathiri/

The suffix /-poole/ (/-പോലെ/- like) and /-maathiri/ (/-മാതിരി/- manner) come under the equality/similarity suffix. /-പോലെ/ can be included in different categories such as postposition, clitics or as equality suffix.

Eg: (36)

വീട് + പോലെ > വീടുപോലെ	[c]/[cc] + പ് > [c]/[cc]ു
viit + poole > viitupoole	[c]/[cc] + p > [c]/[cc]u
അത് + മാതിരി > അതുമാതിരി	[c]/[cc] + മ് > [c]/[cc]ു
ath + maathiri > athumaathiri	[c]/[cc] + m > [c]/[cc]u

PS3. Posterior duration past suffix /-muthal~/ and /-thott/

The suffixes /-mutal~/ (/-മുതൽ/- from) and /-thott/ (/-തൊട്ട്/- starting with) can be used as posterior duration past suffix.

Eg: (37)

നാളെ + മുതൽ > നാളെമുതൽ	എ + മ് > no change
naale + muthal~ > naalemuthal~	e + m > no change
രാവിലെ + തൊട്ട് > രാവിലെതൊട്ട്	എ + ത് > no change
raavile + thott > raavilethott	e + th > no change

PS4. Cause expressing suffix /-muulaM~/, /-kaaraNaM~/, /-nimiththaM~/

Reason or 'because of' can be expressed by using the cause expressing suffixes /-മൂലം/, /-കാരണം/, /-നിമിത്തം/.

Eg: (38):

ക്യാൻസർ + മൂലം > ക്യാൻസറുമൂലം	ർ + മ് > റു
kyaan~saR~ + muulaM~ > kyaan~sRumuulaM~	R~ + m > Ru
പറഞ്ഞത് + കാരണം > പറഞ്ഞതുകാരണം	[c]/[cc] + ക് > [c]/[cc]ു
paRanjnjath + kaaraNaM~ > paRanjnjathukaaraNaM~	[c]/[cc] + k > [c]/[cc]u
അത് + നിമിത്തം > അതുനിമിത്തം	[c]/[cc] + ന് > [c]/[cc]ു
ath + nimiththaM~ > athunimiththaM~	[c]/[cc] + n > [c]/[cc]u

PS5. Exclusion suffix /ozhicc/ and /ozhike/

/-ozhicc/ (/-ഒഴിച്ച്/) and /-ozhike/ (/-ഒഴികെ/) has similar meanings as "except".

Eg: (39)

അവൾ + ഒഴികെ > അവളൊഴികെ	ൾ + ഒ > ളൊ
avaL~ + ozhike > avaLozhike	L~ + o > Lo

PS6. Anterior duration past suffix /-vare/

The suffix used to express anterior duration past is /-vare/ (/-വരെ/) which means "until" or "till".

Eg: (40)

കർപ്പൂരം + വരെ > കർപ്പൂരംവരെ	ം + വ് > no change
kaR~ppuuraM~+ vara > kaR~ppuuraM~vara	aM~ + v > no change

PS7. 'Through out' suffix /-ottaake/ and /-ottukk/

/-ottukk/ (/-ഒട്ടുക്ക്/) and /-ottaake/ (/-ഒട്ടാകെ/) means "completely" or "wholly" or "through out". Usually comes with nominative case.

Eg: (41)

നാട് + ഒട്ടുക്ക് > നാടൊട്ടുക്ക്	[c]/[cc] + ഒ > [c]/[cc]ൊ
naat + ottukk > naatottukk	[c]/[cc] + o > [c]/[cc]o
നാട് + ഒട്ടാകെ > നാടൊട്ടാകെ	[c]/[cc] + ഒ > [c]/[cc]ൊ
naat + ottaake > naatottaake	[c]/[cc] + o > [c]/[cc]o

PS8. Referencing suffix /-kuRicc/, /-paTTi/, /-saM~bandhicc/

Referencing suffixes /കുറിച്ച്/, /പറ്റി/ and /സംബന്ധിച്ച്/ are used to say something "about".

Eg: (42)

അവനെ + കുറിച്ച് > അവനെകുറിച്ച്	എ + ക് > no change
avane + kuuRicc > avanekuuRicc	e + k > no change
പരിക്ക് + പറ്റി > പരിക്കുപറ്റി	[c]/[cc] + പ് > [c]/[cc]ു
parikk + paTTi > parikkupaTTi	[c]/[cc] + p > [c]/[cc]u
അവനെ + സംബന്ധിച്ച് > അവനെസംബന്ധിച്ച്	എ + സ് > no change
avane + saM~bandhicc > avanesaM~bandhic	e + s > no change

PS9. 'Not even' suffix /-ott/

This suffix, /-ഒട്ട്/ shows a meaning related with the concept of "not even a little".

Eg : (43)

അവൾ + ഒട്ട് > അവളൊട്ട്	ൾ + ഒ > ളൊ
avaL~ + ott > avaLott	L~ + o > Lo

PS10. 'Instead of' suffix /pakaraM~/

This postposition /-പകരം/ gives the meaning of "instead of":

Eg: (44)

അവന് + പകരം > അവനുപകരം	[c]/[cc] + പ് > [c]/[cc]ു
avan + pakaraM~> avanupakaraM~	[c]/[cc] + p > [c]/[cc]u

PS11. Superior contact location suffix /-meel~/, /-meele/, /-miithe/, /-mukaLil~/

The suffixes /-മേൽ/, /-മേലെ/ (on) and /-മീതെ/, /-മുകളിൽ/ (above) gives the meaning of superior contact location.

Eg: (45)

തലക്ക് + മീതെ > തലക്കുമീതെ	[c]/[cc] + മ് > [c]/[cc]ു
thalakk + miite > thalakkumiithe	[c]/[cc] + m > [c]/[cc]u
മണിക്ക് + മേൽ/മേലെ >മണിക്കുമേൽ/മണിക്കുമേലെ	[c]/[cc] + മ് > [c]/[cc]ു
maNikk + meel~/meele> maNikkumeel~/maNikkumeele	[c]/[cc] + m > [c]/[cc]u

PS12. Inferior contact location suffix /-thaazhe/, /-kiizhe/, /-kiizhil~/, /-atiyil~/

Inferior contact location suffixes /-താഴെ/, /-കീഴെ/,/-കീഴിൽ/, /-അടിയിൽ/ gives the meaning "under" or "below".

Eg: (46)

അവന് + താഴെ > അവനുതാഴെ	[c]/[cc] + ത് > [c/cc]ു
avan + thaazhe > avanuthaazhe	[c]/[cc] + th > [c]/[cc]u
കട്ടിലിന് + കീഴെ > കട്ടിലിനുകീഴെ	[c]/[cc] + ക് > [c]/[cc]ു
kattilin + kiizhe > kattilinukiizhe	[c]/[cc] + k > [c]/[cc]u
കട + കീഴിൽ > കടക്കീഴിൽ	അ + ക് > ക്ക
kuta + kiizhil~ > kutakkiizhil~	a + k > kka

PS13. <u>'After' Suffix /SeeshaM~/</u>

/SeeshaM~/ (/-ശേഷം/) means "after" or "remainder"

Eg: (47)

ഒട്ടിച്ചതിന് + ശേഷം > ഒട്ടിച്ചതിനുശേഷം	[c]/[cc] + ശ് > [c]/[cc]ു
otticcthin + SeeshaM~ > otticcthinuSeehaM~	[c]/[cc] + S > [c]/[cc]u

PS14. <u>Benefactive suffix /aayi/, /veeNti/</u>

Benefactive suffix /-ആയി/ and /-വേണ്ടി/ gives the meaning "for"

Eg: (48)

അവൾക്ക് + ആയി > അവൾക്കായി	[c]/[cc] + ആ > [c]/[cc]ാ
avaL~kk + aayi > avaL~kkaayi	[c]/[cc] + aa > [c]/[cc]aa
കക്ഷിക്ക് + വേണ്ടി > കക്ഷിക്കുവേണ്ടി	[c]/[cc] + വ് > [c]/[cc]ു
kaxikk + veeNti > kaxikkuveeNti	[c]/[cc] + v > [c]/[cc]u

PS15. <u>'Towards' suffix /-neere/</u>

This suffix /-നേരെ/ possess the meaning "towards", "at" etc.

Eg: (49)

കുട്ടിക്ക് + നേരെ > കുട്ടിക്കുനേരെ	[c]/[cc] + ന് > [c]/[cc]ു
kuttikk + neere > kuttikkuneere	[c]/[cc] + n > [c]/[cc]u

PS16. <u>'From' suffix /-ninn/</u>

/-ninn/ (/-നിന്ന്/) denotes "from".

Eg: (50)

കൂടെ + നിന്ന് > കൂടെനിന്ന്	എ + ന് > no change
kuute + ninn > kuuteninn	e + n > no change

PS17. <u>Possession suffix /-pakkal~/</u>

/-പക്കൽ/ gives the meaning "with".

Eg: (51)

അവളുടെ + പക്കൽ > അവളുടെപക്കൽ	എ + പ് > no change
avaLute + pakkal~ > avaLutepakkal~	e + p > no change

PS18. <u>'In front of' suffix /-munnil~/,/-mumpil~/,/-mumpaake/</u>

"Infront of" or "before" is the meaning for the suffixes /-മുന്നിൽ/, /-മുമ്പിൽ/ and /-മുമ്പാകെ/.

Eg: (52)

കോടതി + മുമ്പാകെ >കോടതിമുമ്പാകെ kootathi + mumpaake > kootathimumpaake	ഇ + മ് > no change i + m > no change
അവൾക്ക് + മുമ്പിൽ > അവൾക്കുമുമ്പിൽ avaL~kk + mumpil~ > avaL~kkumumpil~	[c]/[cc] + മ് > [c]/[cc]ു [c]/[cc] + m > [c]/[cc]u
അനുഭവങ്ങൾക്ക് + മുന്നിൽ >അനുഭവങ്ങൾക്കുമുന്നിൽ anubhavangngL~kk + munnil~> anubhavangngL~kkumunnil~	[c]/[cc] + മ് > [c]/[cc]ു [c]/[cc] + m > [c]/[cc]u

PS19. <u>'Within' suffix /-uLLil~/, /-akaththh/ and /-akaM~/</u>

/-ഉള്ളിൽ/, /-അകത്ത്/ and /-അകം/ gives the meaning "within". They express spacial meanings.

Eg: (53)

മേശക്ക് + ഉള്ളിൽ > മേശക്കുള്ളിൽ	[c]/[cc] + ഉ > [c]/[cc]ു
meeSakk + uLLil~ > meeSakkuLLil~	[c]/[cc] + u > [c]/[cc]u
പൂരിക്ക് + അകത്ത് > പൂരിക്കകത്ത്	[c]/[cc] + അ > [c]/[cc]a
puurikk + akaththh > puurikkakaththh	[c]/[cc] + a > [c]/[cc]a
മണിക്ക് + അകം > മണിക്കകം	[c]/[cc] + അ > [c]/[cc]a
maNikk + akaM~ > maNikkakaM~	[c]/[cc] + a > [c]/[cc]a

PS20. Comparative suffix /-kaaL~/

The suffix /-കാൾ/ gives the meaning of "'than". /-kaaLuM~/ (/-കാളും/) and /-kaattiluM~/ (/-കാട്ടിലും/) shows similar meaning.

Eg: (54)

അവളെ + കാൾ > അവളെക്കാൾ	എ + ക് > ക്ക
avaLe + kaaL~ > avaLekkaaL~	e + k > kka

PS21. 'Without' suffix /-kuutaathe/

As the name suggest, 'without' suffix (/-കൂടാതെ/) gives the meaning "without".

Eg: (55)

അമ്മയെ+ കൂടാതെ > അമ്മയെകൂടാതെ	എ + ക് > no change
ammaye + kuutaathe > ammayekuutaathe	e + k > no change

PS22. 'On' suffix /-puRaththu/

'On' suffix /-പുറത്ത്/, meaning "outside".

Eg: (56)

വീട്ടിന് + പുറത്ത് > വീട്ടിനപ്പുറത്ത്	[c]/[cc] + പ് > [c]/[cc]ുപ
viittin + puRaththu > viittinuppuRaththu	[c]/[cc] + p > [c]/[cc]u (p)

PS23. 'Around' suffix /-cuTTuM~/

Around suffix (/-ചുറ്റും/) meaning "around".

Eg: (57)

വീട്ടിന് + ചുറ്റും > വീട്ടിനചുറ്റും	[c]/[cc] + ച് > [c]/[cc]ു
viittin + cuTTuM~ > viittinucuTTuM~	[c]/[cc] + c > [c]/[cc]u

PS24. Punctual past suffix /-mun~p/ and /-mump/

The punctual past suffix (/-മുൻപ്/ and /-മുമ്പ്/) means "before".

Eg: (58)

ഞായറാഴ്ചക്ക് + മുൻപ് > ഞായറാഴ്ചക്കുമുൻപ്	[c]/[cc] + മ് > [c]/[cc]ു
njaayaRaazhccakk + mun~p> njaayaRaazhccakkumun~p	[c]/[cc] + m > [c]/[cc]u

PS25. 'In between' suffix /-itayil~/

'In between' suffix is /-itayil~/ (/-ഇടയിൽ/) which means "between".

Eg: (59)

ചെടികളുടെ + ഇടയിൽ > ചെടികളുടെയിടയിൽ	എ + ഇ > യി
cetikaLute + itayil~ > cetikaLuteyitayil~	e + i > yi

PS26. 'Behind' suffix /-pinnil~/

Behind suffix is /-പിന്നിൽ/. The meaning is "behind".

Eg: (60)

നേതാവിന് + പിന്നിൽ > നേതാവിനുപിന്നിൽ	[c]/[cc] + പ് > [c]/[cc]u
neethaavin + pinnil~ > neethaavinupinnil~	[c]/[cc] + p > [c]/[cc]u

There are certain other postpostional suffixes such as 'Through' suffix, cause expressing / instrumental suffix, frequentative suffix and 'using' suffix. These postpositions agglutinate with noun and verb forms. So they have been classified as a sub category of dual functional suffix which is discussed from PDFS1 to PDFS4 in section 3. 1. 2. 2.

3.1.2 Dual Functional Suffixation

Dual functional suffixes are those suffixes which are agglutinated with nouns, verbs, noun phrases, verb phrases, postpositional phrases etc. The different steps involved in identifying the root word using dual functional suffixes can be explained as:

Eg: (61) *thiirumaanikkaaththathaaN* (തീരുമാനിക്കാത്തതാണ്)

Analysing the word from right to left, the first suffix /-aaN/ (/-ആണ്/) is obtained. It is a dual functional suffix. So the remaining portion can be its nominal or verbal form. On further analysis the system identifies the next suffix /-ath/ (/-അത്/) as a dual functional suffix. Eliminating this suffix, the word becomes *thiirumaanikkaaththa* (തീരുമാനിക്കാത്ത). Now the suffix is /-aaththa/ (/-ആത്ത/). This is a verb suffix. After removing this suffix, the root word is identified. Since there is no further suffixes it can be concluded that root word is *thiirumaanikk* (തീരുമാനിക്ക്) and it is a verb having causative suffix.

Step 1:	തീരുമാനിക്കാത്തതാണ്	-	തീരുമാനിക്കാത്തത്	+	ആണ്	
	thiirumaanikkaaththathaaN	-	thiirumaanikkaaththath	+	aaN	
					dual functional suffix	
Step 2:	തീരുമാനിക്കാത്ത	+	അത്	+	ആണ്	
	thiirumaanikkaaththa	+	ath	+	aaN	
			dual functional suffix +		dual functional suffix	
Step 3:	തീരുമാനിക്ക്	+	ആത്ത	+	അത്	+ ആണ്
	thiirumaanikk	+	aaththa	+	ath	+ aaN
	verb		verb suffix +		dual functional suffix +	
					dual functional suffix	

3.1.2.1 Dual Functional Suffixes

DFS1. Existential/Existential copula suffix /-uNt/

/-ഉണ്ട് / (have), the existential copula, can be used to denote the existence of something already established. This suffix can also be used as a being verb. Usually this will come with positive sentences.

Eg: (62)

പദ്ധതി + ഉണ്ട് > പദ്ധതിയുണ്ട്	ഇ + ഉ > യു
paddhathi + uNt > paddhathiyuNt	i + u > yu
എഴുതിയിട്ട് + ഉണ്ട് > എഴുതിയിട്ടുണ്ട്	[c]/[cc] + ഉ > [c]/[cc]ു
ezhuthiyitt + uNt > ezhuthiyittuNt	[c]/[cc] + u > [c]/[cc]u

DFS2. Negative/negative verbal suffix /-alla/ and /-illa/

/-അല്ല/, /-ഇല്ല/ (not that, not having) act as negative verbs correspond to the copulas /-aaN/ and /-uNt/ which usually come with negative sentences. Both occur in tag questions. /-illa/ is also used to deny the existence of something and /-alla/ shows the denial of the attachment of a given quality to an entity. For the negative future form, /-illa/ is added to the infinitive or to the verb stem.

Eg: (63)

കുട്ടി + അല്ല > കുട്ടിയല്ല	ഇ + അ > യ
kutti + alla > kuttiyalla	i + a > ya
നടന്ന് + അല്ല > നടന്നല്ല	[c]/[cc] + അ > [c]/[cc]a
natann + alla > natannalla	[c]/[cc] + a > [c]/[cc]a
കുട്ടി + ഇല്ല > കുട്ടിയില്ല	ഇ + ഇ > യ
kutti + illa > kuttiyilla	i + i > ya
നടന്ന് + ഇല്ല > നടന്നില്ല	[c]/[cc] + ഇ > [c]/[cc]a
natann + illa > natannilla	[c]/[cc] + i > [c]/[cc]a

DFS3. Instrumental case/conditional suffix /-aal~/

/-aal~/ (/-ആൽ/) is considered as the instrumental (*prayojika*) case marker. It is used to show instruments with which action is done or parts of the body which are used in doing the action. In Malayalam /-aal~/ can be used to show the components with which things are made. In passive voice, the doer of the action and in causative sentence, the medial agent takes the instrumental marker [Geethakumary 2002]. This suffix also comes with past tense stem.

Eg: (64)

രാമൻ + ആൽ > രാമനാൽ raaman~ + aal~ > raamanaal~	ൻ + ആ > നാ n~ + aa > naa
വന്ന് + ആൽ > വന്നാൽ vann + aal~ > vannaal~	[c]/[cc] + ആ > [c]/[cc]ാ [c]/[cc] + aa > [c]/[cc]aa

DFS4. And co-ordination/Future model suffix /-uM~/

/-uM~/ (/-ഉം/) can be used to co-ordinate the words. It is added with nominal bases, noun suffixes, verb forms etc.

Eg: (65)

വീട് + ഉം > വീടും viit + uM~ > viituM~	[c]/[cc] + ഉ > [c]/[cc]ു [c]/[cc] + u > [c]/[cc]u
വന്ന് + ഉം > വന്നും vann + uM~ > vannuM~	[c]/[cc] + ഉ > [c]/[cc]ു [c]/[cc] + u > [c]/[cc]u

DFS5. Or co-ordination/Interrogative suffix /-oo/

/-oo/ (/-ഓ/) suffix can be co-ordinated with nouns, adjectives and adverbs. /-oo/ as an interrogative participle can be added to a finite verb.

Eg: (66)

മാധവൻ + ഓ > മാധവനോ maadhavan~ + oo > maadhavanoo	ൻ + ഓ > നോ n~ + oo > noo
പാടാൻ + ഓ > പാടാനോ paataan~ + oo > paataanoo	ൻ + ഓ > നോ n~ + oo > noo

DFS6. Quotative participle suffix /enn/

The reported speech is usually marked by the quotative participle /-enn/ (/-എന്ന്/), which will follow the string (nominal and verbal) representing the reported utterance.

Eg:(67)

സുന്ദരൻ + എന്ന് > സുന്ദരനെന്ന്	ൻ + എ > നെ
sundran~ + enn > sundaranenn	n~ + e > ne
എഴുതി + എന്ന് > എഴുതിയെന്ന്	ഇ + എ > യെ
ezhuthi + enn > ezhuthiyenn	i + e > ye

DFS7. Adjectival/ be-relative participle suffix /-uLLa/

/-ഉള്ള/ means 'having'. Commonly seen with existential and locative sentences. /-uLLa/, the relative participle of /-uNt/ (/-ഉണ്ട്/) is used to adjectivalise a number of grammatical forms such as adverbs, infinitives, noun bases with different cases etc.

Eg: (68)

പഠിത്തം + ഉള്ള > പഠിത്തമുള്ള	ം + ഉ > മു
paTiththaM~ + uLLa > paTiththmuLLa	aM~ + u > mu
എഴുതിട്ട് + ഉള്ള > എഴുതിയിട്ടുള്ള	[c]/[cc] + ഉ > [c]/[cc]ു
ezhuthitt + uLLa > ezhuthiyittuLLa	[c]/[cc] + u > [c]/[cc]u

DFS8. Equation/ Finite verb suffix /-aaN/

/-aaN/ (/-ആണ്/) the equative copula, can act as a finite verb form which can be added to major constituents of the sentence such as noun phrase, adverb phrase and postpositional phrase. It carries a certain degree of emphasis.

Eg: (69)

സിനിമയെന്ത് + ആണ് > സിനിമയെന്താണ്	[c]/[cc] + ആ > [c]/[cc]ാ
sinimayenth + aaN > sinimayenthaaN	[c]/[cc] + aa > [c]/[cc]aa
പറഞ്ഞ് + ആണ് > പറഞ്ഞാണ്	[c]/[cc] + ആ > [c]/[cc]ാ
paRanjnj + aaN > paRanjnjaaN	[c]/[cc] + aa > [c]/[cc]aa

DFS9. Hearsay suffix /-pooluM~/

/-pooluM~/ (/പോലും/) means "even". It is used to indicate certain lack of authority for an assertion like /-atree/ (/-അത്രേ/).

Eg: (70)

കുട്ടികാർ + പോലും - കുട്ടികാരുപോലും kuuttukkaR~ + pooluM~ - kuuttukaarupooluM~	ർ + പ് > രു R~ + p > ru
പറയാൻ + പോലും - പറയാൻപോലും paRayaan~ + pooluM~ - paRayaan~pooluM~	ൻ + പ് > no change n~ + p > no change

DFS10. Predicted future time suffix /-uLLuu/

/-uLLuu/ (/-ഉള്ള്/) is a marker used with past tense and present tense verb forms. The suffix also shows the existence of an element.

Eg: (71)

അനുവാദം + ഉള്ള് > അനുവാദമുള്ള് anuvaadaM~ + uLLuu > anuvaadamuLLuu	ം + ഉ > മു aM~ + u > mu
വഴങ്ങുക + ഉള്ള് > വഴങ്ങുകയുള്ള് vazhangnguka + uLLuu > vazhangngukayuLLuu	ക + ഉ > യു ka + u > yu

DFS11. Existential interrogative participle/ Negative interrogative participle suffix /-illee/, /-allee/

A warning may be given by the use of an imperative followed by /-allee/. /-അല്ലേ/ and /-ഇല്ലേ/ suffixes are usually used in circumstances where the expecting answer is 'yes'.

Eg: (72)

മണ്ണ് + ഇല്ലേ > മണ്ണില്ലേ maNN + illee > maNNillee	[c]/[cc] + ഇ > ഇ [c]/[cc] + i > [c]/[cc]i
ഓട് + അല്ലേ > ഓടല്ലേ oot + allee > ootallee	[c]/[cc] + അ > no change [c]/[cc] + a > no change

DFS12. Ordinal/Future model suffix (/-aaM~/)

Ordinals are formed in Malayalam by adding the suffix /-aaM~/ (/-ആം/) or /-aamatte/ (/-ആമത്തെ/) to cardinal numbers. Suffix /-aaM~/ can also be considered as a future model suffix. This is also added to a verbal stem which conveys permission.

Eg: (73)

ഒന്ന് + ആം > ഒന്നാം	[c]/[cc] + ആ > [c]/[cc]ാാ
onn + aaM~ > onnaaM~	[c]/[cc] + aa > [c]/[cc]aa
വര് + ആം > വരാം	[c]/[cc] + ആ > [c]/[cc]ാാ
var + aaM~ > varaaM~	[c]/[cc] + aa > [c]/[cc]aa

DFS13. <u>Pronomalization suffix /-ath/</u>

In pronoun category, /-ath/ (അത്) is third person singular neuter. According to the context, pronomalized form goes with verb and nouns take pronoun.

കുട്ടി + അത് > കുട്ടിയത്	ഇ + അ > യ
kutti + ath > kuttiyath	i + a > ya
പറഞ്ഞ് + അത് > പറഞ്ഞത്	[c]/[cc] + അ > [c]/[cc]a
paRanjnj + ath > paRanjnjath	[c]/[cc] + a > [c]/[cc]a

3.1.2.2 Postpositional Suffixes

Certain postpositions will act as dual functional suffixes. They are listed below.

PDFS1. <u>'Through' suffix /-vazhi/ /-maaR~ggaM~/</u>

/-വഴി/ and /-മാർഗ്ഗം/ represents the meaning "through".

Eg: (74)

തൊണ്ട + വഴി > തൊണ്ടവഴി	അ + വ് > no change
thoNta + vazhi > thoNtavazhi	a + v > no change
വന്ന + വഴി > വന്നവഴി	അ + വ് > no change
vanna + vazhi > vannavazhi	a + v > no change

PDFS2. <u>Cause expressing/ Instrumental suffix /-koNt/</u>

/-koNt/ (/-കൊണ്ട്/) (with/by/due to/within). This is a derived form of the verb koL- (കൊൾ-) which carries the meaning "'to hold", "to experience", "to come in contact" etc. It comes with noun and past participle forms.

Eg: (75)

അവളെ + കൊണ്ട് > അവളെക്കൊണ്ട്	എ + ക് > ക്ക്
avaLe + koNt > avaLekkoNt	e + k > kk
ഓടി + കൊണ്ട് > ഓടിക്കൊണ്ട്	ഇ + ക് > ക്ക്
ooti + koNt > ootikoNt	i + k > kk

PDFS3. Frequentative suffix /-thooRuM~/

/-തോറും/ which means "each" or "every". It comes after nominative case and future participle forms.

