CHAPTER I

INTRODUCTION
1.1. INTRODUCTION

This research proposes a structure for a computational bilingual lexicon with Malayalam as the source language. Importance has been given to the linguistic and lexicographic aspects, since these constitute the main task in developing a lexicon. The aim is to develop a bilingual lexicon of two Indian languages, Malayalam, a Dravidian language as the source language and Hindi, a North Indian language belonging to Indo-Aryan family, as the target language. The final goal is a multipurpose lexical data which can be used irrespective of any particular grammar formalism. To demonstrate this, a structure is proposed to be used as a lexicon lookup for machine aided translation between these languages.

Computational lexicon, hereafter referred as COLEX, is a lexicon used by a machine for a wide range of applications in the field of Natural Language Processing (NLP). Machine Readable Dictionary (MRD), Electronic Dictionary (ED), Computational Lexicon (CL), etc. are the terms used synonymously to refer it. It is one of the important and much needed tools in any natural language processing system. In one or the other form, it is a pre-requisite for all natural language based systems, ranging from Machine Translation (MT), text processing and generation, language teaching and language learning, etc. to development of computational systems like office automation, speech analysis and synthesis, natural language interfaces, information retrieval, and more generally a supporting tool for communications.

Natural Language Processing is one of the major research areas under the domain of Artificial Intelligence (AI). It is an interdisciplinary study where Computer Science, Cognitive Science, Psychology and Linguistics are integrated. NLP can be broadly defined as the study of computational and mathematical modelling of various aspects of natural language for its analysis and generation. The tools of work in NLP are grammar formalism, lexicon and data structure, algorithms for representing and analysing knowledge, reasoning mechanisms, etc.

The availability of a dictionary in computer readable form offers an extremely
valuable resource which can be tapped for the above mentioned purposes by performing morphological or syntactic parsing, semantic processing, recognizing written and spoken words, etc. Machine readable dictionary facilitates automatic extraction of information for a system. Here the lexically relevant information is deployed in a structure that is a succession of independent entries, so ordered, that any entry can be found through an explicitly statable procedure. A machine lexicon acts as a tool for automatic translation and helps in learning the target language. For all these purposes the lexicon should possess information at various levels—phonological, morphological, syntactic, semantic and pragmatic. Developing a lexicon to meet all these requirements is the first and important step in any NLP system. According to modern grammatical theories, lexicon is the main source of information for discourse analysis. If a fully represented lexicon is ready, then development of a natural language based system is considered more than half-way over. In the light of the above mentioned facts, in the proposed work to develop a lexicon, only a structure for the lexical data of Malayalam-Hindi is attempted, as the construction of a full lexicon is obviously the task of a group of lexicographers and computational linguists. The computational implementation and the architectural supports are done with the help of computer scientists.

1.2. LEXICON AND LEXICOGRAPHY

A dictionary (the term dictionary and lexicon are synonymously used hereafter) is a lexical data base which provides information about the vocabulary of a language. It specifies for each word of the language its part of speech, grammatical category and syntactico-semantic features and context sensitive information (pragmatic). The term dictionary is derived from the Latin word dictionarium and the term lexicon from Greek lexicon. There are a number of sources available directly related to lexicon which provide lexical information. These are glossary, concordance, thesaurus, encyclopedia, etc.

---

Lexicography, the art of dictionary making, has received attention for long time. Lexicography is generally defined as writing or compiling a dictionary. It is a part of applied linguistics. The basic concern of lexicography is *word* which is dealt with in different branches of linguistics, viz. phonetics, morphology, grammar and stylistics. India has a rich tradition in lexicography dating back to 5th century B.C. with the *Nirukta* of Yaska. The term *nighantu* has been employed to connote the meaning of lexical data or dictionary. Many dictionaries have been produced since then in most of the Indian languages of which some are available today.

1.2.1. Types of Dictionaries

Lexicons are of different types. Depending on the number of language/s they can be termed monolingual dictionary, bilingual dictionary, trilingual dictionary, multilingual dictionary. Depending on the information they incorporate, lexicons are called concise dictionary (e.g., Award Concise English Dictionary), etymological dictionary (e.g., Dravidian Etymological Dictionary), reference dictionary, synchronic dictionary, diachronic dictionary, etc. Depending on the types of users, dictionaries are called learners' dictionary (e.g., Oxford Advanced Learners' Dictionary), teachers' dictionary, foreign language learner's dictionary, etc. Depending on the size there are: pocket dictionary, enlarged dictionary, dictionary in many volumes (e.g., Oxford English Dictionary; 20 Volumes.), etc. Reverse dictionary, morphological dictionary, paradigm dictionary, etc. are some other types. Spelling dictionary (e.g., Cassil's Spelling dictionary, Dictionary of spelling: British and American, by Michel Waist, etc.), pronunciation dictionary (Daniel Johns' English Pronunciation Dictionary), synonym/antonym dictionary, rhyming dictionary, usage dictionary, etc. make a set of other types. Computer dictionary, electronic dictionary etc. are comparatively new in this field, related to modern technology.

1.2.1.1. Monolingual, Bilingual and Multilingual Dictionaries

---


A monolingual dictionary, like Oxford Advanced Learner’s Dictionary, COBUILD’s Dictionary of English Language, etc., Sabdatharavali (Malayalam-Malayalam Dictionary.) for the human user usually contains information about parts of speech, irregular inflected form, definition of meaning in the same language, and often some etymological information or details about the origin of words. A monolingual dictionary is consulted for understanding the same language. In a monolingual dictionary the meaning of a word is given in the same language either by a synonymous word or by an explanation or definition.

For example the word *kai* means *karam* (hand) or by an expression equivalent to 'the part of body used to do work', etc.

A bilingual dictionary is consulted for transforming into and understanding a second language. The contents of a bilingual one include among others the category of the item and its equivalent in the target language. There are two kinds of bilingual dictionaries. One is meant for native (first) language speakers/users which they use to get the target language equivalent. And the other is meant for the second language users. Thus, a Malayalam-Hindi dictionary for Malayalam speakers is meant for the former group while Malayalam-Hindi dictionary for Hindi speakers is for the latter. In both the dictionaries the user mostly refer for the target language equivalent. Both these dictionaries are different with respect to the information they provide. Developing bi-lingual dictionary for different users is a tedious job, particularly the selection of entries from both the languages. Bilingual dictionaries carry a list of translation equivalents in its target language. In the case of idioms, an example sentence is given from which the reader is expected to infer correct usage. Bilingual dictionaries are more in number compared to other types of dictionaries.

For a multilingual lexicon the basic information required are category and syntactic expansions of the item in the source language and the grammatical and categorical equivalents in the target languages. Multilingual dictionary has one source language and more than one target languages. Dictionaries having more than two languages are less in number compared to bilingual dictionaries because of fever users.
1.2.2. Other Lexical Sources

Glossary is a collection of vocabulary of technical or rare words, not in ordinary use, which are arranged in alphabetical order with definitions given. Glossary may be given appended with the text or given separately. Glossary discusses a word whereas a dictionary defines it. Only selected domain vocabulary is available in a glossary. Some of them to be mentioned are: Harry Show’s Dictionary of Literary Terms, Patric Murray’s Literary Criticism - A glossary of Terms, Bhadasastra Nighantu-Malayalam-English, Literary Terms English-Malayalam\(^5\), etc.

Thesaurus is a dictionary of synonyms. Here one looks for a proper word when he/she knows the meaning. It gives suitable word for the meaning. Roget's (1852) Thesaurus is the first in this field. A number of thesauruses are available in the market today.

Concordance is an alphabetical list of the words in a book with reference to the passage where they occur, or a similar list for the works of an author. Concordance of the Bible, or a Veda, concordance of words, phrases, and passages in the dramatic works of Shakespeare, etc. are examples. It is considered as a type of dictionary of quotations. Development of large machine readable corpora are on for all Indian languages sponsored by DOE, Govt. of India. Concordance can be made from these corpora by proper computer programs.\(^6\)

1.2.3. Lexicography in Malayalam

Malayalam lexicography started in 1746 with Arnose Padiri’s dictionary, 'Dictionarium Malabaricum'. Richard Collins' 'Malayalam Nighandu' (1865) is the first monolingual dictionary, and it was printed at CMS press, Kottayam, Kerala. Since

\(^5\) Kerala Bhasha Institute, Trivandrum has published a number of glossaries in Malayalam from different disciplines.

then a lot of dictionaries have been published or are being published. Among them Dr. Gundert's (1872) and S. Padmanabha Pilla’s (1932) are most accepted. The Lexicon Department of Kerala University is publishing a comprehensive Malayalam Lexicon in volumes and is yet to be completed. The latest one is "Sabdasagaram" in 3 volumes, by B.C. Balakrishnan, et al. (1991) published by D.C. Books, Kottayam. Bilingual lexicons with Malayalam as target language are more in number than the other way round. A number of Hindi-Malayalam bilingual lexicons are available. But only one or two Malayalam-Hindi dictionaries are available in the market so far.