Eg: (76)

ആണ്ട് + തോറും > ആണ്ടുതോറും aaNt + thooRuM~ > aaNtuthooRuM~	[c][cc] + ത് > [c][cc]ു [c]/[cc] + th > [c]/[cc]u
കാണും + തോറും > കാണുംതോറും kaaNuM~ + thooRuM~ > kaaNuM~thooRuM~	ും + ത് > no change aM~ + th > no change

PDFS4. 'Using' suffix /-vecc/

/-vecc/ (/-വെച്ച്/) means at/using/among/in/placing.

Eg: (77)

മകനെ + വെച്ച് > മകനെവെച്ച് makane + vecc > makanevecc	എ + വ് > no change e + v > no change
മറച്ച് + വെച്ച് > മറച്ചുവെച്ച് maRacc + vecc > maRaccuvecc	[c]/[cc] + വ് > [c]/[cc]ു [c]/[cc] + v > [c]/[cc]u

3.2 Morphophonemic Changes

"സംഹിതയാൽ ഉണ്ടാകുന്ന വർണ്ണങ്ങളുടെ ഉച്ചാരണഭേദം ആകുന്നു സന്ധി." [വ്യാകരണമിത്രം]

"saM~hithayaal~ uNtaakunna vaR~NNangngaLute uccaaraNabhedaM~ aakunnu sandhi" [vyaakaraNamithraM~]

"(വർണ്ണ)യോഗജന്യം വികാരം സന്ധി"

[കേരളപാണിനീയം 1071 M.E]

"(vaR~NNa)yoogajanyaM~ vikaaraM~ sandhi"

[keeraLapaaNiniiyaM~ 1071 M.E]

Sandhi rules are phonological alternations that are triggered at junctures at junctions of words or morphemes. The name sandhi comes from Sanskrit and means "juncture". It doesn't designate one particular phonological process. Instead, it is a non-specific cover term for any kind of sound mutation that occurs at the edges of words and morphemes, and that is triggered in environments created by morphological or syntactic concatenation operation [Philip 2007].

Malayalam grammar has categorised sandhi rules into different types [Seshagiriprabhu 1983] [Varma 1999]. According to its consonant-vowel pair based categorisation, there are *svara sandhi* (svaraM~ + svaraM~), *svara vyanjjana sandhi* (svaraM~ + vyanjjanam), *vyanjjana svara sandhi* (vyanjjanam + svaraM~), *vyanjjana sandhi* (vyanjjanaM~ + vyanjjanaM~). Here *svaram* is vowel and *vyanjjanaM~* is the consonant. The morphophonemic changes occurring in the end phonemes of the word and the initial phoneme of the suffix is used to identify root.

The noun suffixes, verb suffixes, postpositional suffixes and dual functional suffixes identified can be categorized and grouped according to their initial phoneme. Table 3.6 list the groups with their corresponding suffixes. For example, 'Group a' will list the suffixes starting with *a* (അ). Here 'c' denotes the consonants and 'cc' denotes consonant clusters.

Table 3.6: List of suffix groups

Index	Suffix group	Suffixes
1	Group a	/-athth/,/-allathe/,/-allaathth/,/-akaththl/,/-akaM~/, /-alla/, /-allee/ /-aNee/, /-appet-/, /-aruth-/, /-ath/, /-atte/, /-ava/, /-avee/, /-ave/, /-atiyil~/, /-aNaM~/,/-al~/, /-av/, /-a/, /-akaththl/ , /-ava/
2	Group aa	/-aakaaM~/, /-aakil~/, /-aan~/, /-aanjnja-/, /-aaNt/, /-aaRaayi/, /-aaRuNt/, /-aath/, /-aathe/, /-aaththa-/, /-aakatte/, /-aaya/,/-aayi/, /-aaL~/, /-aaN/, /-aavunn/, /-aavuu/ , /-aayirikkaaM~/, /-aaM~/,/-aar/, /-aamatte/
3	Group i	/-i/, /-ikk-/, /-irikk/, , /-itayil~/, /-itt/, /-(i)ppikk/, /-il~/, /-ileekk/, /-iluute/, /-il~ninnu/, /-illaathe/, /-illaaththa/, /-illa/, /-illee/
4	Group u	/-uka-/, /-ukayaaN/, /-uLLil~/, /-unnu/, /-unnuNt/, /-uvin~/, /-uvaan~/, /-ute/, /-uNt/, /uM~/, /ulla/, /-ullu/
5	Group e	/-e/, /-etuththu/, /-ennath/, /-ethire/, /-enth/, /-enthin/,/-ethire/, /-engng/,/-evite/,/-eppooL~/, /-engngne/,/-ethra/, /-enn/, /-ennath/, /-enthee/,/-ethramaathraM~/
6	Group ee	/-eekkaaM~/, /-eekkuM~/, /-eene/, /-eeNta/, /-eeNti/, /-eeth/
7	Group o	/ozhicc/, /ozhike/, /ottukk/, /-ot/, /-oththl/,/-oppaM~/, /-ott/, /-ottaake/
8	Group oo	/-oolaaM~/, /-oo/, /-ooLu/, /-oot/
9	Group cha	/-U/

10	Group c	/-kaaL~/, /-kaaraNaM~/, /-kiizhil~/, /-koNt/, /-koNtiri-/, /-kal~/, /kaL~/, /-kuute/, /-kuuti/, /-koNtirikke/,/-kootaathe/,/-kotukk/, /-kuRicc/, /-kuutaa/, /-keezhe/ /-cu/, /-cuTTuM~/ /-maaRggaM~/, /-meel~/, /-meele/, /-meethe/, /-muulaM~/, /-mpooL~/, /-mump/, /-mumpaake/, /-mumpil~/, /-mun~p/ , /-munnil~/, /-muthal~/ ,/-mukaLil~/, /-maathiri/, /-maaR~/, /mathi/, /-maaR~ggaM~/ /-neere/, /-nimiththaM~/, /-ninn/, /-njnju/, /-nookk/, /-Ntu/, /-NaM~/, /-nnu/, /-neeR~kk/ /-pakkal~/, /-paTTilla/, /-pinnil~/, /-pooL~/, /-poole/, /-poozhekkuM~/, /peer/,/-peeR~/, /-pp/, /-ppooL~/, /-puRaththi/, /pakaraM~/, /-paTTi/, /-pooluM~/ /-thaazhe/, /-thooRuM~/, /-thott/, /-ththaM~/, /-ththu/, /-thu/, /-ttu/, /-tammil/, /-thanne/, /thutangngi/ ,/-ththaM~/ /-vare/, /-varin~/, /-vazhi/, /-vecc/, /veeNti/ /-saM~bandhicc/ /SeeshaM~/
11	Group cc	/-TTu/, /- kkaaM~/, /-kk/, /ngngL~/, /-ccu/, /-kkuu/
12	Group nte	/-nte/

The morphophonemic changes, discussed in 3.1.1.2 and 3.1.1.5, occuring when these suffixes (Table 3.6) are concatenated with root endings phonemes (Table 3.7) are diagrammatically shown in Figure 3.4(a) to Figure 3.4 (i). In these hierarchical diagrams, zero level node represents the end phoneme of the root word, first level nodes represent the suffix groups and second level nodes represent the morphophonemic changes. Verb suffixes are exempted from analysis.

Table 3.7: List of root ending phonemes

n~	N~	L~	R~	l~
a	aa	i	ii	oo
e	ee	u	uu	ai
[c]	[cc]	aM~		

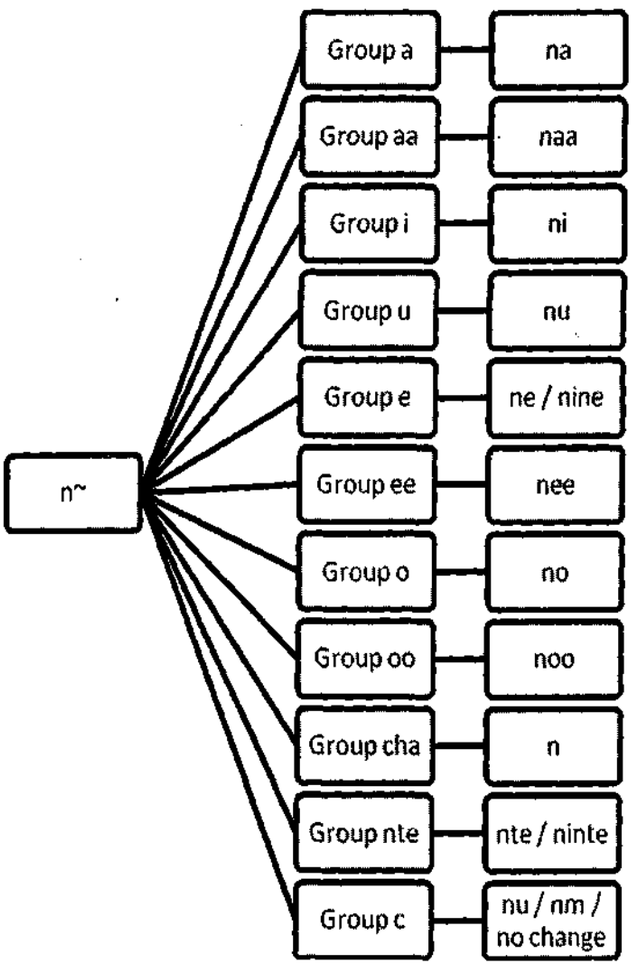


Figure 3.4 (a): Pictorial representation of morphophonemic change between the root ending phoneme n~ (ൻ) and suffix groups

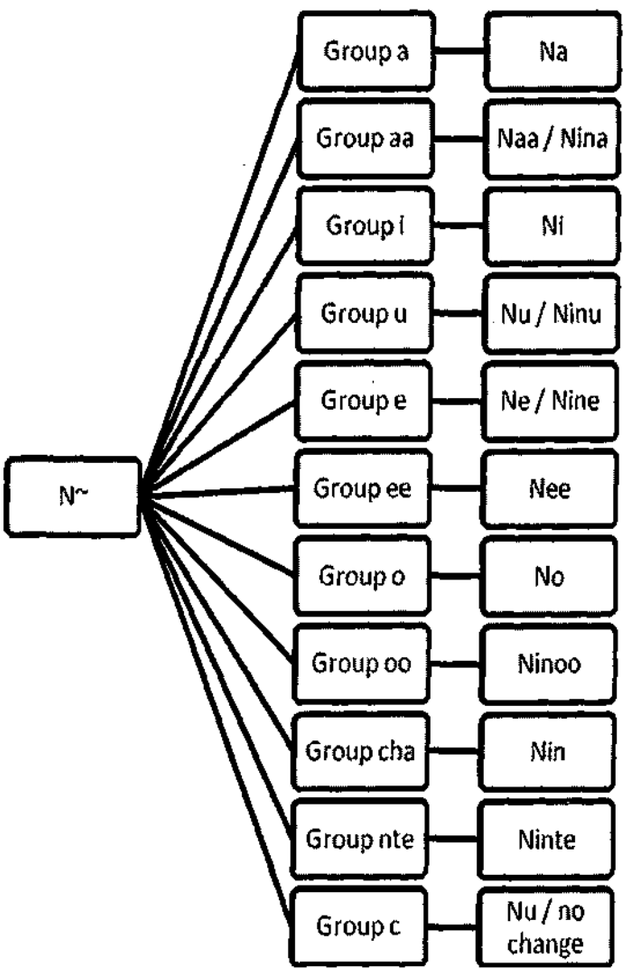


Figure 3.4(b): Pictorial representation of morphophonemic change between the root ending phoneme N~ (ൻ) and suffix groups

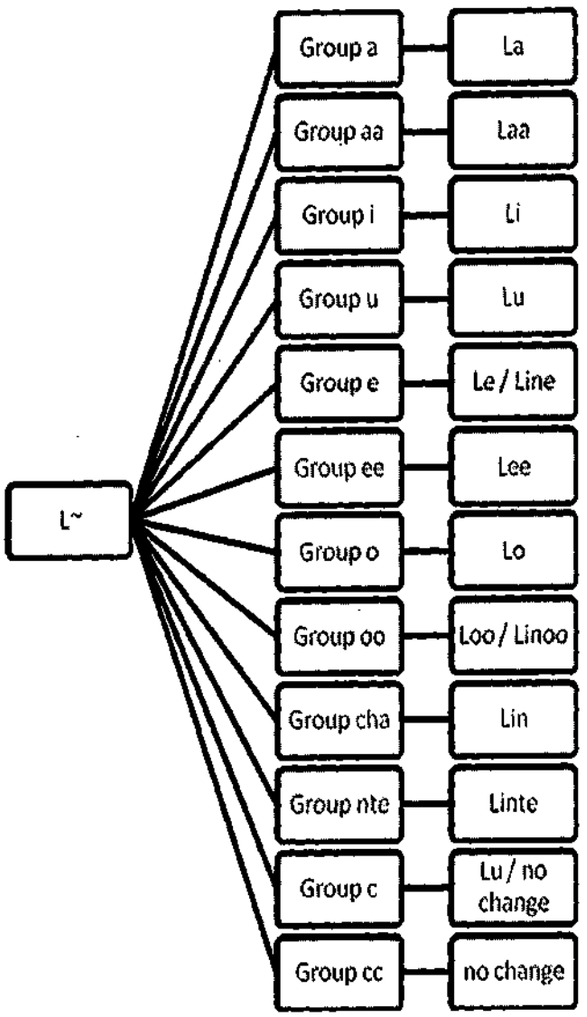


Figure 3.4 (c): 'Pictorial representation of morphophonemic change between the root ending phoneme L~ (ω̃) and suffix groups

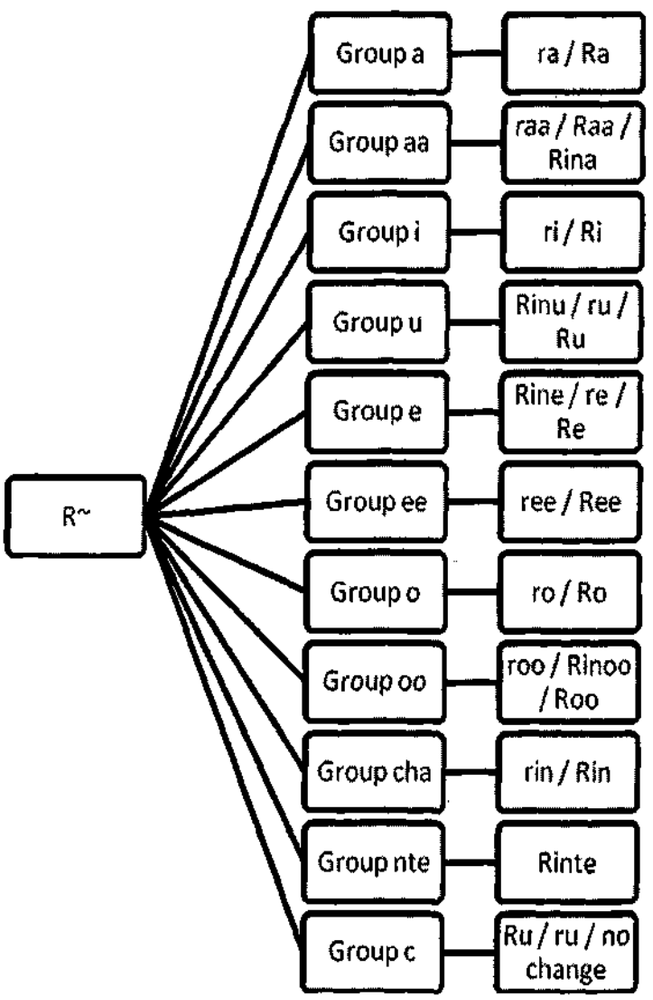


Figure 3.4(d): Pictorial representation of morphophonemic change between the root ending phoneme R~ (ಽ) and suffix groups

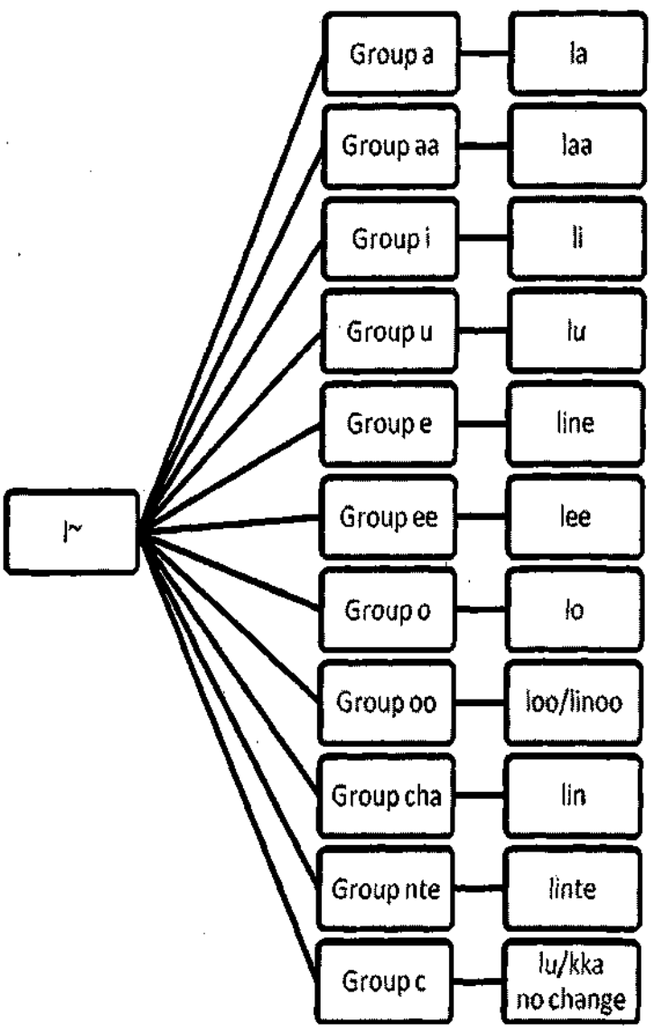


Figure 3.4 (e): Pictorial representation of morphophonemic change between the root ending phoneme ಱ್ (ൾ) and suffix groups

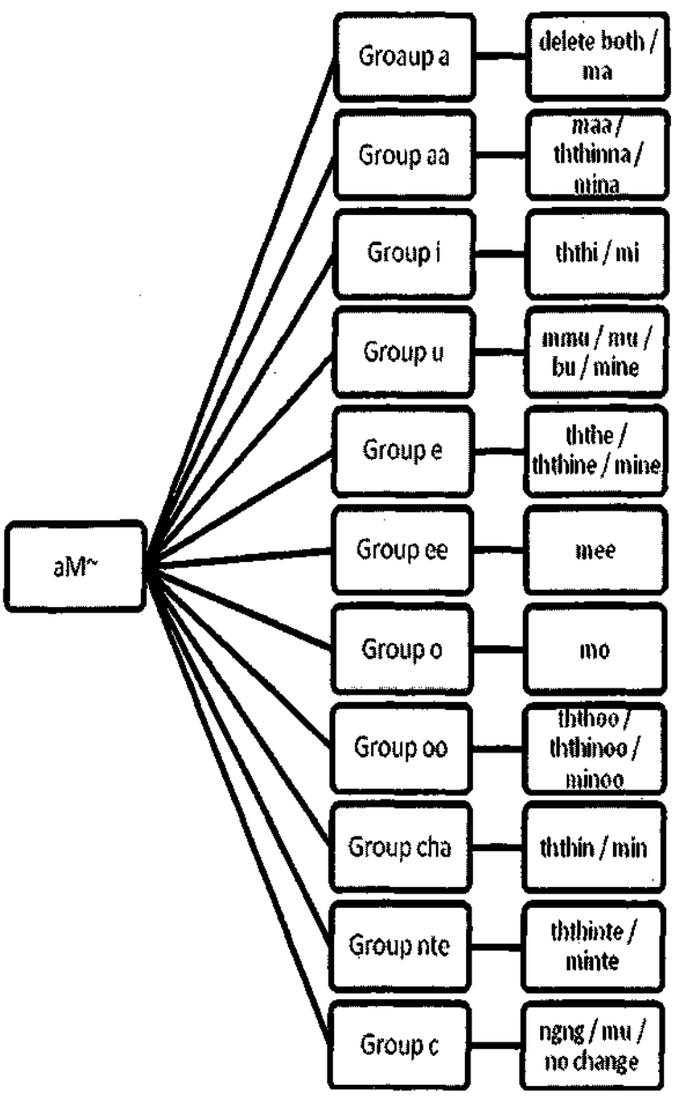


Figure 3.4 (f): Pictorial representation of morphophonemic change between the root ending phoneme aM~ (ಅಂ) and suffix groups

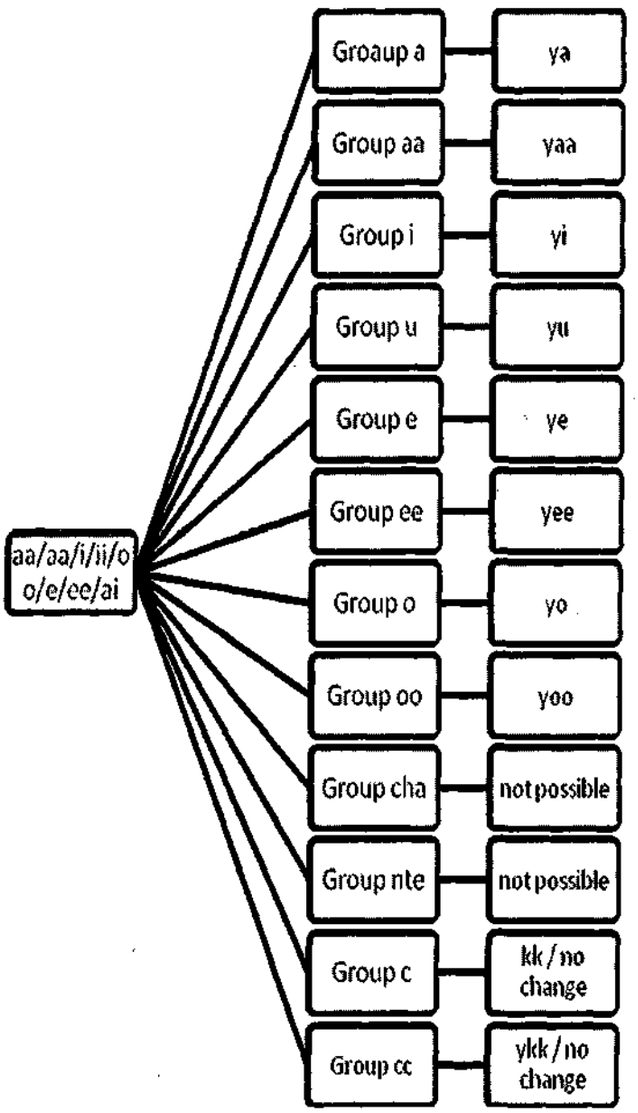


Figure 3.4 (g): Pictorial representation of morphophonemic change between the root ending phoneme starting with vowels except u/uu and suffix groups

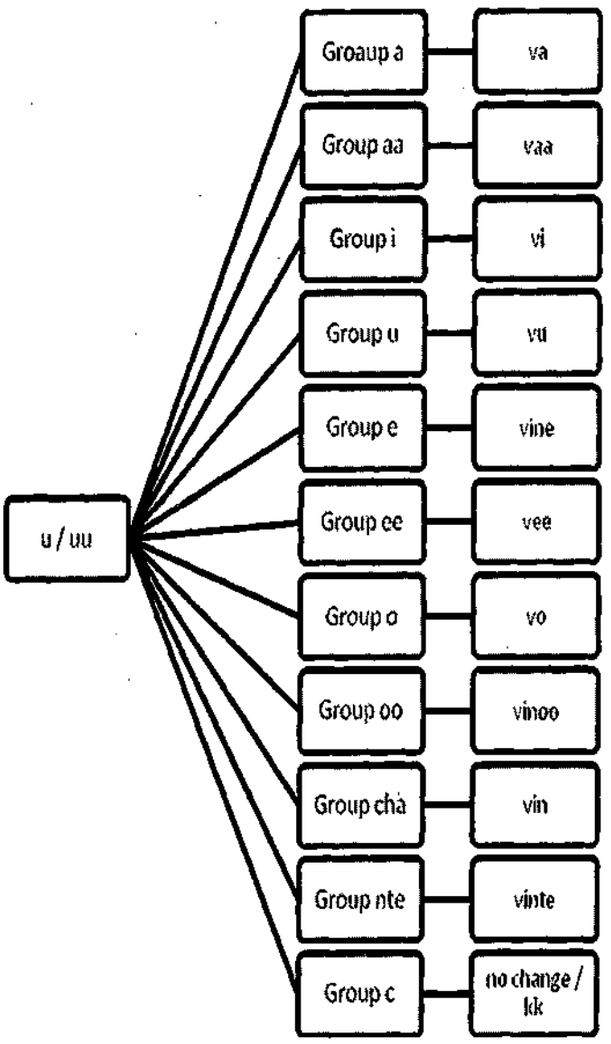


Figure 3.4(h): Pictorial representation of morphophonemic change between the root ending phoneme u/uu and suffix groups

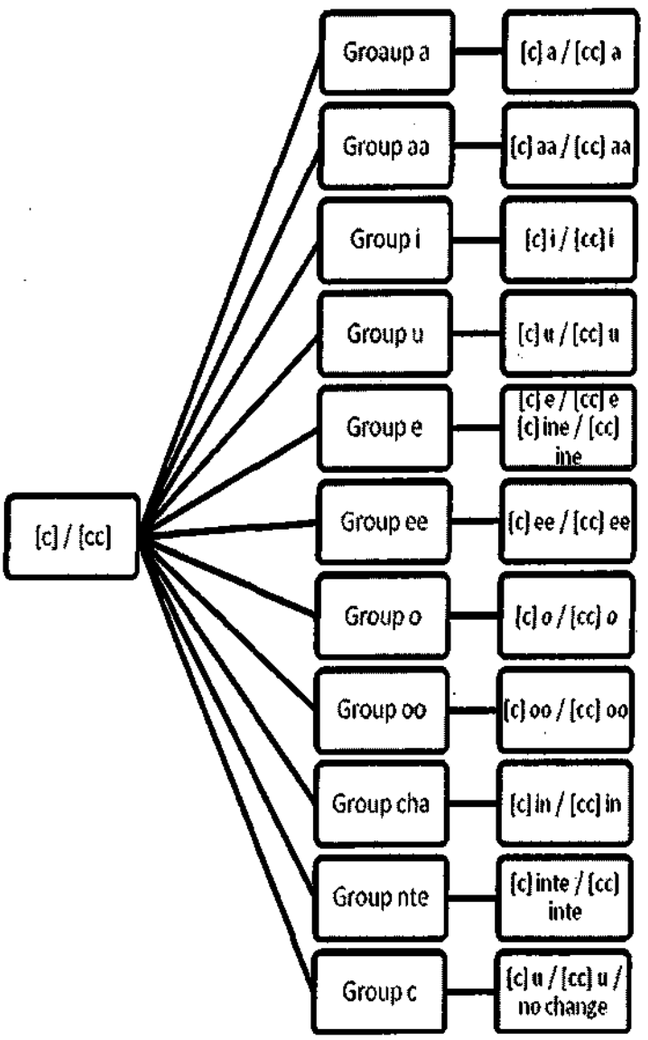


Figure 3.4 (i): Pictorial representation of morphophonemic change between the root ending phoneme [c]/[cc] and suffix groups

In this work, it is attempted to analyse the words structurally and identify the lexical entries in the corpus, concentrating on its orthographic pattern only. Suffixes inflected with words are identified and classified according to their occurrence. Efforts are made to study the pattern of variation during the process of morphophonemic change. Only a limited number of suffixes and examples are considered here. But more words are analysed and the results obtained were used for deriving morphophonemic rules which is discussed in the next chapter. The observations made from the analysis are given in the following section.

3.3 Observations

1. The data for language analysis is taken from the current day language in use. The common words, new words, technical terms, borrowed words are included for this study along with morphophonemic patterns.
2. The locative case marker, /-il~/, is usually used for specifying the location. That is, place names are suffixed with /-il~/ as in പത്തനംതിട്ടയിൽ (paththanaM~thitta- a place in Kerala). But it was observed that in some places /-il~/ suffix is replaced by /-aththʼ/ (/-അത്ത്/) suffix. Instead of /-il~/ suffix, /-aththʼ/ suffix is used because place names ending in /aM~/ usually shows this morphophonemic variation. Eg: thiruvananthapuraM~- thiruvananthapuraththʼ, kollaM~ - kollaththʼ, vakkaM~ - vakkaththʼ (തിരുവനന്തപുരം - തിരുവനന്തപുരത്ത്, കൊല്ലം-കൊല്ലത്ത്, വക്കം-വക്കത്ത്). This suffix is also used to specify time and position. Eg: mazhayaththʼ, mukhaththʼ (മഴയത്ത്, മുഖത്ത്).
3. One of the inferences from the analysis is that when a *Question word suffix* (wh-questions) is agglutinated, the word to which it is suffixed is always a noun. In case a verb requires suffixation of a questain word, the verb is normalised using /ath/ (/-അത്/) and then wh-question word is suffixed. Eg: paRanjnjathenthʼ (പറഞ്ഞതെന്ത്)
4. When sandhi has to be tackled computationally, the grammatical feature of the root does not play any role. The initial and final syllables are responsible for sandhi change (in some cases previous syllables should also be considered).
5. Not only case, gender, number, tense markers can decide the category of word, certain other suffixes like question word suffix, debitive suffix etc. can also decide whether the word is a noun or a verb.

6. Majority of foreign words in Malayalam are nouns in its root form or they are agglutinated with noun suffixes. If a verb word is borrowed, either it carries its inherent foreign suffix or it will use the forms of native marker (-cey-)(-ചെയ്-) to the verb form. The analysis shows that a large collection of foreign root words can be obtained just removing the noun suffixes.

7. There are certain words like *bisinassukaaran~* (ബിസിനസ്സുകാരൻ - business man), *bisinassukaari* (ബിസിനസ്സുകാരി - business woman), *bisinassukaaR~* (ബിസിനസ്സുകാർ - business persons) which have suffixes /-kaaran~/, /-kaari/, /-kaaR~/ (കാരൻ, കാരി, കാർ). In order to get the root word, these suffixes may also be included in noun suffix category. Similarly certain other root words like /-ഒന്ന് (/-onn/) can be added as a suffix by considering their frequency of occurrence in words. Besides the commonly occuring suffixes explained in this chapter, additional 8 noun suffixes, 9 verb suffixes and 6 dual functional suffixes such as /-ee/(/-ഏ/), /-maathraM~/ (/-മാത്രം/), /-aak/ (/-ആക്/), /-engkil~/ (/-എങ്കിൽ/) etc. are also added to increase the precision and accuracy of the system. So a total of 132 suffix categories are manually identified.

8. Occurrence of the verb, with past tense marker /-u/ is having a very low frequency in corpus. So in the identification of verb, '/-u/' is not used as a criteria for verb identification. The suffixes taken for identifying the past tense markers are given in section 3.1.1.4 in Table 3.5 as VS32. For more accurate results /-ceythu/ (/-ചെയ്തു/) can be considered as a single suffix in the cases like *tookk ceythu, rimuuv ceythu* (ടോക്ക് ചെയ്തു, റിമൂവ് ചെയ്തു) etc.

9. Using the suffix /-ath/ (/-അത്/), a verb can be changed to a noun. When this suffix is at the end of the word, that word is a normalised verb. Eg: paRanjnjath, ootiyath (പറഞ്ഞത്, ഓടിയത്).

10. The suffixes identified manually can also be computationally identified by sorting the words in the corpus according to the end syllable. Programs can be written to categorize and name them according to their grammatical features thereby develop a computational lexicon without manually inputting the suffixes.

11. More postpositional suffixes can be identified for suffixation. 158 postpositions are listed by grammarians like Gundert (1851), George Matthan (1863), A. R. Raja Raja Varma (1917), Seshagiriprabhu (1919) and L.V Ramaswami Ayyar (1936), S. Radhakrishnan Mallassery (1994) etc.

12. The computational grammar developed here can be used as an integral part in morphological analyser, morphological generator, coining technical terminology and many other natural language processing tasks.

13. Using the list of suffixes, a noun suffix dictionary, a verb suffix dictionary, pronoun dictionary etc. can be developed. As a future attempt, one can map these suffixes with their meaning and enrich the dictionaries.

14. The work can be extended and used for processing multiword strings. Eg: If *apakatanila, alpajnjaanaM~, apaayasuucana* (അപകടനീല, അല്പജ്ഞാനം, അപായസൂചന) are list of words, first check from right end, whether those words are in noun/verb dictionary. If present, remove it and apply the rules so that the output will be *apakataM~, nila, alpaM~, jnjaanaM~, apaayaM~, suucana* (അപകടം, നീല, അല്പം, ജ്ഞാനം, അപായം, സൂചന). Morphophonemic changes listed in this chapter can be further used for formulating rules for processing multiwords.

15. The dual functional suffixes are given here as a separate category for easy and accurate analysis. It can be simply termed as 'suffix' and generalise it for any word identification.

16. A structural analysis about noun, pronoun, verb, postpostions can be further carried out for understanding more about the neighbouring phonemes and morphemes, their pattern of formation and functions.

17. Since the memory or storage space is not a problem with computational lexicon the user can view them as a single word arranged with all its information.

18. The results of above analysis proved that corpus is the best resource for 1) obtaining all the words in use 2) studying the morphology and generating rules 3) studying the grammatical features 4) identifying the exceptions 5) for language research.

19. The list of suffixes and concepts are not exhaustive were some of the language specific constructions are not included in the list which requires further studies.

Grammar provides the rules for proper use of language. For developing language tools a comprehensive and well formulated grammar which is capable of analysing every aspect of the language with implications of computational techniques are required. This chapter gives an analysis of word formation for understanding the morphophonemics of the language. Morphophonemics involves an investigation of the phonological variations within morphemes, usually marking different grammatical functions [w22]. Morphophonemic rule has the form of a phonological rule, but is restricted to a particular morphological environment [w23]. Next chapter discusses how the morphophonemic change is represented as morphophonemic rule and it is implementation in a program called RWI to extract the root word.

Chapter IV

IMPLEMENTATION OF COMPUTATIONAL GRAMMAR FOR ROOT WORD IDENTIFIER

In chapter III, the structure and functions of Malayalam words are presented. In this chapter, the computational grammar analysed in the previous chapter is written as morphophonemic rules. These rules are implemented using a Root Word Identifier program which identifies the words, root form of words, suffixes along with their grammatical features. The root words obtained from RWI are compared with a language model made out of a corpus of Malayalam language, to check whether it is a native or foreign word. This comparison is done by a back end processing which is explained in the next chapter. The different steps involved in the pre-processing are discussed in detail. Computational lexicon and related results obtained from the system are also explained.

4.1 Pre-processing and Root Word Identifier (RWI)

The Root Word Identifier is a program designed to automatically remove the inflected part and derive the root of the word using morphophonemic rules. It can be adopted in many Natural Language Processing applications related with morphological level of analysis such as spell checker, language identifier etc. For example, if a corpus consists of native words and foreign words and are inflected with suffixes, a RWI removes these suffixes and transform it into its root word. Consider the word *bukkil~* (ബുക്കിൽ- in the book). Here */-il~/* (/-ഇൽ/) is native suffix and *'bukk'* (ബുക്ക്) is a foreign word. After removing the suffixes the phonemic composition of a word can be accurately studied.

In the context of Malayalam, it is assumed that the root word is the part of the word obtained by eliminating the suffixes and reforming the word with morphophonemic changes. The RWI removes the inflected suffixes, producing string of characters and can then transform it into original root form. The input data is taken in conventional Malayalam script rather than in their phonemic transcription forms. The system needs only a set of rules without any dictionary assistance.

The pre-processing of text data is diagrammatically shown in Figure 4.1. The rectangles / cylinders show the files and ovals show the processing.

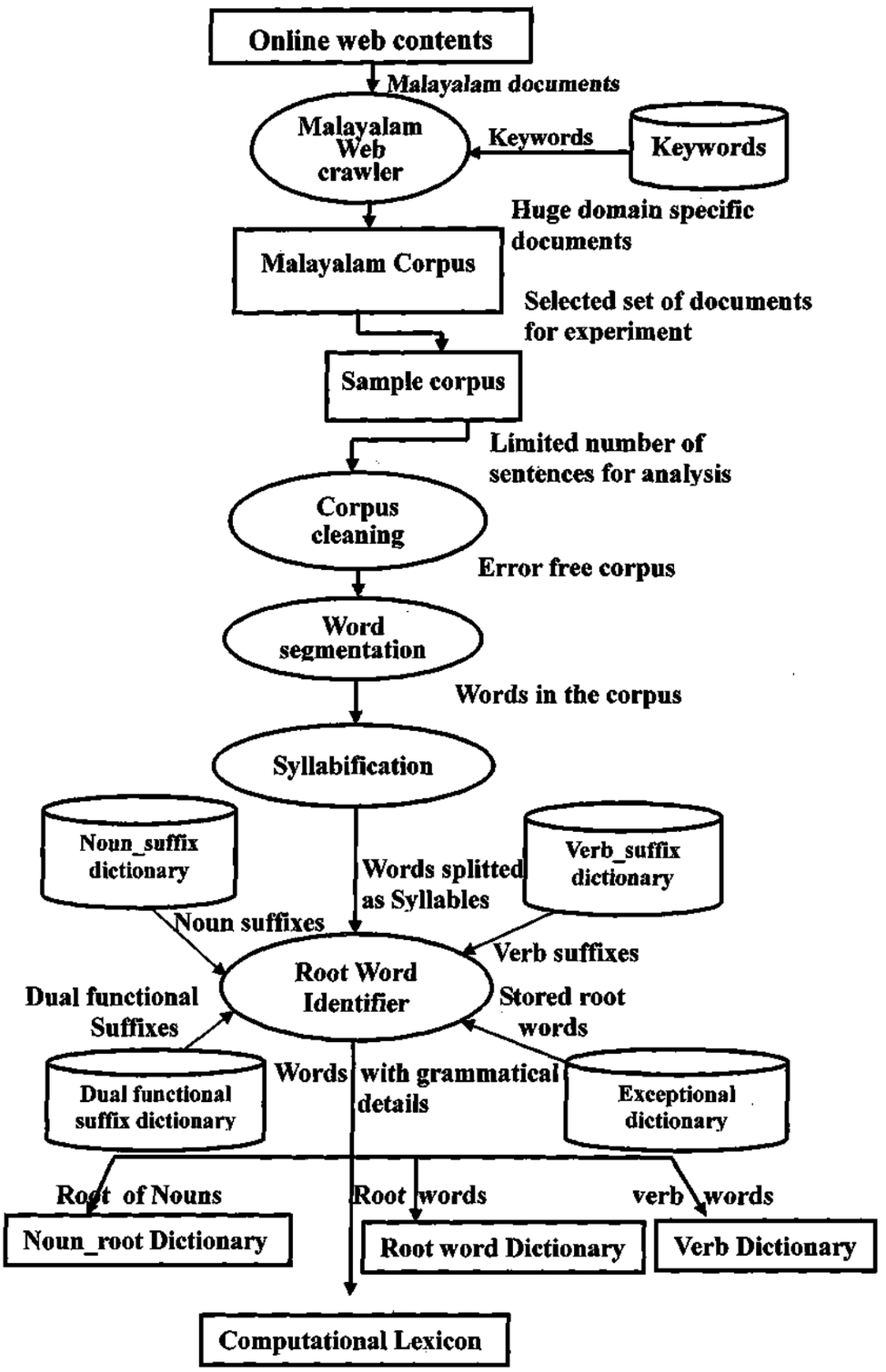


Figure 4.1: Block diagram of pre-processing of text data and RWI

The web crawler brings a domain specific corpus with meta data as explained in section 2.4. as data for this work. Random selections of sentences from the "Malayalam corpus" were collected for training the system. The sample corpus for the study consists of words from the domains of technology, health, law, film, lyrics of songs and religion. The total number of 23,045 words is given to the RWI. In the corpus cleaning as shown in Figure 4.1, the typographical forms which are having less relevance in this context free analysis are removed. Those typographical forms include punctuation marks, comma, colon, semicolon, exclamation mark, question mark, quotation marks, numbers, roman numerals, currency symbols, different types of brackets, dash, hyphen, ellipses, dot, asterisk, alternative signs (eg:/p/), white space etc. After this, the sentences are splited into legitimate words. This is referred to as word segmentation [McEnery 2006]. In the next stage, the process of syllabification will split the words into constituent syllable. Since Malayalam is a syllabic language, the words are stored as monosyllabic, disyllabic, trisyllabic upto eighteen syllabic words as shown in Table 4.1.

Table 4.1: Examples for syllabled words

Syllabic type	**Examples**
Monosyllabic words	naaM~ (നാം), see (സീ)
Disyllabic words	aM~ gaM~ (അം ഗം), a kkaM~(അ ക്കം)
Trisyllabic words	aa vee SaM~ (ആ വേ ശം), aa Svaa saM~(ആ ശ്വാ സം)
Quadrisyllabic words	bhaM~ gi yaa kki (ഭം ഗി യാ ക്കി), mo bai li l~(മൊ ബൈ ലി ൽ)
Penta syllabic words	ma la yaa Li yuM (മ ല യാ ളി യും)
Hexa syllabic words	du R~ ba la maa ya (ദു ർ ബ ല മാ യ)
Septi syllabic words	aM~ ga raa shtra ngng L~ kk (അം ഗ രാ ഷ്ട്ര ങ്ങ ൾ ക്ക്)
Octo syllabic words	ai kya raa shtra sa bha yu te (ഐ ക്യ രാ ഷ്ട്ര സ ഭ യു ടെ)
Ennea syllabic words	svaa bhaa vi ka ni ya ma ththi nte (സ്വാ ഭാ വി ക നി യ മ ത്തി ന്റെ)
Deca syllabic words etc.	aM~ gii ka ri kka ppe tu nna thi n (അം ഗീ ക രി ക്ക പ്പെ ടു ന്ന തി ന്) etc.

As shown in Figure 4.1, Noun_suffix dictionary contain suffixes categorized as nouns suffixes from NS1 to NS24 and postpositions categorized as PS1 to PS26. Dual functional_suffix dictionary contains suffixes categorized as dual functional suffixes from DFS1 to PDFS4 which is discussed in previous chapter. The Verb_suffix dictionary contains only the verb suffixes (VS1 to VS42). Exceptional dictionary contains certain root words which are listed in Table 4.3 as exceptional words.

The input to the RWI module is the words generated after syllabification. The various dictionaries such as Noun_suffix dictionary, Verb_suffix dictionary etc. are used by RWI module for obtaining the suffixes. Matching morphophonemic rules are applied to the syllabled words iteratively removing the suffixes until the root word is obtained.

The morphophonemic rules are derived from the morphophonemic analysis which is described in section 3.2. The general form of the rule is as follows.

4.2 General form of Morphophonemic Rule

Assuming that the processing is from the right end, a word W can be represented in the form

W = PXYZ where

Z is the suffix part,
Y is the link morph,
X is the replacement phoneme
P is the stem

During the process of Root Word Identification, the replacement phoneme will transform to root ending phoneme.

That is $X \rightarrow R$

where R is the root word's end phoneme.

The Root Word Identifier removes the suffix part (Z) and link morph (Y) thereby changing the replacement phoneme (X) to new root ending phoneme (R). This transformation can retrieve words ending in R. If X is any one of the glide /-y-/ or /-v-/ no transformation is carried on and the glides with suffixes are deleted. If X is same as R, only the suffixes are removed. No transformation or deletion is needed in that condition. Link morph is considered as the nearest phonemes to the suffix part. /-in-/ and /-u-/ are the link morphs. The suffix lists are listed in chapter III. Transformation of X to R is shown in Table 4.2. The [c] represents consonant and [cc] represents consonant clusters.

Table 4.2: Morphophonemic rule implementation

Rule Number	Replacement phoneme (X)	Root ending phoneme (R)	Example inflected word->root word
Rule 1	n (ന്)	n~ (ൻ)	raamane → raaman~ രാമനെ-> രാമൻ
Rule 2	N (ണ്)	N~ (ൺ)	aruNinaal~ → aruN~ അരുണിനാൽ->അരുൺ
Rule 3	L (ള്)	L~ (ൾ)	avaLil~ → avaL~ അവളിൽ->അവൾ
Rule 4	l (ല്)	l~ (ൽ)	sunilinoot → sunil~ സുനിലിനോട്->സുനിൽ
Rule 5	R (ര്)	R~ (ർ)	karaaRinte→ karaaR~ കരാറിന്റെ->കരാർ
Rule 6	r (ര്)	R~ (ർ)	avaril~ → avaR~ അവരിൽ->അവർ
Rule 7	y (യ്)	Delete y Delete (യ്)	ammaye → amma അമ്മയെ->അമ്മ
Rule 8	v (വ്)	Delete v Delete (വ്)	guruvine → guru ഗുരുവിനെ->ഗുരു
Rule 9	thth (ത്ത്)	aM~ (ം)	maraththil~→ maraM~ മരത്തിൽ->മരം
Rule 10	m (മ്)	aM~ (ം)	kaLLamallaathe→ kaLLaM~ കള്ളമല്ലാതെ->കള്ളം
Rule 11	[c]	[c]	kaaSine → kaaS കാശിനെ->കാശ്
Rule 12	[cc]	[cc]	klaassinte → klaass ക്ലാസ്സിന്റെ->ക്ലാസ്സ്

Apart from these general rules, some words which are having the suffixes /-ngngaL~/, /-nte/, /-nmaaR~/, /-kkaL~/ (/-ങ്ങൾ/, /-ന്റെ/, /-ന്മാർ/, /-ക്കൾ/) should take the root endings as /-aM~/, /-n~/, /-n~/, /-v/ (/-ാം/, /-ൻ/, /-ൻ/, /-വ്/). Examples: raajyangngaL~, naayakante, natanmaaR~, vakthaakkaL~ (രാജ്യങ്ങൾ, നായകന്റെ, നടന്മാർ, വക്താക്കൾ). Exceptional case for this rule is kunjnjungngaL~, makkaL~(കുഞ്ഞുങ്ങൾ, മക്കൾ) etc.

The suffixes are matched with words in such a way that the longest pattern will match first. The reason for the same can be explained with an example. Consider the case of suffixes /-aamaththe/ (/-ആമത്തെ/) and /-ththe/ (/-ത്തെ/). Both patterns match the word *raNtaamaththe* (രണ്ടാമത്തെ), but the first one accounts for more characters, reducing the word correctly to *raNtaaM~* (രണ്ടാം), while the second is an incomplete match. So the *longest-matching* technique [Mason 2000] is seen suitable here. Lists of words which cannot be processed with these rules are included in Exceptional_dictionary. Pronouns which can be considered as keywords are stored in this category. Some of the words in the exceptional dictionary are listed below in Table 4.3.

Table 4.3: List of exceptional words

angngane (അങ്ങനെ)	kumaaR~ (കുമാർ)	naan~ (ഞാൻ)
atraththooLaM~ (അത്രത്തോളം)	bil~ (ബിൽ)	maTTannaaL~ (മറ്റന്നാൾ)
kauN~sil~(കൗൺസിൽ)	kaan~ (കാൻ)	pinnaale (പിന്നാലെ)
raavile (രാവിലെ)	ooroo (ഓരോ)	pakkal~ (പക്കൽ)
koozhikkoot (കോഴിക്കോട്)	koccu (കൊച്ചു)	innu (ഇന്നു)

Certain ambiguities can be resolved by giving additional conditional rules. For example, after removing the suffix part, if only one syllable remains, exempt this word from processing and consider the word as a root word. This will increase the accuracy of the system. These exceptional words are categorised as clitics or indeclinables [Subhash 2010].

The words obtained after removing the noun suffixes are identified as noun words and the words which are having verb suffix are treated as verb words. Words having dual functional suffixes are treated as dual functional words. Such words are categorised as nominal or verbal word. Words without having any suffixes are treated are root words. All noun roots are stored in

Noun_ root Dictionary and verb words in Verb Dictionary. Root word Dictionary, Noun_root Dictionary and Verb Dictionary are the other results of RWI system as shown in the Figure 4.1.

4.3 Algorithm for Root Word Identifier

The algorithm for RWI consists of two functions. RootWordIdentifier() and MPRApplication(). It can be seen that RootWordIdentifier() will be calling MPRApplication() and MPRApplication() inturn call the RootWordIdentifier() in recursive way. The program iterates till the root form of the word is encountered and this root word is considered as the head word/lexical entry for the proposed computational lexicon.

```
RootWordIdentifier()
{
    Input a word W
   if (W is in Exceptional_dictionary)
       then store as root word
  else
      separate the suffix S from W
      if (S is in Verb_suffix dictionary)
        then store as a verb word
     else if(S is in Noun_suffix dictionary)
        then apply MPRApplication()
    else if ( S is in Dual functional_suffix dictionary)
       then apply MPRApplication()
       else
       Store W as root word
}
MPRApplication()
{
    Remove S
    Replace the replacement phoneme by the corresponding root ending phoneme from the
    Table 4.2 to form the word W1 and call the RootWordIdentifier() program with W1
}
```

4.4 Results and Discussion

Every word in the given "Malayalam corpus" was analysed as described in the previous section. The details like syllabic structure of the word, its root form and the morphemes/words agglutinated with this root form are systematically stored in the computational lexicon. The grammatical properties of the words such as noun or verb are represented along with them. Example of the computational lexicon generated using RWI and its byproducts are shown in Table 4.4 to Table 4.6.

Example I. **Simple sentence:-** *"amma kuttiye vati koNt aticcu"*
അമ്മ കുട്ടിയെ വടി കൊണ്ട് അടിച്ചു
Mother beaten the child with stick

Table 4.4: Output from Root Word Identifier for a simple sentence

Noun_root Dictionary	Verb Dictionary	Root word Dictionary	Computational Lexicon
ക ട്ടി	അ ടി ച്ചു	അമ്മ	അമ്മ -അ മ്മ <nominative case suffix> [Root word]
		കുട്ടി	കുട്ടിയെ- ക ട്ടി <accusative case suffix> [Noun]
		വടി	ക ട്ടി <nominative case suffix> [Noun] [Root word]
		കൊണ്ട്	വടി - വ ടി <nominative case suffix> [Root word]
			കൊണ്ട് -കൊ ണ്ട് [Cause expressing/instrumental suffix] [Root word]
			അടിച്ചു - അ ടി ച്ചു [Verb]

The above table shows the output obtained when a simple sentence is given to the preprocessing module of the system as a corpus. All words in the sentences, its syllabic structure, the suffixes included in each word according to its places of occurrence and grammatical category of word are shown in the Computational Lexicon. Name of suffixes are given in between two angle brackets (< >). In the above example, there is only one word having the noun

suffix, *kutti* (കുട്ടി). The system will first identify it as noun word and then as a root word. So this word have two entries in both Noun_root dictionary and Root word dictionary. *koNt* (കൊണ്ട്) is a suffix. So it is identified as a root word and stored in Root word Dictionary. The word *aticcu* (അടിച്ചു)is identified only as a verb word by considering the past tense marker /-ccu/ (/-ച്ചു/).

Example II. **Sentence having English words:-**

"*strees hooR~mooNinte pravaR~ththanaM~ kuRaykkaan~ kaRuththa cookleeTTin kazhiyunnu*"

സ്ട്രേസ് ഹോർമോണിന്റെ പ്രവർത്തനം കുറയ്ക്കാൻ കറുത്ത ചോക്ലേറ്റിന് കഴിയുന്നു

Black chocolate can reduce the action of stress hormone.

Table 4.5: Output from Root Word Identifier for a sentence having English words

Noun_root Dictionary	**Verb Dictionary**	**Root word Dictionary**	**Computational Lexicon**
ഹോർമോൺ	കുറയ്ക്കാൻ	സ്ട്രേസ്	സ്ട്രേസ് - സ്ട്രേ സ് <nominative case suffix> [Root word]
ചോക്ലേറ്റ്	കഴിയുന്നു	ഹോർമോൺ	ഹോർമോണിന്റെ - ഹോ ർ മോ ൺ <genitive case suffix> [Noun]
		പ്രവർത്തനം	ഹോ ർ മോ ൺ < nominative case suffix> [Root word]
		കറുത്ത	പ്രവർത്തനം - പ്ര വ ർ ത്ത നം <nominative case suffix> [Root word]
		ചോക്ലേറ്റ്	കുറയ്ക്കാൻ - കു റ യ്ക്കാ ൻ [Verb]
			കറുത്ത - ക റ ു ത്ത <nominative case suffix> [Root Word]
			ചോക്ലേറ്റിന് - ചോ ക്ലേ റ്റ് < dative case suffix> [Noun]
			ചോ ക്ലേ റ്റ് <nominative case suffix> [Root Word]
			കഴിയുന്നു - ക ഴി യു ന്നു [Verb]

The example shown in Table 4.5 contains three English words. The suffixes attached with all these English words are removed and stored in Root word Dictionary. Here only a single sentence is given as corpus. So the grammatical category of two words *pravaR~ththanaM~* (പ്രവർത്തനം) and *kaRuththa* (കറുത്ത) cannot be identified. A bigger corpus helps to identify the grammatical category of these words.