1.3. LEXICOMPUTING AND COMPUTATIONAL LEXICON

As in any other field, computer came to the help of the lexicographer, and good quality and well documented printed dictionary production became easier. Initially, the computer was used as a tool for aiding lexicographic efforts like data collection, type setting, data management, etc. Later, the relation between the computer and the lexicon became necessary in a wide range of areas. The Lexicon in computer called electronic dictionary, developed as a tool in teaching and learning languages through computer, came as the next higher level in the relationship. The computational lexicon, an inevitable part in the natural language processing or computational linguistic field, is its current status.

After the introduction of generative enterprises which revitalized the NLP, the need for syntactic and semantic features (Katz and Foder, 1962) was identified and widely accepted. The need for a related lexicon was also assumed. But, little attention was

---

7 See Somasekharan Nair, P. 1982 and also Preface of Malayalam Lexicon Vol.1.
8 Malayalam-Hindi Dictionary
9 Sinclair, 1987
paid to actually build it as part of theoretical linguistics. A number of attempts were made since then to develop lexical database or its associated works for computer applications.\(^\text{13}\)

This section deals with the computational lexicon, its history, differences and similarities between common and computational lexicons, electronic dictionary and machine readable dictionary, etc. The questions on lexicon and grammar, and the different grammar formalism also form other major aspects of this section. Collection and classification of data, and the contents in a fully represented lexicon, etc. are briefly looked into here and discussed in detail in chapter 4.

1.4. Computational Lexicon - A Brief History

The first proposal for the design of a computerized lexicon was made by the Soviet linguist, Igor Melcuk, and his team who proposed the concept of the explanatory combinatorial dictionary. Pollock (1982) and Yonnakoudakis & Fawthrop (1983) developed a dictionary for spell correction (93,000 words). Tucker & Nirenburg (1984) designed a system as a support to MT, Briscoe (1985), a system for speech synthesis, Coltheart (1981), one for word frequency checking. These are some of the lexical data bases for specific purposes with limited information, developed in the eighties. Cowie (1983) derived a system for analyzing descriptive texts into hierarchically structured knowledge fragments. The lexical systems group at IBM Yorktown Heights has been working for Wordsmith, an automated on-line dictionary system designed to offer browsing functionality. In this system the users can retrieve words, from a number of dictionaries like 7th Collegiate Dictionary, COBUILD, LDOCE, OALD, Collins Thesaurus, etc., which are close to a given word alongwith dimensions such as spelling, meaning and sound. An on-line lexicon with 100,000

\(^{13}\) See Neuhans, Joachin, 1986; Michio Isoda, 1986; John Lafling, 1992; Bottet Ch., et al., 1986; Papegaj, B C, et al., 1986; Brijitte Barnett, et al., 1986; David Farwell, et al., 1993; Pushejosky James, 1993; Roberto Basili et al., 1994; Max Silberztein, 1989; Eric Laport, 1989; Marc Domening, et al., 1986; Abdelmajiid Ben hamadon, 1986; Makoto Nagao, 1986; Maurice Gross, 1989; etc.
words for parsing in the critique is also available.\textsuperscript{14}

Recently this area has received much attention. As most of the earlier system designers have provided only illustrative lexicons of a few hundred words for limited application, making of full fledged lexicons are drawing much attention today either by extending those earlier ones or starting afresh. COBUILD, LDOCE, OALD are the leading computational lexical data available for English developed by Sinclair (1984), Cowie (1986), and Byrd (1987) respectively.\textsuperscript{15} Corpus based dictionary is more informative, scientific and economical. This dictionary presents words according to their uses.

Computational lexicography started in India only in late eighties, and is still in its infancy. Some attempts have been made in this field.\textsuperscript{16} Some dictionaries like Kriya's Contemporary Tamil Dictionary, etc. are produced with the help of computer and are available in the MRD form also. Electronic dictionary developed by CIIL, Mysore (Tamil-Hindi-English Trilingual lexicon for second language learners) is another achievement in this area.\textsuperscript{17} It contains 96 fields of information for entries. These systems are meant for language learning and teaching. A special course on computational lexicography was held at Central University, Hyderabad, in 1994, where related issues were discussed in detail on this emerging topic, in the Indian context.\textsuperscript{18}

The COLEX developed for the Kannada-Hindi machine translation\textsuperscript{19} at IIT Kanpur is

\textsuperscript{14} See Boguraev, B et al. 1989 for more detail.

\textsuperscript{15} See Boguraev, B et al., 1989 and Sinclair, 1986.

\textsuperscript{16} See Vasudev Varma ,et al. (1993), Renjitha (1994), Eswarapilla (1993), etc. No language in India has produced corpus base dictionary. But some efforts are going on in this direction at five centers of the country to develop corpora of million words.

\textsuperscript{17} It was designed at a workshop held at Barathiyar University, Coimbatore(Oct.'93). In this workshop the researcher had the opportunity to participate in session that tried to design Malayalam-Hindi-English trilingual lexicon.

\textsuperscript{18} The researcher participated in the course, sponsored by Dept. of Electronics Govt. of India.

\textsuperscript{19} A DOE project Anusarak for translation of Indian languages to Hindi
the first of its kind in India (about 20000 words). Telugu-Hindi is also in its final stage with 60000 words. Works on Oriya, Bangla, Malayalam, etc. to Hindi also are in progress. The content of the lexicon includes the source to target equivalent and category marker along with morphological tag. They are designed and based on direct approach.\textsuperscript{20} A morphological generator supports generation of required lexical paradigm from the root/stem of a selected item and an analyzer parses lexical forms. The lexicon for \textit{Anglabharati}\textsuperscript{21} uses the concept of language independent interlingual method.

A large number of lexicons for specific purposes have been/are being developed for different purposes, such as for computer assisted language teaching and learning,\textsuperscript{22} for spell checking, grammar checking,\textsuperscript{23} for machine aided human translations and human aided machine translations, etc.\textsuperscript{24}

No lexical data is available in Malayalam so far in MRD form for any NLP use. But, there are three ongoing projects which are developing lexical data for the language. One is the corpus being developed at CIIL, Mysore for Malayalam along with the other Dravidian languages. 20 million words from different domains are put as the target.\textsuperscript{25} And the second one is the Malayalam-Hindi bilingual lexicon of MT Anusarak Project, and the third one is the English-Malayalam for Anglabharati project, both being developed at IIT Kanpur.

1.5. LEXICON AND GRAMMAR


\textsuperscript{21} A DOE project for translation of English to Indian languages.

\textsuperscript{22} Singh, G.V, (1993). Project was Computer Assisted Sanskrit Learning and Teaching (CSTLE), JNU and Pradeep Kumar, 1995 Project on Computer Assisted Hindi Language Learning and Teaching, CIIT & IIT Delhi.

\textsuperscript{23} The Word Processors for Indian languages developed by C-DA, Pane named LIP


\textsuperscript{25} Annamalai, E 1992.
Grammar defines with certain rules how lexical units can be joined to yield a meaning. After the introduction of the transformational and generative grammar which brought a sea change in linguistic theories, several formal theories of syntax were developed and applied for derivation of sentences. The first among them was phrase structure grammar (Chomsky, 1957), which tried to generate sentences from a simple lexicon of grammatical units and a formal grammatical theory with category symbols and their replacements. Transformational generative grammar (Chomsky, 1965) allowed to use the lexicon with more syntactic information and formal grammar rules. Case Grammar (Fillmore, 1963), Generalized Phrase Structure Grammar (Gazdar, 1982), Functional Grammar, etc. asked more from the lexicons than from syntax. Lexical Functional Grammar (Bresnon, 1982), Government and Binding (G.B.) theory (Chomsky, 1981), Tree Adjoining Grammar (Aravind Joshi, 1985), etc. demand most of the information from the lexicon leaving the grammar simply as the rules for regular construction. Thus lexicon became the central component for carrying a majority of the information. Most of the formalism give importance to various aspects of language and each formalism has its own treatment of the lexicon.  

The lexicon used by the machine is different from the conventional published dictionaries both in terms of how they organise and how they represent information. The information available in a conventional dictionary is limited and static. The conventional dictionary arranges words in an alphabetical order. The part of speech category like noun, verb, adjective, adverb, etc. of each lexical item is the first information a user gets. The information about derivational forms, grammatical and semantic information of the derived forms, the phrases and their extensions, citing of all ranges of meanings, with complete information on almost all entries of the language is available only in a comprehensive lexicon such as Oxford Dictionary of English Language (20 Volumes each with more than 1500 pages).

As stated earlier a dictionary in a machine readable form used for various purposes facilitates automatic extraction of information for a given requirement. The lexical data

---

26 See 3.5 for more details.
along with the relevant information is deployed in an ordered structure, which can be extracted through an explicitly statable procedure. The deployment of information is different from the conventional dictionary. For instance, it need not be in alphabetical order. The lexical information is so ordered that the search process would move from morphological to syntactic to semantic parsing.