Example III. **Sentence having a word agglutinated with seven suffixes :-**

amma kuttiyutekuuteyillaaththathukoNtaaNoo avan~ varaanjnjath?

അമ്മ കുട്ടിയുടെകൂടെയില്ലാത്തതുകൊണ്ടാണോ അവൻ വരാഞ്ഞത്?

Is it because his mother is not with him, that the child did not come?

Table 4.6: Output from Root word Identifier for a sentence having words agglutinated with seven suffixes

Noun_root dictionary	**Verb dictionary**	**Computational Lexicon**
അമ്മ കുട്ടി അവൻ	വരഞ്ഞ	അമ്മ - അ മ്മ <nominative suffix> [Root word] കുട്ടിയുടെകൂടെയില്ലാത്ത തുകൊണ്ടാണോ - കു ട്ടി യു ടെ കൂ ടെ യി ല്ലാ ത്ത തു കൊ ണ്ടാ ണ് <ഓ> [dual functional word] <Or co-ordination/ Interrogative suffix> [Noun/Verb] കു ട്ടി യു ടെ കൂ ടെ യി ല്ലാ ത്ത തു കൊ ണ്ട് <ആണ്> [dual functional word] <Equation/ Finite verb suffix> [Noun/verb] കു ട്ടി യു ടെ കൂ ടെ യി ല്ലാ ത്ത ത്<കൊണ്ട്> [dual functional word] <Cause expressing/ Instrumental suffix> [Noun/verb] കു ട്ടി യു ടെ കൂ ടെ യി ല്ലാ ത്ത<അത്> [dual functional word] <pronominalization suffix>

				ക ട്ടി യു ടെ ള ടെ<ഇല്ലാത്ത> [Noun] <negative commitative suffix> ക ട്ടി യു ടെ<ളടെ> [Noun] <suffix showing interior movement > ക ട്ടി<ഉടെ> [Noun] <genitive case suffix> ക ട്ടി <nominative case suffix> [Root word]
		അവൻ	-	അ വ ൻ [Pronoun] [Root word]
		വരാഞ്ഞത്	-	വ രാ ഞ്ഞ<അത്> [dual functional word] <pronominalization suffix> വ രാ ഞ്ഞ [verb]

From the Table 4.6, it is observed that the words agglutinated with more than one or two suffixes can also be identified by the system. In the above example, the longest word is having thirteen syllables with seven suffixes. Apart from its root word, the system identified every word and its grammatical category included in that word. Different stages of word analysis and word formation, when each suffix is removed, can also be obtained from this computational lexicon.

The words in the verb category are words identified as verbs in the sentence. Since verb morphological analysis is not attempted in this work, root form of the verb is not obtained here. But the RWI can be modified to extract verb root with the help of the verb suffixation explained in section 3.1.1.4 and Morphophonemic Rule discussed in section 4.2. Here in this example, the verb word *varaanjnja* (വരാഞ്ഞ) can be processed to get the verb root *var(*വര്) by removing the suffix /-aanjnja/ (/-ആഞ്ഞ/) and using the morphophonemic Rule number 11.

The objective evaluation of the RWI is discussed in chapter VI. The Precision, Recall and F-measure of the RWI system were calculated. The system showed Precision above 90%. So, the

following section discuss about the features of the system, inferences obtained when the system process certain words, limitations shown by the system and future attempts that can be made to improve the existing system.

1. The system developed a huge collection of lexical database of inflected, non-inflected, double, compound, complex, reduplicated and foreign words from the Malayalam corpus.
2. This work uses a corpus of latest Malayalam documents and the generated lexicon includes many new words compared to the old lexicon developed manually by Suranand Kunjan Pillai in the year 1965. There is a drastic change in the vocabulary for current Malayalam in use compared with the old Malayalam usage. Since the system is automatically generating the computational lexicon from the corpus given, a bigger corpus can result in a bigger lexicon with more words.
3. The system is designed in such a way that the suffix helps to decide the grammatical category of a word. For example, if the word is *keeraLaaM~* (കേരളം), the system cannot predict the category, since this word has no suffix. But if anywhere in the corpus, there exist words like *keeraLaththile, keeraLaththil, keeraLaththinte* (കേരളത്തിലെ, കേരളത്തിൽ, കേരളത്തിന്റെ) etc., the system can show that the word *keeraLaaM~* (കേരളം) is a noun by considering the noun suffixes /-il~/, /-e/, /-inte/.
4. A large number of words were processed by the system. Pronouns, postpositions, inflected nouns, non-finite and finite verbs were identified correctly by the system.
5. The system is designed in such a way that its output such as words, root words, suffixes, grammatical categories etc. may be used for developing spell checkers and grammar checkers. The proposed system is similar to a morphological analyser. So the corpus, suffix dictionaries, list of root words and morphophonemic rule can be used for the development of a morphological generator for Malayalam language.
6. This work helps to understand certain common findings about words like (a) plural suffix will not come after the case suffix, (b) The occurrence of a verb in root form is rare, (c) quadrisyllabic words showed higher frequency of occurrence in the list of words in the corpus and trisyllabic words showed higher frequency of occurrence in the list of root words, (d) words having 18 syllables showed low frequency of occurrence in the list of words in the corpus and words having 13 syllables showed low frequency of occurrence in root words (e) The maximum number of suffixes agglutinated with the word in the

corpus is seven, etc. These types of information are useful for further linguistic studies. Such inferences can also be used for deriving rules for grammar checkers.

7. The information obtained from the analysed words help to extract semantic details also. For example, using the Noun_root Dictionary, more pieces of information about the word (animate, inanimate, common noun, proper noun etc.) can be identified.

8. Apart from basic grammatical information, knowledge about the frequency of syllables, words, root words, suffixes, nominal words, verbal words etc. helps to extract more information about words. This can be obtained by adding a code for counting the frequency of them to the RWI.

9. In addition to the twelve morphophonemic rules one more rule can be made. If a word end with /-zh/ (ഴ), after removing the suffix part and link morph, then this replacement phoneme /-zh/ (ഴ) can change to either ൾ, ഴ or ഴ്. (/-L~/, /-zha/, /-zh/). If such three options are given to the proposed system, it fails to select the appropriate root ending phoneme. In the test corpus used here contain only limited number of words having replacement phoneme /-zh/ (ഴ). Hence it was not included in this list of rules. A small percentage of error has occurred because of not adding this rule. Example for the above issue is illustrated using three words from corpus. Three options of the rule are shown in first column where first entry is the replacement phoneme and the second is the possible root ending phoneme.

Words/ Rule model	*appoozhuM~* (അപ്പോഴും)	*kazhakaL~* (കഴകൾ)	*thamizhil~* (തമിഴിൽ)
/-zh/-> /-L~/ (ഴ്-> ൾ)	*appooL~* (അപ്പോൾ) correct root	kaL~ (കൾ) wrong root	thamiL~ (തമിൾ) wrong root
/-zh/->/-zha/ (ഴ്-> ഴ)	appoozha (അപ്പോഴ) wrong root	*kazha* (കഴ) correct root	Thamizha (തമിഴ) wrong root
/-zh/->/-zh/ (ഴ്->ഴ്)	appoozh (അപ്പോഴ്) wrong root	Kazh (കഴ്) wrong root	*Thamizh* (തമിഴ്) correct root

10. Using the morphophonemic rules, majority of words were correctly processed by the system. Certain words which showed exceptions to these rules include words like *vaayil~, mukaLilaththe, paaril~, viittil~* (വായിൽ, മുകളിലത്തെ, പാരിൽ, വീട്ടിൽ) etc. Here, when the system applies the morphophonemic rules, the results obtained is as *vaa* (വാ) instead of *vaay* (വായ്), *mukaLilaM~* (മുകളിലം) for *mukaLil~* (മുകളിൽ), *paaR* (പാർ) for *paar* (പാര്) and *veett* (വീട്ട്) for *veet* (വീട്). Similarly when Rule Number 8 is applied, *kaavil~* (കാവിൽ) becomes *kaa* (കാ) instead of *kaav* (കാവ്) which has no meaning. There are some inflected words like *peeril~* (പേരിൽ) which can have the root form of either *peer* (പേര്) or *peeR~* (പേർ). The system cannot decide which root form has to be selected according to the context.

11. The system showed false result in processing these words. Words like *mayil~, koozhikoot, pinnil~, epril~* (മയില്, കോഴിക്കോട്, പിന്നിൽ, എപ്രിൽ) etc. are having word endings similar to suffixes. The system will automatically remove the word endings considering them as suffixes. Proper nouns also have noun suffix as similar to word endings. Eg: *SriikumaaR, kumaaR* (ശ്രീകുമാർ, കുമാർ). Similarly, consider the case of /-muuLaM~/ (/-മൂലം / -because of) which is used as a postpositional marker. But there are words like *sathyavangmulaM~, samulaM~, than~mulaM~* etc. (സത്യവാങ്മൂലം, സമൂലം, തൻമൂലം) which when removed will deviate from the original word. Likewise, certain Malayalam words are having phonological resemblance with another postpositional suffix /-vare/ (/-വരെ/), like *avare, ivare* etc. (അവരെ, ഇവരെ). Such words are included manually in the exceptional dictionary and they are stored as root word. Exceptional list also contain foreign words which ends with native suffixes. Eg: *sivil~, laTTil~, naanoo, ayyayyoo* (സിവിൽ, ലിറ്റിൽ, നാനോ, അയ്യയ്യോ) etc.

12. The system considers debitive suffix /-NaM~/ (/-ണം/) as a verb suffix. A good percentage of words are analysed correctly. But the noun words having similar word endings like *paNaM~* (പണം-money) *maNaM~* (മണം-smell) were wrongly taken as verb words. The system wrongly identified about seventy six words of similar case. Another issue related with this case, is in the processing of the word *ennitt* (എന്നിട്ട്), for example. The system will consider this word as a verb word by taking the end phonemes as a verb suffix (perfective suffix /-itt/ (-ഇട്ട്)). Actually the word is used to show the meaning "after that".

So it should be stored as a root word only and not as a verb word. These ambiguities can be solved either by making an appropriate selection manually or by developing an exceptional rule, considering its previous syllables.

13. Usually, the verb suffix /-i/ (-ി) is considered as a past tense marker. But for analysis, this condition is not given to identify the verb words since there are so many words ending with /-i/ which are not verbs (eg: *Sipaayi* ശിപായി- attender). So the system cannot precisely identify the words like *pooyi* (പോയി) as verbs. They will be tagged as root words only.

14. The proposed system will take certain words like *mariccavare* (മരിച്ചവരെ) and analyse it as a noun word by considering /-vare/ (/-വരെ/) as the suffix. But this word is a compound word having components *maricca* (മരിച്ച), *avaR~* (അവർ) and *e* (ഏ). Such errors affected the accuracy of the system.

15. When the English word 'battle' is transliterated to Malayalam, it shows similar orthographic pattern with the Malayalam word *baaTTiL~* (ബാറ്റിൽ) having the meaning "in the bat". The system cannot decide whether the word in the corpus is transliterated battle having the meaning "related with war" or "in the bat". If the meaning is "in the bat", the correct root word bat is obtained. But if the word is the transliteration of English word 'battle', the system shows the root word 'bat' which is actually wrong. The correct root word should be the word 'battle'.

16. A small percentage of exception is there in the morphophonemic rules. In Rule number 3, L (ള) changes to L~ (ൾ). But when the system apply this rule, correct root word is not obtained for some words. For example, in the word kaaLakaL~ (കാളകൾ), wrong root word kaaL~ (കാൾ) is obtained instead of kaaLa (കാള).

17. Certain words like *38-aaM~, alaksaaNtaR~ VI, 1907-il~, yuu-2-aaR, bi-47* (38-ാാം, അലക്സാണ്ടർ VI, 1907-ൽ, യു-2-ആർ, ബി-47) which are having importance in analysing a language are lost when typographical forms are removed during the pre-processing stage. So they are not entries in this lexicon.

18. Since the corpus has spelling errors certain words are not properly analysed. For example *angngine* (അങ്ങിനെ) is wrongly spelt. The correct spelling is *angngane* (അങ്ങനെ-like that). This affects the head word in the lexicon. Similarly assumption on word boundary (blank space between words) and removal of typographical forms resulted in a small percentage of error in the output.

19. The system is unable to identify each words and their grammatical properties included in the compound words (eg: *kaRuththaponn* കറുത്തപൊന്ന്), double words (eg: kuutekuute കൂടെകൂടെ), multiwords (eg:*vipaNanamaaRggaM~* വിപണനമാർഗ്ഗം) and compound verbs (eg: *koNtuvarika* കൊണ്ടുവരിക). This is because in this work the study is concentrated only on the morphophonemic change occurring when a word is combined with a suffix. To solve this problem, initially the morphophonemic change occurring when a word is combined with another word should be analysed. Then the rules should be derived to separate these words, so that the features can be identified. The system requires more morphemic, syntactic and semantic information based on the context in which the words are used. In this proposed system, these words will be stored as a simple lexical unit in the root word category.

20. Malayalam, being highly agglutinative and inflected, has got so many suffix or suffixes in between words. The RWI system can process only words ending with suffixes and not ending with words. So such words coming in the corpus are not completely processed. In these cases the system cannot identify the correct root word. For example, the word *asaadhyamaayithiirumoo* (അസാധ്യമായിതീരുമോ) when processed by the system becomes *'asaadhyamaayithiir'* (അസാധ്യമായിതീർ). Further analysis is not possible. So noun suffix /-aayi/ and noun root *asaadhyaM~* (അസാധ്യം) cannot be identified. This can be solved by framing rules to detect words and separate them.

21. The documentation of English language is aligned properly. Hyphens are used when there is a need to split words while writing. But the practice of using hyphens in Malayalam documents is very rare. In this work, when corpus is created using original web documents, the words coming in the margin may be wrongly stored as two words in the corpus file. That is, the corpus input to the system contains documents which are not properly hyphenated. So when the system tries to break the sentences into words, words can be wrongly split. This affects the correct analysis by the system. For example, if the word *kuttikaLil~* (കുട്ടികളിൽ) in a web document is written as *kutti* in the previous line and *kaLil~* in the next line, the word *kaLil~* is considered as a new word and the system will apply the rule. An empty space is obtained when removing the /-il~/ and /-kaL~/ suffixes. Such issues occurring in the corpus if rectified can increase the efficiency of the system.

22. Morphophonemic rules are developed using conventional method of writing. The system cannot analyse words when there is a change in the conventional writing patterns. For example, the word manushyanuveeNti (മനുഷ്യനുവേണ്ടി) when split, it should normally be written as *manushyan veeNti* (മനുഷ്യന് വേണ്ടി - for the man). But recently the medias used to write the word *manushyanu veeNti* (മനുഷ്യനു വേണ്ടി) with space in between. In this corpus such words are seen. But here the rule is not made to find the root word for the word *manushyanu* (മനുഷ്യനു). So the root word *manushyan~* (മനുഷ്യൻ) is not obtained. Instead the root word obtained is the word *manushyanu* (മനുഷ്യനു) which is wrong. If the word *manushyanuveeNti* (മനുഷ്യനുവേണ്ടി) is written clearly without space, the system will remove the suffixes /-veeNti/, the link morph /-u-/ and the correct noun root word *manushyan~* (മനുഷ്യൻ) will be obtained. If the word is written as *manushyan veeNti* (മനുഷ്യന് വേണ്ടി), then also the noun root word *manushyan~* (മനുഷ്യൻ) will be obtained based on morphophonemic Rule Number 1.

23. The proposed system cannot provide any information about the gender of a word which is an important grammatical feature of the Malayalam language. Words in the computational lexicon cannot show whether the word is masculine, feminine or neuter.

24. The rule based approach of identifying words and its grammatical properties can be tried for developing lexicon in other languages. Integrating them, "lexicon banks" can be developed by linking the words and mapping them with linguistic features so that the computer could understand all the features of words and their equivalents in other language.

The algorithm for automatically developing a RWI is presented in this chapter. The analysis and resulting lexicon with different examples are presented. The statistical analysis of phonemes and identification of foreign words are discussed in the next chapter.

Chapter V

STATISTICAL LANGUAGE MODELLING OF WORDS AND THE IDENTIFICATION OF FOREIGN WORDS

A Root Word Identifier provides the grammatical details of words, for understanding the function of a word in the language, as explained in the previous chapter. Another important information required for understanding the meaning and function of a word is its etymological details. Identifying the etymological details of a word is important for a wide range of applications such as annotation, lexicography, coining technical terminology, speech synthesis, machine translation, study on loan word phonology and for the study of foreign influence in a language. In order to transliterate an unknown word correctly or translate a proper name/technical term, it is often useful initially to identify the origin/etymology of the word. Identifying the source language plays a major role in information retrieval and cross-lingual information retrieval systems which helps to improve the indexing of search terms thereby retrieving more documents. Identifying foreign words is similar to the task of language identification [Beesley 1998], in which documents or section of documents are classified according to the language in which they are written. However, foreign word identification is made more difficult by the fact that words are nativized by the target language phonology and the fact that differences in character encodings are removed when words are rendered in the target language orthography [Baker et al. 2008].

In this chapter, the statistical approach using n-gram language modelling [Manning et al. 1999] is presented to analyse the features of foreign words occurring in Malayalam. This study is the first attempt in Malayalam to automatically develop a lexicon which includes the details of foreign words seen in Malayalam documents. The proposed lexicon will contain information about the source language of the word in addition to its grammatical properties. Apart from this a separate lexicon for native and foreign words can be further developed to have a foreign word dictionary or terminology data bank for the Malayalam language.

Etymology is the science of origin and history of words. It is the chronological account of the birth and development of a particular word or element of a word, often delineating its spread from one language to another and its evolving changes in form and meaning [w24]. In order to understand the etymology of words a detailed study on the phonotactic pattern of words is required.

5.1 Phonotactics

Phonotactics defines the permissible syllable structure, consonant cluster and vowel sequences by means of phonological constrains. Phonological constrains are language specific restrictions of phonological units/ segmentals [Asher et al. 1997]. For example, in Malayalam language all consonants in the fricative set occur only in loan words. Similarity, /æ/ occurs only in English loans. There are restrictions in position (initial, medial, final) of syllables also. For example, initial clusters do not occur in native words. Similarity in native Malayalam words, of the ten basic vowels /a, aa, i, ii, u, uu, e, ee, o, oo/, all except short /o/, occur word-finally [Asher et al. 1997].

A study on such restrictions/ phonotactic patterns helps to distinguish native words from foreign words. Several English consonants that occur frequently in English words such as 'm', 'f' etc. are expressed with Malayalam syllables such as 'eM~' (എം), 'eph' (എഫ്) etc. and that combinations do not occur frequently in pure Malayalam words. It is observed that the composition of Malayalam syllables seen in a foreign word is different from that of a pure Malayalam word due to the difference between the phonetic systems of the two languages. For example, the permissible syllable combinations seen in Malayalam words *kaat, kuLaM~* (കാട്, കുളം) etc. such as *kaa* and *t*, *ku* and *LaM~* share similar phonetic features and are frequently found in the phonological system of Malayalam language. The probability of such a combination is rare when a foreign word is transliterated to Malayalam. For example, the words like *phraNtshipp, kaaRD* (ഫ്രൺട്ഷിപ്പ്, കാർഡ്) etc. will have a syllable combination different from the phonetic pattern seen in pure Malayalam words.

In this work the distribution of phonetic patterns of native words is compared with the distribution of phonetic patterns of foreign words in order to identify the foreign words written in Malayalam scripts. Here for accuracy, the phonemes are considered as a cluster tagged as syllables. Cluster is a set of texts which statistically share similar linguistic features. It is the grouping of similar objects [Willett 1988]. The words are already syllabified in the pre-processing stage of RWI discussed in section 4.1.

5.2 Statistical Language Modelling using n-gram

Statistical Natural Language Processing can give statistical inference about the linguistic properties of natural language. For example, predicting the next word when the previous words

are given. The task of predicting the next word can be modelled as an attempt to estimate the probability function of the combination. This sort of task is commonly known as *Shannon game* [Shannon 1951] and is applicable in many natural language tasks. n-gram model can be used to analyse the probability distribution of word (syllable) combination in a language. The statistical and language independent nature of n-gram models proved to be successful in processing Indian languages. The common values of n can be n=1, n=2, n=3, n=4, etc. correspondingly named as unigrams, bigrams, trigrams, four-gram models. A *bigram* [Jurafsky et al. 2000] transition probability means the probability of a sequence of two tags: that is the probability that tag B will occur given that it comes directly after a sequence of tag A. The n-gram representation of the word *ciiph* (ചീഫ് - chief) at the syllable level is given below for the cases n=1, n=2 and n=3.

unigram	:	ചീ, ഫ്
bigram	:	0 ചീ, ചീ ഫ്, ഫ് 0
trigram	:	0 0 ചീ, 0ചീ ഫ്, ചീ ഫ് 0, ഫ് 0 0

where 0 denotes the "padding space" or a special string termination symbol for indicating an orthographic syllable boundary. If the word contains n syllables, n unigram, (n+1) bigrams, (n+2) trigrams etc. are possible.

5.3 Language Modelling for English and Malayalam Words

As shown in Figure 1.1, which shows the architecture of the system, back end processing module generates the n-gram language model for Malayalam and English words from a training corpus. The training corpus is created with 2886 English root words and 11,445 Malayalam root words. Malayalam words are collected from the Malayalam Lexicon [Pillai 1965] and from the Malayalam- Malayalam dictionary *Sabdathaaraavali*. English words are collected from an English dictionary [Namboothiri 1999]. There may be more than one transliteration for the same English word. For example, for the English word 'data', two different transliterations, *DaaTTa* (ഡാട്ട) and *DeeTTa* (ഡേറ്റ) are commonly used in Malayalam text. In the list of English root words, only *DeeTTa* (ഡേറ്റ) is present considering it as the standard form. Other transliterations are ignored.

5.4 Identification of Foreign Words

The algorithm for identifying the foreign word is designed in such a way that statistical information regarding the different syllables in foreign words and Malayalam words obtained

from a training corpus (back-end processing) are used to decide whether or not a given root word is a foreign word or Malayalam word.

Kwon, Jeong and Myaeng (1997) proposed a foreign word detection method for estimating whether a given word is a foreign word. They used the following formula for decision.

$$D(W) = \frac{P(Foreign \mid W)}{P(Native \mid W)} \quad \text{.....(1)}$$

In this work, Native is Malayalam language. So substituting equation (1) for Malayalam language.

$$D(W) = \frac{P(Foreign \mid W)}{P(Malayalam \mid W)} \quad \text{..... (2)}$$

where W is a root word. P (Foreign | W) and P(Malayalam | W) represents the conditional probability that W is a foreign word or a Malayalam word respectively.

Applying Bay's theorem,

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)\,P(B)}{P(A)}$$ [Manning et al. 1999]

and substituting A = W and B = Foreign,

$$P(Foreign \mid W) = \frac{P(W|Foreign)\,P(Foreign)}{P(W)} \quad \text{.....(3)}$$

Similarly, substituting A = W and B= Malayalam,

$$P(Malayalam \mid W) = \frac{P(W|Malayalam)\,P(Malayalam)}{P(W)} \quad \text{.....(4)}$$

Applying equation (3) and equation (4) in equation (2)

$$D(W) = \frac{P(W|Foreign)\,P(Foreign)}{P(W|Malayalam)\,P(Malayalam)} \quad \text{.....(5)}$$

If D(W) > 1, W is decided as a foreign word otherwise W is a Malayalam word.

In equation (5), P(W|Foreign) and P(W|Malayalam) are prior probabilities and are estimated by computing the ratio of foreign words and Malayalam words in the training corpus. For the estimation of P(W|Foreign), the syllables s_i s constituting W are assumed to occur independently from each other. So the following formula is used

$$P(W|Foreign) = \lambda_1 \cdot \prod_{i=1}^{n} P(s_i|Foreign) + \lambda_2 \cdot \prod_{i=1}^{n+1} (s_{i-1}\, s_i|Foreign) \quad \text{.....(6)}$$

where $\lambda_1 + \lambda_2 = 1$ and λ_1 and λ_2 represents the weights given to the two different statistics for estimating the probability namely unigram and bigram. s_0 and s_{n+1} are interpreted as special syllables indicating the start of a word and the end of a word respectively.

Similarly,

$$P(W|Malayalam) = \lambda_1 \cdot \prod_{i=1}^{n} P(s_i|Malayalam) + \lambda_2 \cdot \prod_{i=1}^{n+1} P(s_{i-1}\, s_i|Malayalam) \quad \text{.....(7)}$$

is also estimated.

In this work, n-gram model up to n=3 only is considered assuming that the system can identify the occurrence of a foreign word by looking its syllabic combination up to trigram model.

Algorithm

ForeignWordIdentifier()

{

Input the root word W

Evaluate the n-gram for root word W

Evaluate P(W|Foreign) using equation (6)

Evaluate P(W|Malayalam) using equation (7)

Compute D (W) using equation (5) and decide it as a foreign word if D (W) > 1

}

5.5 Results and Discussion

Every word obtained as root word from the RWI stage is analysed and identified whether the word is English or a Malayalam word. The two lists given below shows some of the words identified as English and Malayalam words stored in the English lexicon and Malayalam lexicon.

Words identified as English words
ടിവി
ടീ
ടീം
ടീച്ചർ
ടീംസ്
ടു
ടെക്നോളജി
ടെൻഷൻ
ടെലികമ്മ്യൂണിക്കേഷൻ
ടെലിവിഷൻ
ടോർട്ട്
ടൗൺഷിപ്പ്
ട്യൂൺ
ട്യൂമർ
ട്രസ്റ്റി
ട്രസ്റ്റ്
ട്രാക്ക്
ട്രാപ്
ട്രേഡ്
ട്രൈബ്യൂണൽ

Words identified as Malayalam words
തലച്ചോറ്
തലമുറ
തലവിധി
തവം
തവണ
താങ്കൾ
താണ്ടി
താത്പര്യം
താന്തോന്നി
താമസം
താര
താരം
താരതമ്യം
താൽപര്യം
താളം
താഴെ
താഴോട്ട്
താഴ്വാരം
തിര
തിരക്കഥ

The system will predict the root word as an un-decidable word, when D (W) = 1. That is, the system cannot predict whether it is an English or Malayalam word. For example, the orthographic pattern *nii* (നീ) is seen in English n-gram for representing the English word "Knee" and is seen in Malayalam n-gram for representing the Malayalam word *nii* having the meaning "you". Similarly *thai* (തൈ) have a meaning "thigh" in English and "sapling" in Malayalam. Same reason with *manth* (മന്ത്) having the meaning "month" in English and "filariasis" in Malayalam. Some of the un-decidable words are listed below.

Words identified as Un-decidable words
നീ
തൈ
മന്ത്

This section will discuss about inferences obtained from the analysis, the use of identifying foreign words in a language and limitations of the system. They include

1. Analysing the results, it is observed that foreign words can exist in its

   (a) Root form (Eg. phaan~ -ഫാൻ),

   (b) inflected form with native suffix (phaanil~-ഫാനിൽ),

   (c) inflected with both foreign and native suffix (kattinginaayi-കട്ടിങ്ങിനായി),

   (d) compound form with foreign and native words (skkuuL~kutti-സ്കൂൾകുട്ടി),

   (e) compound form with native and foreign word (paccakaLaR~-പച്ചകളർ),

   (f) foreign word-foreign word combination (beebisiTTR~-ബേബിസിറ്റർ),

   (g) foreign reduplicatives (TTippTTipp-ടിപ്പ്ടിപ്പ്)

   (h) foreign verb with native verb compound (Daan~suceyyuka-ഡാൻസുചെയ്യുക)

   (i) foreign adjectives combine with the be/become form [Asher et al. 1997] of native language (*ReDiyaayoo? റെഡിയായോ?*-are you ready?).

Since words in patterns (a), (b), (c), (f) and (g) are identified correctly, the result obtained from this work can be used to recognize the remaining compound words. Certain methods like maximum matching segmentation, complete matching segmentation, segmentation using syllable tagging method etc., which gave a satisfactory precision in identifying the compound nouns when used in Korean text [Kang et al. 2002] to identify foreign words can be used here.

2. When the foreign words identified by the system were sorted, words with similar phonemic pattern came together. Similar phonemic combinations coming in the initial position of words were taken as prefix. In the same way similar phonemic combinations coming in the end position of words were taken as suffix. These suffixes and prefixes are the same as that in English language when English words were borrowed into Malayalam language. The words with similar prefixes were grouped into prefix category and those with similar suffixes into suffix category. Table 5.1 shows the prefix and suffix category [Choudhury 2009], their phonemic combination in transliterated word, equivalent English prefix and suffix with example. The name of the categories is given according to the functions they possess.

Table 5.1: Prefix and suffix category of English words seen in Malayalam documents

Prefix category	Phonemic combination of prefix	Prefix	Equivalent English prefix	Example
Negation prefix	നോൺ	/nooN~-/	/non-/	നോൺവെജ് nooN~vejj
	ഇൽ	/il~-/	/il-/	ഇൽലീഗൽ il~liigal~
	ഇൻ	/in~-/	/in-/	ഇൻഎഫിഷ്യന്റ് in~ephishynt
	ഇം	/iM~-/	/im-/	ഇംമൂവബിൾ iM~muuvabiL~
	ഇർ	/iR~-/	/ir-/	ഇർറഗുലർ iR~RagulaR~
	ഡിസ്	/Dis-/	/dis-/	ഡിസ്ഒബേ Disobee
Ironic prefix	അൺ	/aN~-/	/un-/	അൺഡു aN~Du
	ഡി	/Dii-/	/de-/	ഡീകോഡ് DiikooD
Disgrace Prefix	മിസ്	/mis-/	/mis-/	മിസ് ലിഡ് misliiD
	മാൽ	/maal~-/	/mal-/	മാൽഫങ്ഷൻ maal~phngshan~
	സ്യൂഡൗ	/syuuDau-/	/pseudo-/	സ്യൂഡൗഇന്റലക്ട് syuuDauintlakTT

Size/ structure prefix	ആർച്ച്	/aaR~cc-/	/arch-/	ആർച്ച്ബിഷപ്പ് aaR~ccbishapp
	സൂപ്പർ	/suuppaR~-/	/super-/	സൂപ്പർന്യൂക്ലിയർ suuppaR~nyaaccuRal~
	ഔട്ട്	/autt-/	/out-/	ഔട്ട്റൺ auttRaN~
	സെർ	/seR~-/	/sur-/	സെർച്ചാർജ്ജ് seR~ccaaR~jj
	സബ്	/sab-/	/sub-/	സബ്നോർമ്മൽ sabnooR~mmal~
	ഔവർ	/auvaR~-/	/over-/	ഔവർഡു auvaR~Du
	അൺഡെർ	/aN~DaR~-/	/under-/	അൺഡെർഫ്ലഡ് aN~DaR~phuD
	ഹൈപർ	/haipaR~/	/hyper-/	ഹൈപർആക്റ്റീവ് /haipaR~aakTTiv/
	അൾട്രാ	/aL~traa-/	/ultra/	അൾട്രാവൈലറ്റ് /aL~traavailaTT/
	മിനി	/mini-/	/mini-/	മിനികാർ /minicaaR~/
Behavioural prefix	കൌ	/kau-/	/co-/	കൌയേക്സിസ്റ്റ് /kauyeeksisTT/
	കൌണ്ടർ	/kauNteR~-/	/counter-/	കൌണ്ടറാക്ക്റ്റ് /kauNteRaakTT/
	ആന്റി	/aanti-/	/anti-/	ആന്റിബോഡി /aantibooDi/
	പ്രൊ	/pro-/	/pro-/	പ്രൊഅമേരിക്കൻ /proameerikkan~/
Place prefix	ഇന്റർ	/intR~-/	/inter-/	ഇന്റർനാഷണൽ intR~naashaNal~
	ട്രാൻസ്	/traan~s-/	/trans-/	ട്രാൻസ്പ്ലാന്റ് traan~splaant
Time prefix	ഫോർ	/phooR~/	/fore-/	ഫോർട്ടേൽ phooR~TTeel~
	പ്രീ	/prii-/	/pre-/	പ്രീസ്കൂൾ priiskuuL~
	എക്സ്	/eks-/	/ex-/	എക്സ്മിനിസ്റ്റർ eksminisTTR~
	റീ	/Rii-/	/re-/	റീസ്സൽ Riissel~
	പോസ്റ്റ്	/poosTT-/	/post-/	പോസ്റ്റ്ബോക്സ് poosTTbooks

Number prefix	യൂനി/മോണോ	/yuuni-/ /mooNoo-/	/uni/ /mono/	യൂനിലാറ്ററൽ yuunilaaTTaRal~ മോണോക്രോം mooNookrooM~
	ബൈ/ഡൈ	/bai-/ /Dai-/	/bi-/ /di-/	ബൈപാസ് baipaas ഡൈമീറ്റർ DaimeeTTaR~
	ടൈ	/tai-/	/tri-/	ടൈസൈക്കിൾ taisaikkiL~
	മൾട്ടി/ പൊളി	/maL~tti-/ /poLi-/	/multi-/ /poly-/	മൾട്ടിനാഷണൽ maL~ttinaashaNal~ പൊളിക്ലിനിക്ക് poLikLinikk
Other prefixes	ഓട്ടോ	/oottoo-/	/auto-/	ഓട്ടോഗ്രാഫ് oottoograaph
	നിയോ	/niiyoo-/	/neo-/	നിയോരോമാന്റിക്ക് niiyootoomantikk
	പാൻ	/paan~-/	/pan-/	പാൻഅമേരിക്കൻ paan~ameerikkan~
	പ്രോട്ടോ	/proottoo-/	/proto-/	പ്രോട്ടോകോൾ proottookooL~
	സെമി	/semi-/	/semi-/	സെമിസർക്കിൾ semisaRkkiL~
	വൈസ്	/vais-/	/wise-/	വൈസ്പ്രസിഡന്റ് vaisprasiDnt
Suffix category	Phonemic combination of suffix	suffix	Equivalent English prefix	Example
	സ്റ്റർ/യർ	/-sTTaR~/ /-yaR~/	/-ster/ /-eer/	മിനിസ്റ്റർ minisTTaR~ സീനിയർ siiniyaR~
	ലെറ്റ്	/-leTT/	/-let/	ബുക്ക്ലെറ്റ് bukkleTT
	യെറ്റ്	/-yeTT/	/-yet/	സിഗയെറ്റ് sigayeTT
	ഹുഡ്	/-huuD/	/-hood/	മാൻഹുഡ് maan~huuD
	ഷിപ്പ്	/-shipp/	/-ship/	ഫ്രണ്ട്ഷിപ്പ് phraNtshipp
	ഡം	/-DaM~/	/-dom/	കിങ്ഡം kingDaM~
	ക്രെസി	/-kresi/	/-carcy/	ഡമോക്രെസി Dmookresi

	എറി	/-eRi/	/-ery/	ചെയെറി ceyeRi
	ഇങ് /ഇംഗ്/ ഇങ്ങ്	/-ing/ /-iMg/	/-ing/	ഗോയിങ്, ബാത്തിംഗ് gooying, baaththiM~g
	ഫുൾ	/-phuL~/	/-ful/	സ്പൂൺഫുൾ spuuN~phuL~
	ഐയറ്റ്	/-aiyaTT/	/-ite/	നക്സലൈയറ്റ് naksalaiyaTT
	ഇയൻ	/-iyan~/	/-ian/	ഇന്ത്യൻ inthyan~
	ഈസ്	/-iis/	/-ese/	ജാപ്പനീസ് jaappaniis
	ഇസ്റ്റ്	/-isTT/	/-ist/	ബുദ്ദിസ്റ്റ് buddisTT
	ഇസം	/-isaM~/	/-ism/	ക്യാപ്പിറ്റലിസം kyaappiTTalisaM~
	അർ	/-aR~/	/-er/	വർക്കർ vaRkkaR~
	റ്റർ	/-TTaR~/	/-or/	യിൻസ്റ്റക്റ്റർ yin~stakTTaR~
	ആന്റ്	/-aant/	/-ant/	ഇൻഫോൻമാന്റ് in~phoon~maant
	ഈ	/-ii/	/-ee/ /-y/	എംബ്ളായീ, ക്രീമി eM~bLaayii, kriimii
	ഏഷൻ	/-eeshan~/	/-ation/	നാവികേക്ഷൻ naavikeekshan~
	മെന്റ്	/-ment/	/-ment/	അറേജ്മെന്റ് aReejment
	അൽ	/-al~/	/-al/	റഫുസൽ, കൾച്ചറൽ Rphuusal~, kaL~ccRal~
	ഏയ്ജ്	/-eeyj/	/-age/	കവറേയ്ജ് kavaReeyj
	നെസ്സ്	/-ness/	/-ness/	ഹാപ്പിനെസ്സ് haappiness
	ഇറ്റി	/-iTTi/	/-ity/	സാനിറ്റി saaniTTi
	ഫൈ	/-phai/	/-ify/	കോഡിഫൈ kooDiphai
	അയ്സ്	/-ays/	/-ise/	പോപിലറയ്സ് poopiiaRays
	എൻ	/-en~/	/-en/	റയ്പ്പൻ Ryppan~

	ലെസ്സ്	/-less/	/-less/	ഹാംലെസ്സ് haaM~less
	ഇലി	/-ili/	/-ly/	മാനിലി maanili
	ലൈക്ക്	/-laikk/	/-like/	കൌലൈക്ക് kaulaikk
	ഇഷ്	/-ish/	/-ish/	ഫൂളിഷ് phooLish
	ഇക്ക്	/-ikk/	/-ic/	ഹീറോയിക്ക് hiiRooyikk
	ഇവ്	/-iv/	/-ive/	അട്രാക്ടിവ് atraakTTiv
	ഷ്യസ്	/-shyas/	/-ous/	ആമ്പിഷ്യസ് aampishyas
	അബിൾ	/abiL~/	/able/	റീഡാബിൾ RiiDaabiL~
	ന്റഡ് / ഡ്	/-ntaD/ /-D/	/-inted/ /-d/ /-ed/	പോയിന്റഡ്, കാർഡ് pooyintaD, kaaRD
	വേർഡ്സ്	/-veeRDs/	/-wards/	ബാക്ക്വേർഡ്സ് baakkveeRDs
	വൈസ്	/-vais/	/-wise/	എഡ്യുക്കേഷൻവൈസ് eDyukkeeshn~vais
	ക്കൽ	/-kkal~/	/-cal/	ക്ലിനിക്കൽ klinikkal~
	ഫ്	/-ph/	/-ugh/ /-of/	റഫ് Rph
	ഴ്സ്	/-zhs/	/-rse/	കോഴ്സ് koozhs
	ർട്ട്	/-R~tt/	/-rt/	കാർട്ട് kaaR~tt

Similarly, words used for representing cardinal numbers, names of weeks, name of certain places, names of months and measurements used in different languages have similar phonemic combinations. It was observed that such suffixes and prefixes can also be considered for understanding the phonemic structure of foreign words borrowed from other languages like Hindi and Sanskrit. This will also help to identify Hindi loan words and Sanskrit loan words in Malayalam. It can be used to convert numerals to numbers and vice versa. Examples are shown in Table 5.2.

Table 5.2: Words having similar initial or final phonemes seen in English, Hindi and Sanskrit loan words.

English loan words	തേർട്ടീൻ (13)	ഫോർട്ടീൻ (14)	ഫിഫ്ടീൻ (15)
	സിക്സ്റ്റി ടൂ (62)	സിക്സ്റ്റി ത്രീ (63)	സിക്സ്റ്റി ഫോർ (64)
	ടൂ ഹൻഡ്രഡ് (200)	ത്രീ ഹൻഡ്രഡ് (300)	ഫോർ ഹൻഡ്രഡ് (400)
	സൺഡേ (Sunday)	മൺഡേ (Monday)	ട്യൂസ് ഡേ (Tuesday)
Hindi loan words	തേരഹ് (13)	ചൌദഹ് (14)	പന്ദ്രഹ് (15)
	ബാസഠ് (62)	തിർസഠ് (63)	ചൌസഠ് (64)
	ദോ സൌ (200)	തീൻ സൌ (300)	ചാര് സൌ (400)
	രവിവാർ (Sunday)	സോമവാർ (Monday)	മംഗൽവാര് (Tuesday)
Sanskrit loan words	ത്രയോദശ (13)	ചതുർദശ (14)	പഞ്ചദശ (15)
	ദ്വിഷഷ്ടി: (62)	ത്രിഷഷ്ടി: (63)	ചതുഃഷഷ്ടി: (64)
	ദ്വിശതം (200)	ത്രിശതം (300)	ചതുഃശതം (400)
	രവിവാരം (Sunday)	സോമവാരം (Monday)	കുജവാരം (Tuesday)

From the table it is clear that the suffixes like /-iin~/ (/-ഈൻ/), /-ah/ (/-അഹ്/), /-daSa/ (/-ദശ/) etc. can be considered as suffixes used for representing numbers 13, 14, 15 in foreign words like English, Hindi and Sanskrit respectively. In Hindi loan words, the suffix /-vaaR~/ (/-വാർ/) shows the name of week. Corresponding suffix used in Sanskrit for showing the name of the week is /-vaaraM~/ (/-വാരം/). In English it is *Dee* (ഡേ).

3. Comparing and analysing the list of native and foreign words, many sociolinguistic inferences can be traced out. It is possible to interpret the co-relation of culture with the word. For example, if a word in native Malayalam lexicon is *ooNaM~* (ഓണം), the festival of Kerala, it may have a native element with the culture of Kerala. That is, the word is not of foreign origin. This word seems to frequently appear in Malayalam literature than in a foreign literature. Another example is the word in the English lexicon *krismass* (ക്രിസ്മസ്സ്) which is of foreign origin. Similarly the proper names like *deepa* (ദീപ), *paaRu* (പാറു) etc. are of native origin and *krisTTi* (ക്രിസ്റ്റി), *muhammad* (മുഹമ്മദ്) etc. are of foreign origin. Comparative study on name of mythological characters, name of national heroes, name of places etc. seems to convey more information than their mere

referential meaning. Thus the etymological details can help to clarify meanings of lexical unit or give clues to understand obscure meanings. Since they carry some special sense of meaning, they should also be included in the lexical stock of a computational lexicon.

4. Each suffix in English language will have an equivalent suffix in Malayalam language. The present work has both the list of suffixes which are agglutinated with these languages. Using these suffixes, a comparative study concentrating on its semantic criteria, can be done which contributes to the task of English-Malayalam machine translation. For example, if the word is an English word nooN~vejjiTTeeRiyan~ (നോൺവെജ്ജിറ്റേറിയൻ), equivalent Malayalam word can be obtained by substituting for the suffix /nooN~/ (/നോൺ-/) with /-allaaththa/ (/-അല്ലാത്ത/) giving vejjiTTeeRiyanallaaththa (വെജ്ജിറ്റേറിയനല്ലാത്ത-not a vegetarian).

5. It was observed that different spelling systems exist for foreign words when compared with native words. This shows the need for having a detailed study on how to transliterate a foreign word according to its pronunciation and orthography to its native form, when a new word comes to the language. For example, the English word yellow is written as *elloo* (എല്ലോ) in Malayalam script different from the pronunciation in English as *yelau* (യെലൌ).

6. By using the list of foreign words, it is observed that rules can be developed to transliterate each phoneme in foreign word to its equivalent phoneme in source language (here English). Similarly, in reverse also. So, if a word from the English language is given, the equivalent transliterated phoneme in target language for the word can be derived. For example, in English if a word starts with the alphabet A, it will either take *a, aa, e, ee* or *oo* as its possible initial rendering when it was written in target language (like *aaskk* [ആസ്ക് -Ask], *egeen*~[എഗെയിൻ-Again], *ooL*~[ഓൾ-All]). Such a system can be used in search engines which will help to enhance the depth of search thereby retrieving information from both source language and target language. It helps in the process of translating the acronyms easily.

7. Exceptional words contain many English words which are having suffixes similar to their word endings. In this work, such words are manually identified and stored as exceptional words. The RWI program can be modified by first identifying the English words

automatically and then implement the rule for obtaining a more accurate result. So the exceptional words like *kumaaR~* (കുമാർ), *bil~* (ബിൽ), *kaan~* (കാൻ) etc. can be identified as foreign words and exempt them from applying morphophonemic rule.

8. A comparative study using the n-grams of both languages can give more information about the phonems and their probability of distribution such as the phonemes which are common in native/foreign words, the position of phonemes in a word etc. Such an empirical study can be used in areas of designing OCR system, designing fonts, designing keyboards and in text to speech converters.

9. It was observed that prepositions used in English (of, to, at, by etc.) are rarely borrowed by the recipient speakers. Similarly the phrasal verbs like fell down, get through are rare. Idioms and fixed expressions in English are also rarely borrowed. Some of the exceptions seen in Malayalam text are *ooph haant* (ഓഫ് ഹാന്റ് -off-hand), *shoott katt* (ഷോട്ട് കട്ട് short-cut) etc.

10. Using n-gram of native and foreign language, a system can be modelled to identify the spelling errors in words. For example, it is possible to provide rules to identify spelling errors when a word starts with any of the *cill* (ചില്ല്) characters [Appendix III] or when two *cill* come closely. Similarly there are certain syllables which are seen only in foreign words. /æ/ occurs only in English loans. Providing such information as rules, a spell checking system can be developed and integrated to the RWI system for obtaining a more accurate spelling of words in the lexicon.

11. The collection of borrowed words can be used to study the general attitude of recipient speakers towards borrowings, frequency of borrowing patterns in recipient language etc. It is also a good resource for doing research on loan word phonology, loan word translation, study on code-switching and code-mixing [Cantone et al. 2009].

Performance analysis of foreign word identification module discussed in chapter VI showed a Precision of 62.5% and a Recall of 72.99% by the system. Some of the limitations of the system which were noticed when the system tries to identify the foreign word are as given below:

12. It was observed that statistical approach can be successfully applied to the language identification task. But a major drawback to apply the statistical approach to foreign word identification is the requirement for a sufficient amount of labelled training examples. The

manually constructed training corpus for language modelling is too small that this machine learning approach will not get enough data to plot all probabilistic occurrences of syllables. Collecting large list of Malayalam root words and English root words is expensive and time consuming. Due to the same reason, the proposed system wrongly analysed English words as Malayalam words. The sparseness of training data can be overcome by increasing the training corpus.

13. In this proposed system, monosyllablic words are the one which are poorly identified. The reason for the same is that there will not be an equivalent phonotactic pattern for this monosyllable word in the corresponding training corpus. For example, if the input word is *kyu* ക്യൂ (Queue), it should have an equivalent phonotactic pattern in the training corpus, particularly in English n-gram model. Then only the system can predict that the word is an English word. Otherwise the system cannot analyse such words. Training corpus having more monosyllablic words can solve such problems. Some of the other words which the system fails to analyse are *vi* (വി), *ke* (കെ) etc.

14. In this work, all words identified by the system as foreign words are not of English origin. Sometimes the words identified as foreign word have their origin in Sanskrit, Arabic or Latin. For example, the system will take the word *al~ameen~* (അൽഅമീൻ- Alameen) as a foreign word only. It can be modified further to obtain the etymology as Arabic language.

15. The computational lexicon will mark the word *thrii* (ത്രീ) correctly as an English word. But due to different spelling systems existing for transliterated foreign words, it is also seen as *thri* (ത്രി) in the corpus. So the system will tag it as a Malayalam word. Actually it is seen as a prefix of certain Sanskrit words.

16. No attempt is made to find the foreign syllables which are seen only in transliterated English words. If they are found, knowledge about them can be used in transliteration, back transliteration, spell checkers and development of OCR's.

17. Compound words which include both English and Malayalam words showed poor results during the analysis of words such as *e.es.ai sukumaaran~* (ഏ. എസ്. ഐ. സുകുമാരൻ), *appiil~kootathi* (അപ്പീൽകോടതി) etc.

18. RWI fails to identify certain English word whose end phonemes resemble the suffixes of Malayalam words. Those words are wrongly analysed by RWI. Eg: *ReeDiyoo* (റേഡിയോ),

caakkoo (കാക്കോ) etc. This problem can be tackled by first identifying whether it is an English word or Malayalam word and then pass to RWI program for more accurate results.

19. No attempt was made to process the words obtained as un-decidable words. Words like *cellaan~* (ചെല്ലാൻ) having two meanings "To go" in Malayalam language and "a form issued for money transfer" in English are not processed further to find its etymology.

20. A system designed for developing a computational lexicon should give much more information about words like whether the word is countable, uncountable, transitive verb, intransitive verb etc. Variation in language (whether regular or irregular, frequently or rarely used, normal or stylistic etc) and details like whether the word is archaic, regional, obsolete etc. are also not indicated. Hyponymous compounds, case relationships, kingship terms, technical terminologies, co-ordinate compounds, echo compounds, complex compounds, verb-noun compounds, adverb-noun compounds, adjective-noun compounds, verb-verb compounds are not analysed. Since a word carries a bundle of information related to the phonology, morphology, morphophonemics, morphosyntax, lexicography, semantics, syntax, text, grammar, etymology etc., it seems to be a difficult task to capture all its information by considering only its surface form or its orthography [Pinker 1995]. But this work can be extended for retrieving more information about words.

5.6 Further Prospects

The large collection of English and Malayalam root words obtained from this work can be fed to the training corpus thereby increasing the number of English root words from 2886 to 6996 and Malayalam root words from 11445 to 26135. This helps to identify more foreign words without manual assistance. This model can be used for extracting foreign words from any language.

For increasing the accuracy of the system, machine learning methods such as Support Vector Machines [Kruengkrai et al. 2005], Normalised Dot Product [Damashek 1995], K-nearest Neighbour, Relative Entropy [Sibun et al. 1996], Conditional Random Fields [Lafferty et al. 2001] etc. can be tried out. Other applied methods include Decision Trees, Neural Networks and Multiple Regression [Botha et al. 2007] can be experimented to retrieve the etymology of any foreign language. The smoothing techniques [Chen et al. 1999] can also be experimented to obtain more accurate results such as Laplace smoothing used by Dunning [Dunning 1994]. The advanced smoothing methods like Absolute Discounting, Katz Smoothing, Kneser-Ney Smoothing, Pruning Algorithm etc. used for identifying language of the document can be tried to

identify the foreign word. Entropy Based Pruning [Stolcke 2002], Revised Kneser pruning [Siivola et al. 2007] can be used to exclude the n-gram that contributes only little to the modelling accuracy [Vatanen et al. 2010].

This chapter discussed a method for automatically identifying the foreign words, which are borrowed into Malayalam, using minimum monolingual resources, that is, with no bilingual dictionary or transliterated text. Foreign words or out-of-vocabulary words or loan words are those words which are borrowed from other languages, written in native script. Here the English words scripted in Malayalam orthography are considered only as foreign words. A corpus developed from web is used as the resource for data analysis. The root words in the corpus are identified and they are used for identifying the English words. Statistical language modelling using n-gram is used in the task of predicting the phonemic combinations of both foreign and native language. The computational lexicon will have an entry to denote whether the root word is native or foreign word.

The output from the system helps to study the phonotactic patterns of English words. Some of the commonly occurring prefixes and suffixes are classified and listed. This result can be used for developing spell checkers, grammar checkers etc. The system exposes the possible phonotactics of current Malayalam language and also gives more information about different writing/spelling system used for transcribing foreign words.

There should be a unique writing and spelling system for foreign words to make the native language suitable for computational processing and help the common man to write them without spelling errors. This approach is a first step towards that mission. Next chapter evaluate the complete system.

Chapter VI

PERFORMANCE EVALUATION OF THE SYSTEM

For developing a computational lexicon for Malayalam language, a system was developed to identify the lexical items, its grammatical properties and etymology. A monolingual corpus of the current Malayalam language is created using a web crawler. Using this Malayalam corpus, the computational grammar of Malayalam words is designed to understand the structure and features of word. Studying the morphophonemic variations occurring at the boundaries of root word and suffixes, morphophonemic rules are derived and these rules are implemented using a RWI to extract root words with its grammatical properties. The system also identifies the English loan words in Malayalam documents which gave important information to derive the etymology of words. An elaborate performance evaluation of the developed system is presented in this chapter.

The overall system analyses the input corpus and generates the computational lexicon with linguistic knowledge about Malayalam words. A sample computational lexicon generated by the system for Malayalam language is explained in section 6.1. An elaborate performance evaluation for RWI module and foreign word identification module, using some performance metrics is presented in section 6.2.