A machine lexicon is open-ended and does not limit information collection. Taking advantages of modern technology the size of the lexicon 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 encyclopedic or pragmatic knowledge, apart from the conventional information, and also the same for the target languages as well. These informations in the lexicon can be organized in such a way that the machine could be able to understand and retrieve them for any given purpose.

The basic purpose of a bilingual lexicon is to correlate the lexical units of one language with the lexical units of another language which are equivalent in their lexical meaning. Mere lexical transfer or choosing of an appropriate target lexical item for a source language is not enough. For the language pairs such as Hindi and Malayalam which belong to different language families, the lexical transfer is not so straight. It often happens that a single lexical item of source language may correspond to a complex expression of target language, and vice versa. Further, some lexical items may appear in more than one category and chances of structural changes are also possible.

1.6. CONTENTS OF A LEXICON

Words are the basic unit of language at the syntactic and semantic levels which act as the ultimate building blocks of a sentence. Words belong to different syntactic categories which determine the distribution, i.e., the context in which they can occur. The lexicon in any system plays an important, dynamic and necessary part in the syntactic and semantic fields.
The content of various types of dictionaries determined by the purpose and use. The types of dictionaries such as phonological, morphological, syntactical, semantic, etc. depend upon the phonology, morphology, meaning, etc. of the items. Root word is used to be given as the basic entry in a common purpose dictionary because it is impossible to put all inflectional or derivational forms. Since Malayalam is an agglutinative language and is written in a continuous manner at the compound and phrasal levels without any space between the lexical items, chances of making mistakes are more while separating lexical items. If the separation is not done carefully 'ghost words' which have no existence in the language will appear.\textsuperscript{27} But unlike in Malayalam, Hindi uses space between words, even between a verb and an auxiliary.

Contents and organization of a COLEX are the main issues before us which are discussed in the following section. Usually words and category information are available in all COLEXes, apart from morphological and semantic knowledge for recognizing words.

1.7. COLLECTION AND CLASSIFICATION OF DATA

The first question which arises in the development of a lexicon is collection of the data and its classification. Data collection is the most important and tedious task in dictionary making. The usual methods applied for lexical data collection are either directly to take the data from the fields and informants or to resort to selected literary works and texts, existing dictionaries and other machine readable resources, etc.

Two general resources used for making computational lexicon data are the common printed dictionaries and other available machine readable texts. Another method being applied recently to get data is to depend on large machine readable corpora. Here, words can be tagged or sorted either automatically or manually. An aligned bi/multilingual data can be used for a bi/multilingual dictionary.

\textsuperscript{27} As discussed by B.C. Balakrishana Pilla (1972).
Data collection depends on a number of factors like size of the dictionary, contents, purpose, etc. One of the most widely accepted criteria for selection of entries for a dictionary is frequency count. Classification of entries is done primarily on alphabetical basis, and then on the basis of part of speech, sub-category, so on and so forth. Morphological classification is done on the basis of aspects like root, stem, suffixal possibility, etc. Similarly, there can be different criteria for domain specific or etymological classifications.

1.8. LEXICAL DESIGN

As said earlier, the structure of a lexicon depends upon its use. Every system of NLP has its own ideas and conventions with which it is formed. A number of grammatical theories are used for language analysis and each one has its own specific formalism. Depending on the theory the information in the lexicon varies.

The following grammatical or syntactic information is primarily needed in a COLEX: category, subcategory, inflectionally different forms, phrasal forms and possible compounds. Depending on the grammatical theories, other information like argument structure, thematic role, etc. are to be included. Pragmatic and encyclopedic information like grammatical differences, lexical ambiguity, transfer ambiguity, category ambiguity, lexical gap, homography and polysemy, anaphora, etc. are also needed.

The design of an appropriate lexicon involves the following problems to be tackled:

- What are the units of entries and the information to be incorporated?
- How should lexical information be prepared?
- What is the domain? Is it specific or general or extensible?
- What is the size required for the lexicon for a given application, and what is the technique used to limit the size?
- What is the organization mode of bilingual lexical items?

---

Section 3.6.3
* What should be the kind of techniques applied for generating a multilingual lexicon? Can the lexicon be sufficiently prepared from an existing one, and what are the additions needed?

* What is the technique of organization of lexicon to make the access faster?