6.1 Computational Lexicon

The proposed system developed a computational lexicon having 21,446 lexical entries. As the size of the corpus increases, this lexical database will also increase. If the developed system is fed with a small corpus containing a single sentence as given below, the following linguistic information are generated as shown in Figure 6.1.

samiikRRthaahaaraththil~ proottiin~, minaaRal~, viTTaamin~ iva atangngiyirikkunnu

(സമീകൃതാഹാരത്തിൽ പ്രോട്ടീൻ, മിനറൽ, വിറ്റാമിൻ ഇവ അടങ്ങിയിരിക്കുന്നു –

A balanced diet contains proteins, minerals, vitamins etc)

Lexical entry	Linguistic information		Details
സമീകൃതാഹാരം	Syllable structure	:	സ മീ കൃ താ ഹാ രം
	Etymology		മലയാളം
	Grammatical category		നാമം
	Root form		സമീകൃതാഹാരം
	Inflected forms	N	**സമീകൃതാഹാരത്തിന്റെ**
			സമീകൃതാഹാരം <Genitive case suffix>
		N	**സമീകൃതാഹാരങ്ങൾ**
			സമീകൃതാഹാരം <Plural suffix>
		N	**സമീകൃതാഹാരമല്ലാത്ത**
			സമീകൃതാഹാരം <Negative adjectival suffix>
		N	**സമീകൃതാഹാരത്തിൽനിന്ന്**
			സമീകൃതാഹാരം <Abative case suffix>
		N	**സമീകൃതാഹാരമായ**
			സമീകൃതാഹാരം <Relative adjectival suffix>
		N	**സമീകൃതാഹാരങ്ങളുടെ**
		N	**സമീകൃതാഹാരങ്ങൾ**<Genetive case suffix>
			സമീകൃതാഹാരം <Plural suffix><Genetive case suffix>
		N	**സമീകൃതാഹാരത്തെകുറിച്ച്**
		N	**സമീകൃതാഹാരത്തെ**<Referencing suffix>
			സമീകൃതാഹാരം <Accusative case suffix><Referencing suffix>
		N	**സമീകൃതാഹാരത്തിനുവേണ്ടി**
		N	**സമീകൃതാഹാരത്തിന്** <Benefactive suffix>
			സമീകൃതാഹാരം <Dative case suffix><Benefactive suffix>
		DFW	**സമീകൃതാഹാരങ്ങളും**
		N	**സമീകൃതാഹാരങ്ങൾ**<And co-ordination/Future model suffix>
			സമീകൃതാഹാരം<Plural suffix><And co-ordination/Future model suffix>
		N	***സമീകൃതാഹാരത്തിൽ***
			സമീകൃതാഹാരം <Locative case suffix>
പ്രോട്ടീൻ	Syllable structure	:	പ്രോ ട്ടീ ൻ
	Etymology		ഇംഗ്ലീഷ്
	Grammatical category		നാമം
	Root form		പ്രോട്ടീൻ
	Inflected forms	N	**പ്രോട്ടീനിന്റെ**
			പ്രോട്ടീൻ <Genetive case suffix>
		DFW	**പ്രോട്ടീനിലാണോ**
		DFW	**പ്രോട്ടീനിലാണ്** <Or co-ordination/Interrogative suffix>

		N	**പ്രോട്ടീനിൽ**<Equation/Finite verb suffix> <Or co-ordination/Interrogative suffix> പ്രോട്ടീൻ <Locative case suffix><Equation/Finite verb suffix> <Or co-ordination/Interrogative suffix>
		N	**പ്രോട്ടീനുകളുടെ**
		N	**പ്രോട്ടീനുകൾ**<genitive case suffix>
			പ്രോട്ടീൻ <plural suffix><genitive case suffix>
			പ്രോട്ടീൻ
			Root word
മിനറൽ	Syllable structure	:	മി ന റ ൽ
	Etymology		ഇംഗ്ലീഷ്
	Grammatical category		നാമം
	Root form		മിനറൽ
	Inflected forms	N	**മിനറലുകളിൽ**
		N	**മിനറലുകൾ**<Locative case suffix>
			മിനറൽ<Plural suffix><Locative case suffix>
		DFW	**മിനറലിന്റെയും**
		N	**മിനറലിന്റെ**<And co-ordination/Future model suffix>
			മിനറൽ<Genitive case suffix><And co-ordination/Future model suffix>
		DFW	**മിനറലിലാണോ**
		DFW	**മിനറലിലാണ്**<Or co-ordination/Interrogative suffix>
		N	**മിനറലിൽ**<Equation/Finite verb suffix><Or co-ordination/Interrogartive suffix>
			മിനറൽ <Locative case suffix><Equation/finite verb suffix>
			<Or co-ordination/Interrogative suffix>
			മിനറൽ
			Root word
വിറ്റാമിൻ	Syllable structure		വി റ്റാ മി ൻ
	Etymology		ഇംഗ്ലീഷ്
	Grammatical category		നാമം
	Root form		വിറ്റാമിൻ
	Inflected forms	N	**വിറ്റാമിനില്ലാതെ**
			വിറ്റാമിൻ<Negative commitative suffix>
		DFW	**വിറ്റാമിനിലാണോ**
		DFW	**വിറ്റാമിനിലാണ്**<Or co-ordination/Interrogative suffix>

		N	വിറ്റാമിനിൽ<Equation/Finite verb suffix><Or co-ordination/Interrogative suffix>
			വിറ്റാമിൻ <Locative case suffix><Equation/Finite verb suffix> <Or co-ordination/Interrogative suffix>
			വിറ്റാമിൻ
			Root word
ഇവ	Syllable structure		ഇ വ
	Etymology		മലയാളം
	Grammatical category		നാമം-സർവ്വനാമം
	Root form		ഇവ
	Inflected forms	N	ഇവയുടെ
			ഇവ <Genitive case suffix>
		N	ഇവയ്ക്ക്
			ഇവ <Dative case suffix>
		N	ഇവയിലൂടെ
		N	ഇവയിൽ<Path locative suffix>
			ഇവ<Locative case suffix><Path locative suffix>
			ഇവ
			Root word
അടങ്ങിയിരിക്കുന്ന	Syllable structure		അ ട ങ്ങി യി രി ക്ക ന്ന
	Etymology		മലയാളം
	Grammatical category		ക്രിയ
	Root form		not identified
	Inflected forms	V	അടങ്ങിയിരുന്നിട്ട്
		V	അടങ്ങിയിരിക്കുക
		V	അടങ്ങിയിരിക്കാത്ത്
		V	അടങ്ങിയിരിക്കുവാൻ
		V	അടങ്ങിയിരിക്കുമ്പോൾ
		V	അടങ്ങിയിരുന്ന
		V	അടങ്ങിയിരുന്നാളു
		V	*അടങ്ങിയിരിക്കുന്ന*

Figure 6.1: Sample Computational Lexicon generated

The words seen in the lexical entry are the words which the computational lexicon considers as the head words in the lexicon. They are the root words and verb words in the corpus. The linguistic information about each lexical entry such as syllable structure, etymology, grammatical category, root form and the inflected forms are systematically shown in the figure. The knowledge about the syllable structure of the word is helpful for a non-native person to pronounce the word easily. Space is used to mark the set of syllables of the corresponding words. This will also help to understand how the syllables are distributed in a word. Etymology denotes whether the word is English or Malayalam. This information helps to know more about the origin of the word. It also helps to know about the spelling or the orthographic pattern when an English word is transliterated to Malayalam. One of the major challenges faced by the Malayalam language computing tools is the lack of automatic identification of out-of-vocabulary words such as the person name, place name, abbreviations etc. Using this computational lexicon, if the word is identified as foreign word or Malayalam word, efficiency of language tools will increase drastically. If the knowledge about the source language is obtained, transliteration and translation become more accurate.

The grammatical category of the root word, shown in Figure 6.1 denotes whether the root word is a noun or a verb. This is the most important information for retrieving its meaning. Apart from the basic grammatical category, sub categories like pronouns are also shown with the word as in the case of *iva* (ഇവ) in the Computational Lexicon. Postpositions are also indicated if present in the sentence. Root form of the word is the derived root word from the inflected forms. This information is useful for concordance, key-word-in-context, frequency count etc. One can infer all the inflected forms occurred in the corpus related with the particular lexical entry with their grammatical category. N shows that the inflected form of the word is a noun word, DFW as dual functional word and V as verb word. In this example, *samiikRRthaahaaraththil~* (സമീകൃതാഹാരത്തിൽ) is appended to the existing list of 9 similar words when the input corpus is fed to the system. All the 12 words which can inflect with the root word *samiikRRthaahaaraM~* (സമീകൃതാഹാരം) can also be seen in the computational lexicon. Apart from the list of inflected words, all suffixes agglutinated with the word according to the position they take, words obtained after removing the suffixes, name of suffixes, can also be infered. The different inflected verb forms of *atangngiyirikkunnu* (അടങ്ങിയിരിക്കുന്നു) gives the possible verb combination of similar words seen in the corpus.

From this computational lexicon, special words having some linguistic peculiarities like words having the highest number of suffixes, words having maximum dual functional suffixes, words having maximum syllables, words having single syllable, most frequently occurring foreign word etc. can be easily retrieved.

6.2 Objective Evaluation

An important recent development in NLP has been the use of much more rigorous standards for the evaluation of NLP systems. The evaluation measures [Manning et al. 1999] like Precision, Recall and F-measure are often used to evaluate the efficiency and effectiveness of retrieval process. These performance matrices are used for an objective evaluation of the performance of the developed system. The matrices are defined as

1. Precision (P) = $\frac{\text{Number of words correctly recognized as a category}}{\text{Total number of words recognized as that category}}$

= $\frac{tp}{tp+fp}$

where tp is the true positives which is the number of words correctly recognized and fp is the false positives/false acceptances which are the wrongly recognized words as that category.

2. Recall (R) = $\frac{\text{Number of words correctly recognized as a category}}{\text{Total number of words of that category in the test set}}$

= $\frac{tp}{tp+fn}$

where fn is the false negatives/false rejections which are correct words in a category but are not recognized as that category by the system.

3. F-measure (F) = 2*P*R/(R+P) where P is the Precision and R is the Recall

6.2.1 Objective Performance Analysis of RWI Module

A small corpus containing 24,055 words from the domain of health, film, technology, law, lyrics of songs and religion were used for testing the performance of the RWI. The typographical forms and foreign language words, which have less relevance in this work are removed from the test corpus and the remaining 23,045 words were used to test the system. RWI program will identify the words in the corpus and gives the root form of the word with its grammatical category such as noun, verb or dual functional.

Noun words are words having noun suffixes. It was observed that the maximum number of noun suffixes agglutinated with a word is three in the test corpus. 137 words were identified as noun words having three noun suffixes. Majority of noun words in the corpus have only one noun suffix.

Dual functional words are words having dual functional suffixes. The maximum number of dual functional suffixes agglutinated with a word is three and 77 words are having three dual functional suffixes.

Root words are words which are obtained by removing noun suffixes and dual functional suffixes. They also include words without having any suffixes. Pronouns, postpositions, certain suffixes and words in the Exceptional_dictionary, are also considered as root words.

Verb words are words having verb suffix. Lexical entries are the root words and verb words. Performance metrics obtained for root words, noun words, dual functional words, verb words, and total lexical entries in the lexicon are given as follows.

I. Noun words

Total number of noun words in the test set = 7375

Total noun words processed by the RWI system = 6975

Number of noun words correctly recognized = 6612

Precision	=	94.79%
Recall	=	89.64%
F-measure	=	92.14%

II. Dual Functional words

Total number of dual functional words in the test set = 4728

Total dual functional words processed by the RWI system = 4463

Number of dual functional words correctly recognized = 4359

Precision	=	97.66%
Recall	=	92.19%
F-measure	=	94.84%

III. Root words

Total number of root words in the test set = 19,575

Total root words processed by the RWI system = 19,500

Number of root words correctly recognized = 18,997

Precision	=	97.42 %
Recall	=	97.04 %
F-measure	=	97.22%

IV. Verb words

Total number of verb words in the test set = 2805

Total verb words processed by the RWI system = 2493

Number of verb words correctly recognized = 2449

Precision	=	98.23%
Recall	=	87.3%
F-measure	=	92.43%

V. Lexical entries

Total number of lexical entries in the test set = 22,380

Total number of lexical entries identified using RWI program = 21,993

Total number of lexical entries correctly recognized= 21,446

Precision	=	97.51%
Recall	=	95.8%

The system showed a high percentage of precision and recall values. A small percentage of words were unprocessed and a small percentage was processed wrongly. Some of the reasons are already discussed in section 3.3 and section 4.4.

When evaluating the overall performance of the RWI system, verb words are having a high precision. Its false positive is less. That is, words which are wrongly selected as verb category is less. So one can conclude that the verb morphology used in RWI system helped to recognize the majority of processed words correctly.

The verb words gave the lowest recall performance when compared to the recall obtained for noun words, dual functional words and root words. That is, false negative is high for verb words. This shows that many of the words in the verb category are not recognized as verb by the RWI system. So the recall obtained for lexical entries also showed a decreased value. Addition of more verb suffixes to the existing list of verb suffixes, as listed in Table 3.5 can increase the recall rate of the system.

In the case of noun words, the precision obtained is only 94.79%. This is because certain words have phonemes seen at the end of the word similar to noun suffix. The system will wrongly select such words in noun category, instead of storing them as root words. So false positives increases and precision decreases. In order to decrease false positives, foreign words included in them can be identified using the proposed system and store them as root words. This will helps to increase the precision of both root words and noun words.

The proposed system is an efficient system and comparable to the different existing morphological analysers. A morphological analyser developed for Malayalam verbs [Saranya 2008] identifies a word having suffix /-uM~/ only as a verb word. But in certain cases, this future tense marker /-uM~/ may act as an And co-ordination suffix. In our approach the system will identify a word with /-uM~/ suffix as a dual functional suffix showing that the word with this suffix can act as a verb or as a noun. The morphological analyser developed using Apertium toolbox [Vinod et al. 2011] need a morphological dictionary. In our system, there is no need to have an inbuilt dictionary or manual assistance for collecting words. More number of noun words and verb words are identified using our system. They used only 1000 words for evaluation purpose. We used around 24000 words for this purpose. The efficiency of their morphological analyser depends on the manually built dictionary. Our system will automatically enrich the lexical list by giving larger corpus. The system designed and implemented for automatic creation of an Arabic lexicon [Al-Shalabi et al. 2004] identifies the gender of the word. Even though our system did not tried to find the gender of the word, more linguistic information like syllabic structure, etymology of words are automatically identified.

6.2.2 Objective Performance Analysis of Foreign Word Identification Module

Root words obtained from the RWI are fed to the system for deciding whether the word is a foreign word or a Malayalam word. The work is limited by the identification of English words only. The test corpus contains 4110 English words, 51 Tamil words (only film names are counted), 17 Arabic words and one Telugu word in addition to Malayalam words. The proposed system identified Tamil words as Malayalam words and Arabic words as English words and Telugu words as un-decidable word. The performance matrices of the foreign word identification module are as follows.

Total number of root words taken for experiment = 18800
Total English words in test set = 4110
Number of English words processed by the system= 4800
Number of English words correctly recognized = 3000

Precision	=	62.5%
Recall	=	72.99%
F-measure	=	67.75%

The system wrongly showed certain English words as Malayalam words. Wrong prediction of compound words, words from other languages etc. caused fp to increase and precision to decrease. If a mono syllable word is not having an equivalent phonemic pattern in any of the training corpus, the system cannot identify its category. This causes fn to increase and resulted in a poor recall. Other reasons for the decrease in these values have already been discussed in section 5.5.

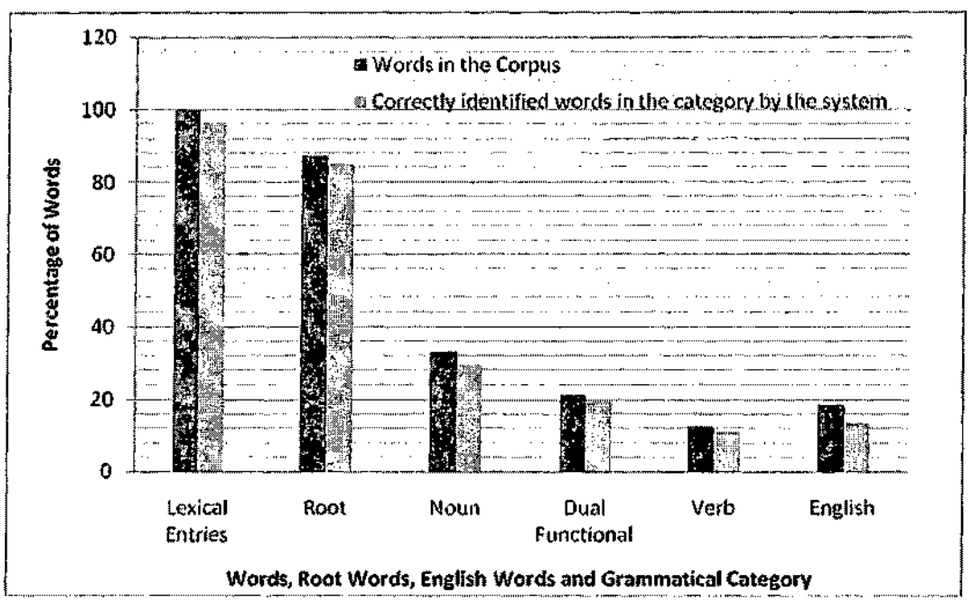


Figure 6.2: Comparison of actual lexical category (Ground Truth) in the corpus and that identified by the system

Figure 6.2 shows the difference in the actual number of words identified in each category in the corpus with the output generated by the system. It can be seen that about 96% of the lexical items in the corpus are included in the computational lexicon. That is, majority of words in the vocabulary are correctly processed. The system also identified nearly 85% of root words. The percentage of root words will increase if the root form of verb words is also identified.

As seen in the Figure 6.2, only a small percentage of English words are in the test corpus because the corpus documents are taken from different domains such as technology, health, law, film, lyrics of songs and religion.

In the task of identification of foreign words, only a precision/recall of 62.5%/72.99% is obtained. The precision / recall obtained in identifying foreign words in Hebrew language is 80%/82% [Goldberg 2008] and in Korean language is 84.33%/92.01% [Kang et al. 2002]. They are using long test samples and many latest statistical techniques to improve the accuracy of identification like smoothing techniques as discussed in section 5.6. Without using such techniques, Hebrew language obtained a precision/recall value of 58.7%/64.9% and Korean language obtained a precision/recall value of 62.89%/70.02% only. These results are comparable to our results.

6.3 Application of the Developed Corpus Based Computational Lexicon

Corpus based computational lexicon which is automatically generated for Malayalam language, by the proposed system, has many advantages when compared with traditional lexicon and existing machine readable dictionaries. Some of the features of the developed system is that it can provide the linguistic information of Malayalam words, which can be processed by the computers such as syllabic structure, etymology, root form of the word and its grammatical category, various inflected forms of the root word, suffixes agglutinated with word, name of each suffixes, their position of occurrence and the grammatical category of each word obtained after removing the suffixes. These details about the Malayalam words can be used to develop certain language tools where a lexicon need as the lexical database. Methodologies discussed in this proposed work can be used to create much larger corpus and for generating larger lexical database. Such a lexicon having information about millions of words can be used for developing high entity tasks like machine translation, lexicographic modifications, coining of technical terminologies, developing transliterated dictionary, speech synthesis, studying foreign influence in a language [George 1972] etc. All these task need computational lexicon for efficient processing. The applications in related areas are briefly explained in the following section.

1. **Lexicography:** Lexicographers can use this lexicon to revise the words in the existing dictionaries, on-line vocabulary builders and traditional lexicon thereby including new words to them. Nowadays manually entered dictionaries are used for Malayalam computing tasks such as in Parts-Of-Speech tagging, chunking, Named Entity Recognition, identifying lexeme for building lexicon, developing transliterated Malayalam and English dictionaries, morphological analyser and generator, cross-lingual information retrieval system etc. They can be made more efficient by incorporating this computationally developed lexicon.

In addition to the details provided for a root word in the lexicon generated as above, the registers or subject domains, listed in Appendix II, can also be added to the word. It will provide the information about the domain/domains from which the word is seen in the corpus. The categorization of these registers can also be used as objects in clustering technique which place similar objects in the same group and dissimilar objects to different groups. Using statistical techniques like frequency count and concordance, information such as commonly used words, context of each lexical entry in the lexicon etc. can be retrieved by a lexicographer. This will help to extract the meaning of words.

2. **Aid for Grammarians:** The computational grammar developed here can be compared and evaluated with the existing grammars. It can also be used for testing hypothesis derived from various grammatical theories. It is known that the existing grammar fail to reflect the language in current use, since it varies day by day. Computational linguist can use the information obtained from this analysis to extend the study on grammar with more examples.

3. **Machine Translation:** The above approach of designing a computational lexicon can contribute to develop a full fledged translation system. In the stages of formalisation of translation equivalents, automatically identified morphological details can give more accurate results than manually tagged ones. By acquiring the meaning, in addition to the above details of words, one can directly use it for translation process. The process of translation also requires lexical mapping and grammatical mapping in which the words in the source language are translated to target language. If the source or target language is Malayalam one can utilize the lexical and grammatical information about the Malayalam words from this work. The lexical ambiguity and constituent mapping can be dissolved to some extent by the use of a corpus based lexicon.

During the process of Machine Translation using bilingual aligned corpus, the lexicon can be used to identify the equivalents in its target languages.

4. **Syntactic analysis:** A sentence parser structure consists of either morphological or grammatical relations, or both and a dictionary of items. In an understanding mode, the lexicon (dictionary of items) can be used to recognize words and embedded clauses, to construct kernel structure and to build transformational expressions. In a generating mode the lexicon can be used to consult when a noun or a verb phrase is built or when a connective transformation is applied or when a question is answered. So the above results can be used for the syntactic analysis of Malayalam language.

5. **Semantic and anaphoric annotation:** The grammatical information about words from the lexicon can be used for semantic and anaphoric annotations. Morph features required for annotation can be easily obtained from the proposed lexicon.

6. **Text to speech system:** An automatic detection of a shift from one language to another in written text is required for developing text to speech system. If the system is turned on in Malayalam mode and if it encounters an English word, the system can recognize that word as an English word and it can then automatically switch to English lexicon having its pronunciation (if the pronunciation of head words are also included with the Computational lexicon).

The foreign word identification module can generate separate collection of Malayalam root words and English root words. The list of them can be used in applications like corpus annotation, processing of parallel corpus, enhancing search engine, frequency calculation, concordance of words, lexical collocation, lemmatisation of English words, study on word formation, manual annotation, chunker, parser, study on transliteration and even in developing spell checkers and grammar checkers. Some of the areas where the separate list of English and Malayalam words used are discussed in the following section.

A. **Phonological analysis:** A comparative study on phonological segments (characters) of both native and foreign words can be used for understanding the restrictions of these segments in native and foreign words, restriction of characters in different positions of word (initial, medial, intermediate) etc. For example: /-pha/ (-ഫ) occurs only in English words. Statistical distribution of different segments and segment types (vowels, consonant, consonant clusters etc) helps to extract more information about Malayalam scripts. Statistical modelling of words using n-gram discussed in chapter V can be used for phonological analysis.

B. **Coining technical terminology:** From these lexicons it is easy to find out the frequency of foreign and native words in a document separately. If the frequency is high for a particular foreign word, one can conclude that it is the most suitable and appropriate term which the society prefer to use. Such information can be taken into consideration when coining technical terminologies and developing term banks and knowledge base systems.

C. **Automatic tagging of foreign words:** In many grammatical tag sets developed for Dravidian languages there are untagged words or out-of-vocabulary words (words which are not tagged with grammatical properties). Such untagged words include person name, item name, scientific terminologies, foreign words, technical words etc. From the above work, it is observed that many proper names and technical terms are automatically tagged as foreign words. The methodology implemented in the work can be used to trace such words. The work can be applied to other language pairs like English-Hindi, English-Tamil etc. for identifying the influence of English.

D. **Information from root form of words:** Words in its root form can be used to extract character level frequency, syllabic level frequency, word level frequency etc. The frequency of words in various lexical classes such as noun and verb can also be traced out. Such valuable information can be used for developing "usage-based dictionary".

Lexicographer can call each root word and develop concordance program to establish the relationship between co-occurring words by analysing their underlying mutual information. This helps to extract phrases, idioms and collocations and there distribution across text types more systematically so that one can implement WebCorp [w25] like application for Malayalam language.

One can customise this lexical database and study can be conducted on particular areas of interest. For example, the list of noun words can be used for studying the name of places. Students, researchers and textbook designers can use this database for collecting the vocabulary of Malayalam according to their use.

The identified root words can be directly used to develop morphological generator which will generate the inflectional and derivational variants of words. Using the list of suffixes and rules, more accurate results can be obtained for the morphological generator than that manually compiled.

E. **Influence of English in native language:** Comparative study on native and foreign words can reveal the use of *Tatsama* and *Tadbhava* words [Joseph 1984] [Girish 2005] in the language. Frequency of such data can reveal how rapidly the foreign words invades the native language. Such a study is having importance in this era of globalization and localization.

F. **Developing bilingual dictionary:** The vocabulary list of both languages can be used in the process of developing bilingual dictionary.

G. **Spell checkers and Grammar checkers:** The phonotactic distributions of native and foreign words are analysed using n-gram statistical method. By analysing the list of n-gram it is possible to trace which phonemic combination will occur in the words in a language and which will not occur. n-gram can also provide the contextual details of a phoneme. Such knowledge can contribute much when developing language tools for checking the spelling of words. When developing grammar checkers, computational lexicon itself can be considered as a back end processor so that details like root words, its grammatical category, suffixes, the different categories in which a word can function (as a noun only, as a verb only, as a noun and a verb) etc. can be systematically retrieved. Computational grammar and morphophonemic rules that have been developed can be directly used in grammar checkers.

H. **Transliterated dictionaries:** The society is now facing a crucial problem of how to transliterate a foreign word when it is written in its native form. Such a dictionary is not available to clear the spelling of foreign words. To resolve this problem, a linguistic analysis on phonemic transformation is required. The work helps to develop a transliteration dictionary and a converter which converts an English word to its equivalent Malayalam transliterated form.

I. **Pronunciation dictionary:** A computational linguist can develop a pronunciation dictionary using the head words in the lexicon by including the pronunciation of each syllable of word so that the user can understand how they are pronounced. The lexicon provides all the orthographic variations of a particular word. This data helps to extract the pronunciation-orthography pairs useful for speech research.

It can be observed from the evaluation of the proposed system for the generation of a computational lexicon that the performance of the system is comparable to similar works in literature for other languages and it outperform the existing manual technologies for making dictionary in Malayalam. The chapter also discussed some of the applications of the proposed computational lexicon. Many further applications can be explained using a computational lexicon which can contribute to automatic language processing.

Chapter VII

CONCLUSION

This work aimed to automatically develop a corpus based computational lexicon for Malayalam language. A fully automatic system is developed which extracts words in the documents collected from the World Wide Web and systematically stores it in a lexicon along with its grammatical category. It also classifies words by their etymological details which are considered as one of the major task in the field of Language Technology.

A Malayalam corpus is automatically created for developing the computational lexicon. Using a web crawler, documents from World Wide Web are automatically stored into respective subject domains according to certain keywords. Computational grammar for the lexicon is designed. The suffixes agglutinated with various grammatical categories in Malayalam are analysed and stored in respective files, for supporting the proposed system. A study on morphophonemic change in root words with suffix agglutination is done and morphophonemic rules are derived. These rules are used to separate the root form of the words and to decide the grammatical categories of words.

The system also identifies the English words in the Malayalam corpus. A back end processor is developed for this, which uses a training corpus containing the n-gram models for English and Malayalam language. This language model is compared with the n-gram model of the root word and the system will identify whether the word is native or foreign.

As expected, the generated lexicon provides the linguistic information such as the syllable structure, etymology, root form of the word, syntactic category, inflected forms of the root word, suffixes agglutinated with the word, name of all suffixes, word obtained after removing each suffix, etc. Since the World Wide Web is used as the resource for corpus creation, this developed lexicon is a repository for commonly used words, technical words, colloquial words, new words and commonly occurring foreign words in Malayalam documents.

An objective evaluation of the system is also done using the statistical measures like precision, recall and F-measure for different grammatical category. It is found that for most of the grammatical categories of words, more than 90% values are obtained for these measures. In the case of English words, the value for these measures is found to be around 70% only, due to inadequate training data.

The inference obtained from the above work has already been discussed in corresponding chapters. Some of them are

- The suffix attached with the root form decides the grammatical category of the root word
- It is sufficient to identify whether the root word is a noun or a verb. Other grammatical categories like adjective, adverb etc. can be identified from the computational lexicon since it contains all possible inflected forms of the root word.
- The manually developed systems for corpus creation, syllabic tagging, parts-of-speech tagging, root word identification and foreign word identification were successfully done with the proposed system automatically.
- Morphophonemic rules helped to derive the root words or head words in the lexicon. Grammatical category of the word doesn't play any role in this rule.
- Using the proposed system, which root word will inflect with which suffix and which root word will not, can be identified.
- Distribution of phonemes helped to trace the ethmology of words.
- Distribution of phonemic patterns of both Malayalam and English words can be studied using statistical language modelling using n-gram.
- Different spelling system exists for foreign words than native words.
- Technical terms, place names, person name etc. which are borrowed from English language can be obtained from the list of foreign words.
- Many wrongly processed words obtained from RWI module can be processed correctly if it is a foreign word.
- The output obtained from the proposed system can be used to increase the training corpus therby making the whole system more efficient.

This developed system for automatic lexicon generation incorporates the corpus based, rule based and statistical approaches. Since a large amount of automatically retrieved documents are used, the generated lexicon is more informative than that generated from a small amount of manually annotated information. This automatic system can incrementally update the lexicon using richer corpora in future. Because of the lack of any published documents in Malayalam language for automatic generation of a corpus based computational lexicon with ethymological details, this work can be treated as the first of its kind, to the best of our knowledge.

Further improvements of the system are possible in the different stages, and a few suggestions for the same are given in the respective chapters on each of the stages of the whole system.

LIST OF PUBLICATIONS

1 Subhash, Meera, Shanavas, S.A. and Wilscy, M., "A Statistical Identification of English Loan Words in Malayalam Documents", In the Proceedings of ICON-2011: 9th International Conference on Natural Language Processing", Macmillan Publishers-Advanced Research Series, India, Pp 91-95, 2011.

2 Subhash, Meera, Wilscy, M. and Shanavas, S.A., "A Rule Based Approach for Root Word Identification in Malayalam Language", International Journal of Computer Science and Information Technology (IJCSIT), Academey and Industry Research Collaboration Centre (AIRCC), Vol 4(3), Pp 159-166, June 2012.

Papers Presented in Seminars and Conferences

Paper Presented	Name of Seminar	Conducted By	Year
The Sapir-Whorf Hypothesis and Malayalam Computing	National seminar on Recent Trends in Philosophical and Linguistic Theories	Department of Linguistics and Department of Philosophy, University of Kerala, Trivandrum	2008
Identification of Clitics in Malayalam- A corpus Based Approach (Best Paper Award-2009)	38th All India Conference of Dravidian Linguists	Dravidian Linguistics Association, Trivandrum	2009
Lexical information extraction from a Malayalam corpus using statistical language modelling	37th All India Conference of Dravidian Linguists	Dravidian Linguistics Association, Trivandrum	2009
Creation of Corpus	Workshop on an introduction to Natural Language Processing	Linguistic Data Consortium for Indian Languages, Central Institute of Indian languages, Mysore and Department of Linguistics, Kerala University, Trivandrum	2009
Linguistic Analysis Methods	Workshop on an introduction to Natural Language Processing	Linguistic Data Consortium for Indian Languages, Central Institute of Indian languages, Mysore and Department of Linguistics, Kerala University, Trivandrum	2010

Resolving Language Barrier through Information Communication and Technology	97th Indian Science Congress	Indian Science Congress Association, University of Kerala,Trivandrum	2010
Manuscript – A resource for Corpus- Image to Text	National Seminar on Problems and Prospects of manuscriptology	Oriental Research Institute and Manuscripts Library, University of Kerala, Trivandrum	2010
Identification of Foreign words in Malayalam Documents – A computational Approach	International seminar on Malayalam Language and Globalisation	Southern Regional Language Centre, CIIL, Mysore and Department of Linguistics, University of Kerala, Trivandrum	2010
Malayalam Computing	Short term course in Language Technology	University Grants Commission, Academic staff college, University of Kerala, Trivandrum	2010
Techniques for Grammatical Information Retrieval from a Malayalam corpus	National Seminar on Malayalam grammatical theories- Tradition to Present	State Institute of Languages and Department of Linguistics and Malayalam, TRCML, Trivandrum	2010
Corpus- A Tool for Translation	National Seminar on Translation and Translation studie	Centre for Translation and Translation Studies, Department of Hindi, University of Kerala, Trivandrum	2010
Integration of Corpus in Language Testing and Evaluation	National Seminar on Evaluation of Language Teaching in Secondary schools	Centre for Testing and Evaluation, CIIL, Mysore and Department of Linguistics, University of Kerala, Trivandrum	2011
Corpus Based Lexicon- A Repository for Language Learners	National Seminar on Language Acquisition/Learning	Southern Regional Language Centre, CIIL, Mysore and Department of Linguistics, Trivandrum	2011
A Statistical Identification of English Loan Words in Malayalam Documents	9th International Conference on Natural Language Processing (ICON-2011)	NLP Association, India; IIIT Hyderabad; LDC-IL, CIIL Mysore; and AU-KBC Research Centre, Anna University,Chennai	2011

Appendix I

List of sources available for data collection

1. Educational Book (School / college / technical education / professional/ higher education books)
2. Reference Books (subject/entertainment books)
3. World Wide Web
4. Literary Works- Novels, Drama, short story etc.
5. Biographies & Essays
6. Religious Scripts
7. Encyclopaedia/Dictionary/lexicon
8. Autobiography
9. Press Reportage (Newspaper, Magazines, Journals, Periodicals, Pamphlets, Manifestos, Editorials, Press Reviews)
10. Cinema/Drama Script
11. Personal Letters (from different age groups)
12. Diary Personal
13. Media (Television, Radio)
14. Advertisements
15. Notices and brochures
16. Medical Reports
17. Government leaflets
18. Quiz Questionnaire
19. Manuscripts
20. Handwritten Wills
21. Inscriptions
22. Narrations
23. Discourse
24. Account Registers
25. Catalogues, almanacs, recipes, ceremonial materials, addresses, publication titles, product labels, registration plates, maps (eg: dialect map), identity cards, tickets, application forms (for obtaining entity of person, places, objects etc).
26. Written Statements of Arguments, Dialogues, Monologues, Conversations, Discussions (Formal / Informal), Interviews, Commentary, Prepared Oration

Each register will have the details such as:

- Branches of that particular domain (eg: physics: quantum physics):
- Information about each branches of physics
- History, philosophy about physics
- Inventions, discoveries, inventors, scientific instruments, prizes, awards
- Definitions, Laws, theories, hypothesis, phenomenon related to physics
- Clark table, measurements, standard value, details of lab experiments & apparatus
- Physics of toys, physics of automation (Robert, car,crane..), physics of sports, materials in use.
- Influence of physics to the society
- Audio & visual datas about physics
- News, literary works, publications, conferences,
- Text books, bibliographies about each domain.

APPENDIX II

Taxonomy of resigers for developing Malayalam corpus

(വിഷയവ്യാപ്തി)

The hierachical list of domains provides highly systematic and effective means for documenttation storage. The list is exhaustive and can be extended. Section 2.4 discussed about the registers.

1. പ്രകൃതിശാസ്ത്രം (Natural Science)
   - 1.1 കാർഷികം
     - 1.1.1 കൃഷികൾ
       - 1.1.1.1 ഉദ്യാനകൃഷി
       - 1.1.1.2 ഔഷധകൃഷി
       - 1.1.1.3 പഴകൃഷി
   - 1.2 ജീവശാസ്ത്രം
     - 1.2.1 ശാഖകൾ
       - 1.2.1.1 ജനിതകശാസ്ത്രം
       - 1.2.1.2 ജീവതത്ത്വശാസ്ത്രം
       - 1.2.1.3 പായൽ പഠനശാസ്ത്രം
       - 1.2.1.4 ബാക്ടീരിയാപഠനം
       - 1.2.1.5 വൈറസ് പഠനം
       - 1.2.1.6 സാമൂഹികജീവശാസ്ത്രം
       - 1.2.1.7 ഭൗമാതീത ജീവപഠനം
   - 1.3 ജന്തുശാസ്ത്രം
     - 1.3.1 മൃഗങ്ങൾ
     - 1.3.2 പക്ഷികൾ
     - 1.3.3 മത്സ്യങ്ങൾ
     - 1.3.4 പ്രാണികൾ
     - 1.3.5 ഇഴജന്തുക്കൾ
     - 1.3.6 കീടങ്ങൾ
     - 1.3.7 ഷഡ്പദങ്ങൾ
     - 1.3.8 ജന്തുരസതന്ത്രം
   - 1.4 സസ്യശാസ്ത്രം
     - 1.4.1 സുഗന്ധവ്യഞ്ജനങ്ങൾ
     - 1.4.2 പച്ചക്കറികൾ
     - 1.4.3 ഇലവർഗ്ഗങ്ങൾ
     - 1.4.4 കിഴങ്ങുവർഗ്ഗങ്ങൾ
     - 1.4.5 പഴവർഗ്ഗങ്ങൾ
     - 1.4.6 ഔഷധസസ്യങ്ങൾ
     - 1.4.7 വൃക്ഷങ്ങൾ
     - 1.4.8 ധാന്യ-പരിപ്പ്- കുരുക്കൾ
     - 1.4.9 കുമിളുകൾ

3.8.2 ഗാനങ്ങൾ

3.8.2.1 പ്രണയഗാനം

3.8.2.2 സിനിമ/ നാടക/സീരിയൽ ഗാനം

3.8.2.3 ആത്മീയഗാനം

3.8.3 കാവ്യങ്ങൾ

3.8.3.1 വിലാപകാവ്യം

3.8.3.2 മഹാകാവ്യം

3.8.4 കവിതകൾ

3.8.5 പാട്ടുകൾ

3.8.5.1 നാടൻപാട്ട്

3.8.6 മറ്റു ഗാനവിഭാഗങ്ങൾ

3.8.6.1 ഗീതം

3.8.6.2 പദ്യം

3.8.6.3 സംഗീതം

3.8.6.3.1 സംഗീതോപകരണശാസ്ത്രം

3.9 ഹസ്തലിഖിതശാസ്ത്രം

3.10 ശിലാലേഖനവിദ്യ

4 സാമ്പത്തികശാസ്ത്രം- വ്യാപാരം (Finance - Marketing)

4.1 വാണിഭം

4.1.1 ശാഖകൾ

4.2 വാണിജ്യം

4.2.1 ശാഖകൾ

4.3 വ്യവസായം

4.3.1 ശാഖകൾ

4.4 ബാങ്കിങ്

4.4.1 ശാഖകൾ

4.5 സാമ്പത്തികം

4.5.1 ശാഖകൾ

4.5.1.1 അർത്ഥശാസ്ത്രം

4.5.2 നികുതി

5 . ഭരണം (Management)

5.1 പ്രതിരോധം

5.2 വിജ്ഞാപനം

5.3 സർക്കാർ

5.3.1 ഭരണകൂടങ്ങൾ

5.4 അന്താരാഷ്ട്രം

5.5 ദേശീയം

5.6 ഭരണനടത്തിപ്പ്

Appendix III

Phonetic transliteration scheme [W9e]

സ്വരങ്ങൾ svarangngaL~							
അ	ആ	ഇ	ഈ	ഉ	ഊ	ഋ	
a	aa	i	ii	u	uu	RR	
	ാ	ി	ീ	ു	ൂ	ൃ	
എ	ഏ	ഐ	ഒ	ഓ	ഔ	അം	അഃ
e	ee	ai	o	oo	au	aM ~	aH
ൊ	ോ	ൈ	ൊ	ോ	ൌ	ം	ഃ
വ്യഞ്ജനങ്ങൾ vyanjjanangngaL~							
ക്	ഖ്	ഗ്	ഘ്	ങ്		ങ്ക	ന്റെ
k	kh	g	gh	ng		nk	nte
ച്	ഛ്	ജ്	ഝ്	ഞ്		ഠ	ക്ഷ
c	ch	j	jh	nj		TT	x
ട്	ഠ്	ഡ്	ഢ്	ണ്		ക്യ	വൈ
t	T	D	Dh	N		q	Y
ത്	ഥ്	ദ്	ധ്	ന്		ക്യ	ഞ്ച
th	thh	d	dh	n		Q	nch
പ്	ഫ്	ബ്	ഭ്	മ്			
p	ph	b	bh	m			
യ്	ര്	ല്	വ്	ശ്	ഷ്	സ്	ഹ്
y	r	l	v	S	Sh	s	h
ള്	ഴ്	റ്					
L	zh	R					
ചില്ലുകൾ cillukaL~							
ൻ	ൽ	ൾ	ർ	ൺ			
n~	l~	L~	R~	N~			
ചന്ദ്രകല candrakala							
്							
U							

REFERENCES

[Abdusalam et al. 2006] Abdusalam, F.A., Nwesri, S.M.M. and Tahaghoghi, Falk Scholer, "Capturing Out-of-Vocabulary Words in Arabic Text", In the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), Sydney, Pp 258–266, 2006.

[Achla et al. 2004] Achla, M. Raina, Mukerjee, A. and Shukla, Pushpraj, "A Unified Computational Lexicon for Hindi-English Code-switching", Sangal, Rajeev and Bendre, S.M. (eds.), Natural Language Processing, Allied Publishers, New Delhi, Pp 185-194, 2004.

[Allen 1995] Allen, J., "Natural Language Understanding", The Benjamin/Cummings Publication, Amsterdam, 1995.

[Al-Shalabi et al. 2004] Al-Shalabi, R. and Kanaan, R., "Constructing An Automatic Lexicon for Arabic Language", International Journal of Computing & Information Sciences, Vol 2(2), Pp 114- 128, August 2004.

[Al-Yahya et al. 2010] Al-Yahya, Maha, Hend, Al-Khalifa, Bahanshal, Alia and Al-Helwah, Nawal, "An Ontological Model for Representing Semantic Lexicons: An Application on Time Nouns in the Holy Quran", The Arabian Journal for Science and Engineering, Vol 35, 2010.

[Andrewskutty 1971] Andrewskutty, A.P., "Malayalam- An intensive Course, Dravidian Linguistic Association, Trivandrum, 1971.

[Antony et al. 2010] Antony, P.J., Mohan, S.P. and Soman K.P, "SVM Based Part of Speech Tagger for Malayalam", IEEE International Conference on Recent Trends in Information, Telecommunication and Computing, Pp 339-341, 2010.

[Antworth 1990] Antworth, E., "PC-KIMMO: A Two-level Processor for Morphological Analysis", Dallas, TX: Summer Institute of Linguistics, 1990.

[Asher et al. 1997] Asher, R.E. and Kumari, T.C., "Malayalam- Descriptive Grammars", Routledge, London and New York, 1997.

[Baker et al. 2008] Baker, K. and Brew, C., "Statistical Identification of English Loanwords in Korean Using Automatically Generated Training Data", In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), 2008.

[Baum 1972] Baum, L.E., "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic functions of a Markov Process", Inequalities, Pp 3-8, 1972.

[Beesley et al. 2003] Beesley, K. and Karttunen, L., "Finite State Morphology", CSLI Publications, Stanford, 2003.

[Beesley 1998] Beesley, Kenneth, "Language Identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text", In Languages at Crossroads, Proceedings of Annual Conference of the American Translators Association, Pp. 47-54, 1998.

[Bennett et al. 1986] Bennett, P.A., Johnson, R.L., McNaught, J., Pugh, J. Sagar, J.C. and Somers, H.L., "Multilingual Aspects of Information Technology", Gower, Hants, 1986.

[Bharati et al. 1997] Bharathi, Akshar, Chaitanya, V., Kulkarni, A.P. and Sangal, Rajeev, "Anusaraka Machine Translation in stages", Vivek: A Quarterly in Artificial Intelligence, Vol 10, Pp 22-25, 1997.

[Bharati et al. 1995] Bharathi, A., Chaitanya, V. and Sangal, R., "Natural Language Processing: A Paninian Perspective", Prentice Hall of India, NewDelhi, 1995.

[Bhatt et al. 2011] Bhatt, B. and Bhattacharyya, P., "IndoWordNet and Its Linking with Ontology", In Proceedings of ICON-2011: 9th International Conference on Natural Language Processing, Macmillan Publishers, India, 2011.

[Bhattacharyya et al. 2010] Bhattacharyya, P., Fellbaum, C. and Vossen P. (eds.), "Principles, Construction and Application of Multilingual Wordnets", In the Proceedings of the 5th Global WordNet Conference, Narosa Publishing House, Mumbai, 2010.

[Biber et al. 1998] Biber, D., Susan, C. and Randi, R., "Corpus Linguistics-Investigating Language Structure and Use", Cambridge University press, Cambridge, 1998.

[Bindu et al. 2011] Bindu, M.S and Mary Idicula, Sumam, "Named Entity Identifier for Malayalam using Linguistic Principles Employing Statistical Methods", International Journal of Computer Science Issues (IJCSI), Vol. 8 (5), September 2011.

[Bindu et al. 2011] Bindu, M.S and Mary Idicula, Sumam, "A Hybrid Model for Phrase Chunking Employing Artificial Immunity System and Rule Based Methods", International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 2 (4), October 2011.

[Bindu et al. 2009] Bindu.M.S, Mary Idicula, Sumam, "Analysis of Malayalam Compound Words and Implementation of a Compound Word Splitter Tool using Finite State Models", International Conference on Modeling and Simulation, India, Pp 1-3, 2009.

[Bista et al. 2007] Bista, S.K., Keshari, B., Khatiwada, L.P, Chitrakar, P. and Gurung, S., "Nepali Lexicon Development", PAN Localization, Working Papers 2004-2007, Pp 311-15, 2007.

[Botha et al. 2007] Botham, Reinier, Gerrit and Barnard, Etienne, "Factors that Affect the Accuracy of Text-based Language Identification", In Proceedings of PRASA 2007, Pp 7–10, 2007.

[Caldwell 1956] Caldwell, Robert, "A Comparative Grammar of the Dravidian or South-Indian Family of Languages", Kegan Paul, Trench, Trubner & Co. Ltd., London, 1956.

[Cantone et al. 2009] Cantone, K.F. and Witchin, N., "Code switching at the interface of language-specific lexicons and the Computational System", In the proceedings of the International Journal of Bilingualism, Pp 91-109, 2009.

[Chakrabarti et al. 2004] Chakrabarti, D. and Bhattacharyya, P., "Creation of English and Hindi Verb Hierarchies and their Application to Hindi WordNet Building and English-Hindi MT", Global WordNet Conference (GWC-2004), Czech Republic, January 2004.

[Chakrabarti et al. 1999] Chakrabarti, S., Berg, M.V.D. and Dom, B. "Focused Crawling: A New Approach to Topic-specific Web Resource Discovery", Elsevier Science B.V., 1999.

[Chen et al. 1999] Chen, S.F. and Goodman, J., "An Empirical Study of Smoothing Techniques for Language Modeling", Computer Speech & Language, Vol 13(4), Pp 359–393, 1999.

[Choudhury 2009] Choudhury B. R., "Dynamic Memory English Speaking Course", Fusion books, NewDelhi, 2009.

[Cowie 1989] Cowie, A. P., "Oxford Advanced Learner's Dictionary of Current English (4th ed)",Oxford University Press, Oxford, 1989.

[Crystal 2003] Crystal. D, "The Cambridge Encyclopedia of Language", Cambridge University Press, Cambridge, 2003.

[Damashek 1995] Damashek, M., "Gauging Similarity with n-grams: Language-independent categorization of text", Science, Vol 267, Pp 843–849, 1995.

[Dandapat et al. 2007] Dandapat, S., Sarkar, S. Basu, A., "Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario", In Proceedings of the Association for Computational Linguistic, Pp 221-224, 2007.

[Dash et al. 2003]Dash, Niladri Sekhar and Chaudhuri, B., B., "Relevance of Corpus in Language Research and Application", International Journal of Dravidian Linguistics, Vol 33(2), Pp 101-122, 2003.

[Dash 2005] Dash, Niladri Sekhar, "Corpus Linguistics and Language Technology", Mittal Publications, New Delhi, 2005.

[Dogru et al. 1999] Dogru, S. and Slagle, J.R., "Implementing a Semantic Lexicon", W. Tepfenhart & W. Cyre (Eds.) ,Conceptual Structures: Standards and Practices, In the Proceedings of 7th International Conference on Conceptual Structures, ICCS'99, Springer-Verlag, Berlin, Pp 154-67, 1999.

[Dunning 1994] Dunning, T. "Statistical Identification of Language" ,In the Technical Report MCCS-94-273, Computing Research, Lab, New Mexico State University, 1994.

[Elwell 2006] Elwell, R., "Finite State Methods for Bantu Verb Morphology", In the Proceedings of the Texas Linguistics Society X, Austin, 2006.

[Fellbaum 1998] Fellbaum, C. (ed), "WordNet An Electronic Lexical Database", The MIT Press, Cambridge, London, 1998.

[Fillmore et al. 2001] Fillmore, C. J., Wooters, C., and Baker, C. F., "Building a Large Lexical Databank which Provides Deep Semantics", In Proceedings of the Pacific Asian Conference on Language, Information and Computation, Hong Kong, 2001.

[Gangemi et al. 2006] Gangemi, A., Navigli, R. and Velardi, Poola, "The OntoWordNet project: Extension and Axiomatization of Conceptual Relatives in WordNet", In the proceedings of On the Move to meaningful Internet Systems (OTM2003), Springer-Verlag, Catania, Italy, Pp 820-838, 2003.

[Gangemi et al. 2010] Gangemi, A., Guarino, N., Masolo, C. & Oltramari, A., "Interfacing WordNet with DOLCE: Towards OntoWordNet", Cambridge University Press, London , 2010.

[Geethakumary 2002] Geethakumary, V., "A Contrastive Analysis of Hindi and Malayalam", Language in India, Vol 2, 2002.

[George 1972] George, K.M., "Western Influence on Malayalam Language and Literature, Sahitya Academy, NewDelhi, 1972.

[Girish 2005] Girish, P.M., "The Influence of English on Malayalam Language", Language in India, Vol 5, 2005.

[Goldberg et al. 2008] Goldberg, Y. and Elhadad. M, "Identification of Transliterated Foreign Words in Hebrew Script", In the Proceedings of the 9th international conference on Computational linguistics and intelligent text processing, Springer-Verlag Berlin, Heidelberg, Pp 466-477, 2008.

[Granger et al. 2003] Granger, S. and Tyson, S.P. (eds), "Extending the scope of Corpus-based Research: New Applications, New Challenges", Rodopi, Amsterdam, 2003.

[Grishman et al. 1994] Grishman, R., Macleod C., and Meyers A., "COMLEX Syntax: Building a Computational Lexicon", In Proceedings of Coling, Kyoto, 1994.

[Gupta et al. 2010] Gupta, R., Goyal, P. and Diwakar, S., "Transliteration among Indian Languages using WX Notation", In proceedings of KONVENS 2010, Saabrucken, Germany, September 2010.

[Halliday et al. 2004] Halliday, M. A. K., Teubert, Wolfgang, Yallop, C. and Cermakova, A. "Lexicology and Corpus Linguistics An Introduction", Continuum, London/New York, 2004.

[Hartmann 1983] Hartmann R.R.K. (ed), "Lexicography: Principle and Practice", Academic Press, London and New York: 1983.

[Hartmann et al. 1998] Hartmann, R. R. K., and James, G. "Dictionary of Lexicography", Routledge, London & NewYork, 1998.

[Horacek 2002] Horacek, H., "Aggregation with Strong Regularities and Alternatives", In the proceedings of Second International Natural Language Generation Conference, 2002.

[Ijaz et al. 2007] Ijaz, M., Hussain, S. "Corpus Based Urdu Lexicon Development", In the Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan, 2007.

[Itai et al. 2006] Itai, A., Wintner,S. and Yona.S, "A Computational Lexicon of Contemporary Hebrew". In Proceedings of LREC-2006, Genoa, Italy, May 2006.

[Jackson 2002] Jackson, Howard, "Lexicography: An Introduction", Routledge/Taylor & Francis, London & New York, 2002.

[Jeong et al. 1997] Jeong, K. Kwon,Y. and Myaeng, S.H., "The Effect of a Proper Handling of Foreign and English Words in Retrieving Korean Text," In the Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages1997, Tsukuba-shi, Japan, 1997.