The size of the lexicon depends upon the application. A common purpose lexicon provides frequently used lexical items along with the related information. That is, we can include or delete entries according to our requirements. The basic morphemic root or the basic stem is the minimal unit. The lexical information is provided according to its usage, apart from part of speech and subcategory. The semantic domain can be specified or made general with the help of pragmatic information. Lexicon for specific purposes can be generated or developed from a generalized data structure by some conversion and modification as per the requirements.

As the lexicon has to be an open-ended one, it needs to have a dictionary editor to facilitate adding, updating, deleting and checking of entries at the terminal. Under the editor, specific sets of menus and help devices have to be provided. An interface facility with the help of which the system can be interfaced with the user is necessary. As the aim is to develop a general purpose multilingual lexicon, the system should be made applicable to any grammatical formalism. The lexicon should be extensible, as it would be an important desirable property, and the lexical system can be extended or changed phase by phase, according to the need of the user.

More often than not, a word in a given sense in a language has its counterpart in the other language as well. Hence it is prudent to store in the lexicon against a word in a given sense all its equivalents in the target languages. Each word in the source language has to be disambiguated with regard to all its senses. In other words, a lexicon is a source to store information for disambiguating word sense. In addition to the words in target languages, details such as their morphological information, special grammatical properties, etc. has to be stored. Phrase should also be stored and word clusters should be treated with regular parts of speech. This enables the parser
to treat a single word and a phrase alike, while filling slots for a particular part of speech.

1.9. LEXICAL KNOWLEDGE AND ITS REPRESENTATION IN A COLEX

Knowledge Representation (KR) is the task of providing a formal representation for information and it is increasingly becoming a part of diverse computational systems. It supports reasoning and its significance is very evident in the field of semantics where representing meaning of utterance is fundamental. Moreover, the semantics of natural language constructions provide a long term potential for studies in knowledge representation.

We are more concerned on syntactic and semantic aspects, since phonological\textsuperscript{29} and morphological information\textsuperscript{30} do not take part in the two important fields of syntactic and/or semantic parsing for the generation and understanding of a natural language discourse for which COLEX is primarily used.

The primary function of the COLEX is to assist the parser in translating the input sentence into an internal meaning representation to be produced. The process of applying a grammar to an input sentence to determine its grammaticality and appropriate syntactic structure is known as parsing, and to parse we need a lexicon of all the components of the sentence. Semantics deals with the meaning of words. Modern syntactic theories like Standard Theory (Chomsky, 1965), G.B. theory (Chomsky, 1982) LFG of Bresnan (1982), Lexical semantics of Jakendoff (1986), Fillmore's Case theory (1963), etc. talk in detail about the semantic aspects and the requirements in a lexicon. Most of the theories say that the lexicon should comprise as much of the semantic information as possible. For example, G.B. focuses on the thematic role which is actually a semantic relationship between heads and their syntactic complements. G.B. argues that these relations are to be stored in the lexical

\textsuperscript{29} See section 2.2

\textsuperscript{30} See section 2.4.
entries of potential heads (verb, adjective, noun, etc.).

1.10. APPLICATIONS OF COLEX

A dictionary serves as a device to correlate the components of the structural system, such as grammar which reflects the usage of individual items. An NLP system is able to operate successfully only if it has different information such as morphological knowledge - the internal structure of words, syntactic knowledge - the organization of words into phrases, semantic knowledge - the meaning of words and the combination of meanings to form the meaning of sentences, and pragmatic knowledge - extra-linguistic information. All this information is supposed to be available in a COLEX. So a fully represented lexicon is a must in all systems for the lexical analysis, which follows a purely computer based calculation and consultation.

Linguistic analysis at any level, whether syntactic or semantic, cannot be started without a complete list of words of the text. In the understanding mode, the lexicon is used to recognize words and embedded clauses, to construct kernel structure and to build transformation expressions. In a generating mode, the lexicon is consulted when a noun or verb phrase is built, when a connective transformation is applied and when a question is answered. The lexicon extends the system’s abilities for semantic interpretation too. Virtually, every branch of a system resorts to the lexicon to accomplish its tasks.

1.10.1. Lexicon For Machine Consumption

The major uses of COLEX for machine are in MT and LT/LL. Machine used translation systems are of three types such as Machine Aided Human Translation (MAHT), Human Aided Machine Translation (HAMT) and fully automated Machine Translation (MT). An MT process involves understanding and analyses of words at different linguistic levels. Important processes in MT system are syntactic and semantic parsing. Three strategies are used in MT with respect to lexicon (Hutchins, J. et al., 1992:70-77). They are direct method, transfer method, and the interlingua
method (See the fig.1). The direct approach uses a morphological analysis phase, where there would be some identification of word endings and reversion of inflected form to their uninflected basic forms, and the result fed into a large bilingual dictionary look-up program. Lexical identification is the only main process and that depends on morphological analysis and leads directly to bilingual dictionary look-up providing word equivalents in the target language. The limitation of this approach is that it allows for only word-for-word translation with the same local word-order adjustment.