[Jes´us Gim´enez et al. 2006] Jes´us Gim´enez and Llu´ıs M`arquez., "SVMTtool", Technical manual, Vol 3, August 2006.

[Jivani 2011] Jivani, A. J., "A Comparative Study of Stemming Algorithms", In the proceedings of International Journal of Computer Technology and Application, Vol 2, Pp 1930-1938, 2011.

[Joseph 1984] Joseph, P.M., "*Malayalathile Parakiya Padangal*", Kerala Bhasha Institute, Trivandrum, 1984.

[Jurafsky et al. 2000] Jurafsky, D., and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition", Prentice Hall Series in Artificial Intelligence, 2000.

[Kang et al. 2002] Kang, Byung-Juand and Choi, Key-Sun, "Effective Foreign Word Extraction for Korean Information Retrieval, Information Processing and Management, Vol 38, Pp 91–109, 2002.

[Karttunen et al. 1983] Karttunen, L. and Wittenburg, K., "A Two-Level Morphological Description of English", In the Proceedings of Advanced Computational Linguistics (ACL-83), 23rd Annual meeting, Pp 217-228, 1983.

[Kazakov et al. 1999] Kazakov, D., Manandhar, S. and Erjavec, T., "Learning Word Segmentation Rules for Tag Prediction", S. Dzeroski, S. & P. Flach (Eds.), Inductive Logic Programming, In the 9th International Workshop: ILP-99 Proceedings, Springer-Verlag, Berlin, Pp 152-161, 1999.

[Khaltar et al. 2006] Khaltar, B.O., Fujii, A. and Ishikawa, I., "Extracting Loanwords from Mongolian Corpora and producing a Japanese-Mongolian Bilingual Dictionary", In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Pp 657-664, 2006.

[Kilgarriff et al. 2003] Kilgarriff,A. and Grefenstette, G., "Introduction to the Special Issue on the Web as Corpus", International Journal of Corpus Linguistics, MIT Press, Vol 29(3), Pp 333-347, 2003.

[Klavans 1988] Klavans, J., "COMPLEX: A Computational Lexicon for Natural Language Systems", In the proceedings of Coling 1988, Pp 815-823, 1988.

[Kobayashi et al. 2000] Kobayashi, M. and Takeda, K., "Information Retrieval on the Web", Journal ACM Computing Surveys (CSUR), USA, Vol 32(2), Pp 144-173, June 2000.

[Kohavi 1995] Kohavi R.,"A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection", In the Proceedings of International Joint Conference on AI, Pp 1137–1145, 1995.

[Krishnamurti et al. 1985] Krishnamurti, B.H. and Gwynn, J.P.L., "A Grammar of Modern Telugu", Oxford University Press, Oxford 1985.

[Kroeger 2005] Kroeger, Paul, "Analyzing Grammar: An Introduction", Cambridge University Press, Cambridge, 2005.

[Krovetz 1993] Krovetz, R. "Viewing Morphology as an Inference Process", In the Proceedings of 16th ACM SIGIR Conference, Pittsburgh, Pp 191-202, 1993.

[Kruengkrai et al. 2005] Kruengkrai, C., Srichaivattana, P., Sornlertlamvanich, and Hitoshi Isahara, "Language Identification Based on String Kernels", In Proceedings of ISCIT-2005, Pp 896–899, 2005.

[Kulkarni et al. 2010] Kulkarni, M., Dangarikar, C., Kulkarni, I., Nanda, A. and Bhattacharyya P., "Introducing Sanskrit Wordnet", In the proceedings of the Global Wordnet Conference (GWC10), Mumbai, India, 2010.

[Kumar et al. 2011] Kumar, Dinesh and Rana, Prince, "Stemming of Punjabi Words by using Brute Force Technique", International Journal of Engineering Science and Technology (IJEST), Vol. 3(2), Pp 1351-1357, February 2011.

[Kumar et al. 2010] Kumar, Dinesh and Singh Josan, G., "Parts-Of-Speech Taggers for Morphologically Rich Indian Languages: A survey", International Journal of Computer Applications,Vol 6 (5), Pp 0975-8887, 2010.

[Lafferty et al.. 2001] Lafferty, John and Andrew McCallum, "Conditional Random Fields, Probabilistic models for segmenting and labeling sequence data. In the proceedings of ICML 2001, Pp 282-289, 2001.

[Lim et al. 2006] Lim, H.S., Nam, K., Park, K., and Cho, S. "A Computational Korean Lexical Access Model using Artificial Neural Network, "In Proceeding (ICIC'06) of International Conference on Computational Intelligence and Bioinformatics, Vol Part III , Pp 693-701, 2006.

[Litkowski 2005] Litkowski, K.C., "Computational Lexicons and Dictionaries", Encyclopedia of Language and Linguistics (2nd ed.), Elsevier Publishers, Oxford, 2005.

[Lovins 1968] Lovins, J.B., "Development of a Stemming Algorithm", Journal of Mechanical Translation and Computational Linguistics, Vol 11, Pp 22-31, 1968.

[Lovis et al. 1998] Lovis, C., Baud, R., Rassinoux, A.M., Michel, P.A.and Scherter, J.R., "Medical Dictionaries for Patient Encoding Systems: A Methodology", Artificial Intelligence in Medicine, Vol 14, Pp 201-214, 1998.

[Mair et al. 2000] Mair,C. and Hundt, M., "Corpus Linguistics and Linguistics Theory", Rodopi, Amsterdam, Atlanta, 2000.

[Majumder et al. 2007] Majumder, P., M. Mitra, S. K. Parui, G. Kole, P. Mitra and K. Datta., "YASS: Yet Another Suffix Stripper", ACM, Transaction on Information Systems, Vol 25(4), Pp 18-37, 2007.

[Makkai 1980] Makkai, A., "Theoretical and Practical Aspects of an Associative Lexicon for 20th Century English", in Zgusta (ed), Pp 125-46, 1980.

[Mallassery 1994] Mallassery, S.R., "Postpositions in a Dravidian language-A Transformational Analysis of Malayalam", Mittal Publications, NewDelhi, 1994.

[Mandeep et al. 2008] Mandeep, S., Gurpreet, L. and Shiv, S., "A Part-of-Speech Tagset for Grammar Checking of Punjabi", The Linguistic Journal, Vol 4(1), Pp 6-22, 2008.

[Manju et al. 2009] Manju K., Soumya S., Mary Idicula, Sumam, "Development of a POS Tagger for Malayalam - An Experience" , In Proceedings of the International Conference on Advances in Recent Technologies in Communication and Computing, Artcom, Pp709-713, 2009.

[Manning et al. 1999] Manning, Christopher D. and Schutze H., "Foundations of Statistical Natural Language Processing", The MIT Press, USA, 1999.

[Marcus 1986] Marcus, C. "Prolong Programming: Application for Database Systems, Expert Systems and Natural Language Systems", Addison-Wesley, NewYork, 1986.

[Mariani et al. 2008] Mariani, J. and Adda-Decker, M. "Automatic Language Identification", Language and Speech Processing, Pp 3-44, 2010.

[Mason 2000] Mason, Oliver, "Programming for Corpus Linguistics How to Do Text Analysis with Java", Edinburgh University Press, Edinburgh, 2000.

[McEnery et al. 2006] McEnery, T., Richard, X. and Yukio, T., "Corpus-based Language Studies-An Advanced Resource Book", Routledge, London & NewYork, 2006.

[Menaka et al. 2010] Menaka S., Sundar Ram, R., Vijay and Sobha, L., "Tamil Morphological Analyser", In the Proceedings of Knowledge Sharing Event on Morphological Analyzers and Generators, LDC-IL,CIIL, Mysore, Pp 1-18, 2010.

[Mitkove 2001] Mitkove, R., "The Oxford Handbook of Computational Linguistics", Oxford Press, Oxford, 2001.

[Mohanty et al. 2002] Mohanty, S., Balabantaray, R. C., "Oriya WordNet", In the proceedings of the 1st Global WordNet Conference, Central Institute of Indian Language , Mysore, 2002.

[Murthy 2004] Murthy, Kavi Narayana, "On Automatic Construction of a Thesaurus", In the proceedings of International Conference, CSLT-O-COCOSDA, NewDelhi, Vol 1, Pp 191-194, 2004.

[NairS 1982] Nair, Somesekharan, "*Nighantu Vijnanam*", State Institute of languages, Trivandrum, 1982.

[NairG 2008] Nair, Gopinathan, B., "Collected Papers on Malayalam Language and Linguistics", International School of Dravidian Linguistics, Trivandrum, 2008.

[Namboothiri 1999] Namboothiri, Sankaran K., "Intimate English-English Malayalam Dictionary with Computer Internet and Technical Terms", Computech Publishers, Trivandrum, 1999.

[Narendranath et al. 2011] Narendranath, R. and Paul, Soma., "A Data Driven Implementation of Malayalam Verb Morphology", In the Proceedings of ICON -2011: 9th international Conference on Natural Language Processing, Macmillan publications, India, Pp 189-194, 2011.

[Noikongka et al. 2007] Noikongka, D., Suktarachan, Mukda and Kawtrakul, Asanee, "Semi-Automatic Thai Computational Lexicon Construction", In the proceedings of KULEX, SNLP2007: Computational Lexicon, Concept Hierarchies, Word Classification, Thailand, 2007.

[Oard et al. 2001] Oard, S.W., Levow, G.A., and Cabezas, C.I, "CLEF Experiments at Maryland: Statistical Stemming and Back off Translation", In revised papers from the workshop of Cross-language Evaluation Forum on Cross language information retrieval and evaluation (CLEF), Springer, London, Pp 176-187, 2001.

[Oh et al. 1999] Oh, Jong-Hoon, and Choi, Key Sun, "Automatic Extraction of Technical Terminologies from Scientific Text Based on Hidden Markov Model", Paper presented in the Eleventh Hangual and Korean information Processing in Korean, 1999.

[Oh et al. 2002] Oh, Jong-Hoon and Choi, Key Sun, "An English- Korean Transliteration Model using Pronunciation and Contextual Rules", In COLING 2002, Pp 1–7, 2002.

[Ooi 1998] Ooi, V.B.Y., "Computer Corpus Lexicography", Edinburgh University Press, Edinburgh, 1998.

[Oostdijk et al. 1994] Oostdijk, N. and Hann, P. "Corpus Based Research into Language", Rodopi, Amsterdam, Atlanta, 1994.

[Paice 1977] Paice, C. D., "Information Retrieval and the Computer", Computer monographs, Macdonald and Jane's, London, 1977.

[Parakh et al. 2011] Parakh, M and Rajesha, N., "Developing Morphological Analyzers for Four Indian Languages using a Rule Based Affix Stripping Approach", Language in India, Special Volume: Problems of Parsing in Indian Languages, Pp 13-16, 2011.

[Parameswari 2011] Parameshwari, K., "An Implementation of APERTIUM Morphological Analyzer and Generator for Tamil", Language in India, Special Volume: Problems of Parsing in Indian Languages, Vol 11, 2011.

[Pavel et al. 2006] Pavel, Dewan Shahriar, Hossain, Sarkar, Iqbal, A., Shah, Faisal Muhammad and Khan, M., "Collaborative Lexicon Development for Bangla", In Proceedings of International Conference on Computer Processing of Bangla (ICCPB-2006), Dhaka, Bangladesh, 2006.

[Pedersen et al. 1998] Pedersen, T. and Bruce, R. "Knowledge Lean Word-sense Disambiguation", In the Proceedings of Fifteenth National Conference on Artificial Intelligence (AAAI-98), Tenth Conference on Innovative Applications of Artificial Intelligence, MIT Press, Pp 800-805, 1998.

[Pei et al. 1954] Pei, Mario A. and Frank, Gaynor, "A Dictionary of Linguistics", Philosophical Library, New York, 1954.

[Philip 2007] Philip, Strazny, "Encyclopedia of Linguistics", Kindle edition, Filtzroz Dearborn, Vol 2, 2007.

[Piepenbrock et al. 1995] Piepenbrock, R., Baayen, R. H. & Gulikers, L., "The CELEX Lexical Database (CD-ROM)", Philadelphia: PA: Linguistic Data Consortium, University of Pennsylvania, US, 1995.

[Pillai 1965] Pillai, Suranad Kunjan, "*Malayalam Lexicon*- a Comprehensive Malayalam-Malayalam-English Dictionary on Historical and Philological Principles", University of Kerala, Trivandrum, Vol I, Pp xi-xxx, 1965.

[Pillai 2011] Pillai, Sreekandeswaram. G. Padmanabha., "*Sbdathaaravali*", NBS, 2011(ed).

[Pinker 1995] Pinker, "The Language Instinct: The new science of language and mind", Penguin Books Ltd, Middlesex, England, 1995.

[Porter 1980] Porter, M. F., "An Algorithm for Suffix Stripping", Program, Vol 3, Pp 14-130, 1980

[Rajashekaran 2008] Rajashekharan, N., "Lexicographic Tradition in India", Language Technology Flash, VishwaBharat@tdil, Pp 74-77, 2008.

[Rajendran 2010] Rajendran, S., "Conceptual Lexicon for Knowledge Representation", In the proceedings of Tamil Internet -2010, Coimbatore, Pp 203-210, 2010.

[Ramanathan et al. 2003] Ramanathan, A. and Rao, D., "A Light Weight Stemmer for Hindi", In the proceedings of the 10th Conference of the European Chapter of the Association of Linguistics (EACL), on Computational Linguistics for South Asian Languages, Budapest, 2003.

[Rao et al. 2010] Rao, Uma Maheswar, G. and Parameshwari K., "On the Description of Morphological Data for Morphological Analyzers and Generators: A case of Telugu, Tamil and Kannada", In Morphological Analyzer and Generators, Mona Parakh (ed.), CIIL, Mysore, Pp 114-123, 2010.

[Rogati et al. 2003] Rogati, M., McCarley, S., and Yang, Y., "Unsupervised Learning of Arabic Stemming using a Parallel Corpus", In proceedings of the 41st annual meeting of the association for Computational Linguistics, E.Hinrichs and D.Roth, (eds), Pp 391-398, 2003.

[Rosner et al. 1998] Rosner, M. and Caruana, J. and Fabri, R., "Maltilex: A Computational Lexicon for Maltese", In proceedings of the workshop on Computational Aspects of Semantic Languages, ACL/COLING98, Pp 97-105, 1998.

[Sagarbas 2000] Sagarbas, "A Straight Forward Approach to Morphological Analysis and Synthesis", In the Proceeding of COMLEX 2000, Greece, 2000.

[Saranya 2008] Saranya S.K., "Morphological Analyzer for Malayalam Verbs", Unpublished M.Tech Thesis, Amrita School of Engineering, Coimbatore, 2008.

[Scheffczyk et al. 2006] Scheffczyk, J., Peasem, A. and Ellsworth, M., "Linking FrameNet to the Suggested Upper Merged Ontology", In the proceedings of the conference on Formal Ontology in Information Systems (FOIS), Baltimore, USA, 2006.

[Schryver et al. 2003] Schryver, D. and Maurice. G., "Lexicographers' Dreams in the Electronic-dictionary Age", International Journal of Lexicography, Oxford University Press, Vol 16(2), Pp 143-199, 2003.

[Schuler 2009] Schuler, K., "VerbNet Overview", in NAACL HLT, Tutorials, Boulder, Association for Computational Linguistics, Colorado, Pp 13–14, 2009.

[Selvaraj 2010] Selvaraj A., "Telugu Wordnet", In the proceedings of Global WordNet Conference (GWC10), Mumbai, India, 2010.

[Seshagiriprabhu 1983] Seshagiriprabhu, M., "*Vyakaranamitram*", (4thed.), Kerala Sahithya Academy, Thrissur, 1983.

[Sgarbas et al. 1995] Sgarbas, K., Fakotakis, N. and Kokkinakis, G., "Two Algorithms for Incremental Construction of Directed Acyclic Word Graphs", International Journal on Artificial Intelligence Tools, World Scientific, Vol 4(3), Pp 369--381, 1995.

[Shanavas 1996] Shanavas, S.A., "Structure of a Computational Lexicon of Malayalam", Unpublished PhD Thesis, Jawaharlal Nehru University, New Delhi, 1996.

[Shannon 1951] Shannon, Claude E., "Prediction and Entropy of Printed English", Bell system Technical Journal, Vol 30, Pp 50-64, 1951.

[Sibun et al. 1996] Sibun, Penelope and Reynar, J.C "Language Identification: Examining the issues", In Proceedings of SDAIR'96, Pp 125–135, 1996.

[Siivola et al. 2007] Siivola,V., Hirsim ki, T. and Virpioja, S. "On Growing and Pruning Kneser-Ney Smoothed n-gram Models", IEEE Transactions on Audio, Speech & Language Processing, Vol 15, Pp 1617–1624, 2007.

[Sinclair 1984] Sinclair, J.(ed), "Collins COBUILD English Dictionary (1st edition), HarperCollins, London, 1984.

[Sobha 1999] Sobha, L.,"Anaphora Resolution In Malayalam and Hindi" Unpublished Doctoral dissertation", Mahatma Gandhi University, Kottayam , Kerala, 1999.

[Stolcke 2002] Stolcke, A., "SRILM – An Extensible Language Modeling Toolkit", In Proceedings of ICSLP, Pp 901–904, 2002.

[Subhash 2006] Subhash, Meera, "Malayalam Corpus Processors Lexicalist Model", Unpublished M.Phil Dissertation, University of Kerala, 2006.

[Subhash 2010] Subhash, Meera, "Identification of Clitics- A Corpus Based Approach", Unpublished paper presented in 38th All India conference of Dravidian Linguistics, 2010.

[Sukhadeve et al. 2011] Sukhadeve, P.P and Dwivedi, S.K., "Advancement of Clinical Stemmer: A Study of Behavior of Medical Students" , Language in India, Special volume Problems of Parsing, Vol 11, Pp 45-51, 2011.

[Sundar et al. 2010] Sundar Ram R., Vijay and Sobha, L., "Noun Phrase Chunker using Finite State Automata for an Agglutinative Language", In the Proceedings of the Tamil Internet, Coimbatore, India, Pp 218 – 224, 2010.

[Teubert 2000] Teubert, W., "Corpus Linguistics-a partisan view", International Journal of Corpus Linguistics, Vol 4, Pp 1-16, 2000.

[Teubert et al. 2007] Teubert, W. and Krishnamurthy, R. "Corpus Linguistics- Critical Concepts in Linguistics", Vol I, Routledge Taylor&Francis, London & NewYork, 2007.

[Thiyagarajan et al. 2002] Thiyagarajan, S., Arulmozi, S., Rajendran, S. "Tamil WordNet", First Global WordNet Conference, CIIL, Mysore, 2002.

[Vaidhya et al. 2009] Vaidhya, Ashwini and Sharma, Dipti Misra, "Using Paradigms for certain Morphological Phenomena in Marathi", In the proceedings of 7th International Conference on NLP (ICON-2009), Macmilan Publishers, India Ltd., New Delhi, December 2009.

[Varma 1999] Varma, A. R. Rajaraja, "Keralapaniniiyam: A Treatise on Malayalam grammar", Translated by C. J. Roy, International School of Dravidian Linguistics", Trivandrum, 1999.

[Vatanen et al. 2010] Vatanen, T., Vyrynen, J.J. and Virpioja, S., "Language Identification of Short Text Segments with n-gram Models", In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Pp 3423–3430, 2010.

[Veerappan et al. 2011] Veerappan, R. and Antony, P. J., Saravanan,S. and Soman, K. P., "A Rule Based Kannada Morphological Analyzer and Generator using Finite State Transducer", In the proceedings of the International Journal of Computer Applications, Foundation of Computer Science, New York, USA, Vol 27, Pp 45-52, 2011.

[Vinod et al. 2011] Vinod P.M, Jayan, and Sulochana K. G., "Malayalam Morphological Analyser: A Hybrid Approach with Apertium Lttoolbox", In the Proceedings of ICON-2011:9th International Conference on Natural Language Processing, Macmillan publications, India, Pp 219-224, 2011.

[Wall et al. 2000] Wall, L., Christiansen, T. and Orwant, J., "Programming Perl", O'Reilly media, Shroff Publishers and Distributors Pvt. Ltd., New Delhi, 2000.

[Willett 1988] Willett, P., "Recent Trends in Hierarchic Document Clustering: A Critical Review", Information Processing & Management, Vol 24 (5), Pp 577-597, 1988.

[Wu et al. 2006] Wu, A. and Jiang, Z., "Statistically-Enhanced New Word Identification in a Rule-Based Chinese System", In Proceedings of the Second Chinese Language Processing Workshop, HKUST, Hong Kong, Pp 46-51, 2006.

[Xu et al. 1998] Xu, J. and Croft,W. B. "Corpus-Based Stemming using Co-occurrence of Word Variants", ACM Trans. Inf. Syst. Vol 16(1), Pp 61–81, 1998.

[Zanchetta et al. 2005] Zanchetta, Eros and Baroni, Marco, "Morph-it! A free corpus-based morphological resource for the Italian Language, In the Proceedings of Corpus Linguistics , Birmingham, UK, 2005.

[Zgusta 1971] Zgusta, Ladislav, "Manual of Lexicography", Prague, Hague : Mounton, 1971.

[Zgusta 1980] Zgusta, Ladislav, "Theory and Method in Lexicography: Western and Non-Western Perspectives", S. Caroline: Hornbeam Press, Columbia, 1980.

WEBLIOGRAPHY

[w1] Stevenson, S. "The Computational Lexicon", Computational Linguistics, Department of Computer Science, University of Toronto, Viewed on 10 November 2010, <www.cs.toronto.edu/compling/Courses/courses.html>

[w2] "WordNet-A Lexical Database for English",Princeton University, Viewed on 12 December 2011, < http://wordnet.princeton.edu/>

[w3] Hussain, K. H. and Suresh, P., "Meera Traditional Script: Malayalam font", Viewed on 02 December 2010, <http://savannah.nongnu.org/forum/forum.php?forum_id=5052>

[w4] "Apertium -Free/Open-Source Platform for Developing Rule-based Machine Translation System", Viewed on 09 February 2010 <http://wiki.apertium.org/wiki/Main_Page>

[w5] "Wikigranthasaala", Viewed on march 2009, < http://ml.wikisource.org/wiki/Indulekha>

[w6a] "Wikidictionary", Viewed on 20 December 2011, <http://ml.wiktionary.7val.com/wiki/>

[w6b] "Malayalam Dictionary", Viewed on 19 June 2010, <http://www.dictionary.tamilcube.com/malayalam-dictionary.aspx>

[w6c] "Online Trilingual Dictionary: English-Malayalam-Hindi Dictionary", Resource Centre for Indian Language Technology Solution, CDAC, Trivandrum, Viewed on 15 July 2009 <www.malayalamresourcecenter.org/mrc/ dictionary>

[w6d] "Mashithantu software group", Viewed on 08 March 2010, <http://www.dictionary.mashithantu.com/>

[w6e] "English- Malayalam Dictionary", Viewed on 09 May 2011 <http://malayalam.changathi.com/Dictionary.aspx>

[w7] "Amrita University, Coimbatore", Viewed on 09 January 2012 <http://www.amrita.edu/cen/ computational.php>

[w8] "The Technology and Resource Centre for Malayalam Language", Viewed on January 2010 <http://www.trcml. keralauniversity.edu/>

[w9a] "Centre for Linguistic Computing Keralam", Viewed on March 2012, http://www.clickeralam.org/

[w9b] "Multilingual Technologies", CDAC, India, Viewed on September 2010,< http://www.cdac.in/html/mlingual.asp>

[w9c] "TDIL-Technology Development for Indian Languages", Dept. of IT, Govt. of India, Viewed on September 2010, <http://ildc.in/Malayalam/MLindex. aspx>

[w9d] "Natural Language Processing Association", IIIT Hyderabad, India, Viewed on August 2010, <ltrc.iiit.ac.in>

[w9e] "Swathanthra Malayalam Computing Project", Viewed on April 2011, <http://smc. sarovar.org/>

[w10] "The Free Dictionary", Viewed on January 2009, <http://legal-dictionary. thefreedictionary.com/corpus>

[w11] "Linguistic Data Consortium for Indian Languages", Viewed in January 2012, <http://www.ldcil.org/ resourcesTextCorp.aspx>

[w12] "Bank of English", Viewed on January 2011, <http://www.titania.bham.ac.uk/ docs/svenguide.html>

[w13] "American National Corpus", Viewed on January 2011 http://www.american nationalcorpus.org/

[w14] "Web Crawler", Viewed on March 2009, http://en. wikipedia.org/wiki/Web_crawler

[w15] "Malayalm Unicode Font – All about Unicode Fonts", Viewed on August 2009, http://www.malayalamunicode. com/

[w16] "Writing a simple web crawler in Perl", Viewed on August 2009, http://www. stratos.me/2009/05/writing-a-simple-web-crawler-in-perl/

[w17] Forcada, M. L., Bonev,B., Ortiz S. Rojas et al. "Documentation of the Open-Source Shallow-Transfer Machine Translation platform Apertium", 2010, Viewed on Januaray 2011 http://xixona.dlsi.ua.es/~fran/apertium2 documentation.pdf

[w18] Bertinetto, Pier Marco et al., "Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)", Viewed on September 2010 < http://linguistica.sns.it/CoLFIS/ Home.htm>

[w19] Ramasree, R.J, Kusuma Kumari, P. "Combining Pos Taggers For Improved Accuracy To Create Telugu Annotated Texts For Information Retrieval", Viewed on < http://www.ulib. org/conference/2007/Ramasree.pdf>

[w20] "Vocabulary.com", Viewed on October 2010 <www. vocabulary. com>

[w21] "Prefixsuffix.com-English Language Root Reference", Viewed on February 2010 <http://www.prefixsuffix.com/affixes.php?navblks=1011000>

[w22] "Britannica Online Encyclopaedia", Viewed on March 2009, <http://www.britannica. com/EBchecked/topic/ 392829/ morphophonemics>

[w23] Loos, Eugene. "Glossary of Linguistic Terms", Viewed on March 2009, <http://www. sil.org/linguistics/Glossary Of LinguisticTerms/WhatIsAMorphophonemicRule. html>

[w24] "Dictionary.com", Viewed on 20 March 2011, <http://dictionary.reference.com/>

[w25] "Web Corp_live_concordance the web in real time", viewed on 18 February 2012, <http://www.webcorp.org. uk/live/>