1. Lexicon in Direct Method

Source Lexicon  <------>  Target Lexicon
Grtl. & Sem. inf  <------>  (grtl & Sem. inf)

2. Lexicon in Interlingua method

Source lag. 1-------- ------ Target lag. 1

Interlingua

Source lag. 2-------- ------ Target lag. 2

3. Lexicon in Transfer Method

Source lge. <--> S/T transfer  <--> Target

Fig 1.
Interlingua method bases on a lexical representation which is language independent languages. The advantage of this method is that with a single intermediate structure a number of languages can be used as source and target languages. In the transfer method source and target equivalents get transferred through a transfer module. Here also, the method is language-dependent.

a. Information Needed For Machine Translation

Phonetical and phonological information is mostly required for the systems which use speech and written media. For speech systems, analysis of the input sound, and generation of synthetic speech in the target language are the two tasks.

Morphology is concerned with the structure of words and the ways in which words are formed from basic sequences of phonemes. Two types are distinguished: inflectional and derivational morphology. Words in languages differ in form according to different functions, e.g. nouns in singular and plural (\textit{kutT}i 'child' and \textit{kutTikaL} 'children'). Inflectional morphology is the system that defines the possible variations on a root form and combines it with various endings to form grammatically functional units. Often the root itself undergoes some changes. (\textit{kaLikku}- 'to play' to \textit{kaLiccu} 'played' and \textit{kaLikkunnu} 'is playing'; \textit{kaaNu}- 'to see' to \textit{kaaLca} 'sight', etc.) Derivational morphology is concerned with the formation of root/stem from other roots, often a change of part of speech is involved. For example the word \textit{kaLi} 'playing', or \textit{caaTTam} 'jumping', \textit{paRna} 'to study' to \textit{PaRaayi}, etc. are derived from their roots. Compounding, where whole words are combined into new forms, is the third process in morphology. The meanings of compounds are sometimes obvious from their components (\textit{pantukaLi} 'play with a ball'), sometimes slightly different (\textit{kutTTikaLi} 'play like a child'), and sometimes completely opaque (\textit{kathakaLi} 'a type of dance form in Kerala').
Morphological analyzer\textsuperscript{31} is commonly used to generate or parse a construction taking the morphological information available in the lexicon (Sproat, R. 1992;2-7). With the help of a morphological analyzer along with the grammatical rules the morphemes and their arrangement can be identified. From a single entry it can produce as many number of words as required. This reduces the size of the lexicon considerably because the size of the lexicon may affect the efficiency.

Syntactic and semantic information is required for the process of connecting a text with its meaning. The primary role of syntactic analysis is to decode knowledge that is semantically encoded. An important component of such knowledge is the relationship between the predicates and their arguments. Much of the information about the syntactic encoding of arguments can be placed in a lexicon as arguments and predicates. Classes of predicates with similar encoding usually have some similarity and thus define a kind of lexical semantics based on semantic features that are predicates of syntactic behavior. So the theory of lexical semantics includes a component of lexical knowledge that describes the syntactic encoding of arguments. Valency information, case frames, semantic features and selectional restriction, arguments and thematic information, etc. constitute the semantic information. Among these some are very important and take space in the dictionary.\textsuperscript{32}

**COLEX For Sentence Parser**

Parsing is a process by which an input sentence is assigned a suitable structure. This structure could either directly represent the meaning of the input sentence or it can be used by a semantic interpretation. Parsing process makes use of two components: a parser- procedural component, and a grammar -declarative component. Parser structure consists of either morphological or grammatical relations, or both and a dictionary of items. For example the sentences below have a structure (NP1 + NP2 + V), and the lexical units are (kuTTi 'ledka', skuul 'skuul', poku 'jaa').

\textsuperscript{31} See 5.5 for more detail
\textsuperscript{32} See section 3.6
e.g., kuTTi skuuLil pookunnu
ladka skuul jaata he.
NP (karta) ladka 'boy'
NP (karma) skuul 'school'
V (kriya) jaa 'to go'

A lexicon for syntactic parsing requires only the syntactic information of the respective lexical entries. It is basically category and subcategory information. A lexicon for semantic parsing requires syntactic and semantic information and also pragmatic information.

1.10.2. Lexicon For Human Use

The lexicon for a language learning or a teaching system requires almost all linguistically relevant information at the lexical level. Apart from phono-morpho-syntactico-semantic information, correct usages from classical or standard texts are also necessary.33

1.10.3. Contents of a Common Purpose COLEX

The components of knowledge needed to be incorporated in a lexical design are:

1. Phonetic and phonological properties of each entry to recognize the spoken form and generate words through morphological process

2. Morphological information to derive inflectional and derivational variants from the root

---

33 See section 4.3.2
e.g. 1. avan 'he'- + (numb-pl/gend-mas + case-gen) > avaruTe
2. vaa 'come'- + (caus/pass + TAM) > varuttunnu

3. The word in the base form with its grammatical properties which include part of speech and related attributes

   e.g. 1. avan, proN, mas, Sg, "he"
   2. vaLyu V(tr), 2-arg, agent/object, "to bent"
   3. veegam adv, move, "fast"
   4. nalla adj, qual, "good"

4. Semantic tags to identify various senses and the related words
   e.g., relational tags like synonyms / antonyms / and domain specifier.

5. Sense disambiguation rules for eliminating ambiguity

   1. Word level - Lexical ambiguity
      e.g. kola 1.<vn>."killing", 2.<n> "bundle"
   2. Phrasal
      e.g. puka vali paaTilla (have two meanings)
         1.<complex V> "don't smoke".
         2.<Compd V> "it is easy to smoke"

6. Root in the target language with related information
   e.g. madrasa (Mal.) > madrasa (Hin) "school"

1.11. LIMITATIONS AND DRAWBACKS OF THE AVAILABLE LEXICAL DATA

Computational lexicography is presently receiving much attention and construction of COLEX of natural languages for NLP application is in progress and at different places in India and abroad. Most of the earlier systems designers have provided only
illustrative lexicons of a few hundred words, for limited application. A large number of lexicons for specific purposes are developed/being developed by different groups, for different applications.

COBUILD (Sinclair, 1984), LDOCE (Cowie, 1986), OALD (Byrd, 1987), etc. are the leading computational lexical data available in English for various purposes.\(^{34}\) Computational lexicons in Indian languages are available for only limited applications. Some lexical data like *Kriya's Contemporary Tamil Dictionary*, etc. are available in MRD form too. Electronic dictionary of CIIL, Mysore (Tamil-Hindi-English lexicon for second language learners) needs significant modifications for proper use. All these lexicons are used as a tool for computer assisted language teaching and learning, for spell checking, grammar checking, or machine aided human translations and human aided machine translations, etc.

The COLEX systems like Kannada-Hindi (about 20000 words),\(^{35}\) Telugu-Hindi (60000 words), Oriya-Hindi, Bangla-Hindi, Malayalam-Hindi, etc. being developed does also need modifications to serve the desired purpose. The content of these lexicons include the source to target equivalent and category marker along with limited morphological tagging. All these lexicons have a morphological generator, which support generation of required lexical form from the root/stem of a selected item and an analyzer that parses lexical forms. The lexicon of *anglabharati* project is a multilingual one with English as source language and different Indian languages as the targets.

All these lexical systems are having limited number of lexical items with limited knowledge and limited purposes. Converting these lexical data for other purposes is more tedious than making afresh. That is why making of a full fledged lexicons for multi-purposes is the need of the hour.

---

\(^{34}\) See Boguraev, B et al., 1989 and Sinclair, 1986

\(^{35}\) see Narayanan V, 1993
1.12. THE STRUCTURE PROPOSED FOR THE PRESENT LEXICAL SYSTEM

It is identified that at least five broad types of knowledge are potentially relevant to NLP system. They are:

1. Phonological knowledge concerning the sound system and structure of words and utterances of the system for the text to speech synthesis.
2. Morphological knowledge about the internal structure of words.
3. Syntactic knowledge concerning the arrangement of words into phrases and sentences.
4. Semantic knowledge concerning the meanings of words and combination of these meanings to form the meaning of sentences.
5. Pragmatic or encyclopedic knowledge to specify the domain and delimit the semantic extension.

The following are the general attributes which each lexical unit proposes to carry.

a. Word in a general or domain specific sense.
b. Frequency of use.
c. Category specifications.
d. Defining semantic features.
e. Levels of semantic categories.
f. Semantic relation with other word groups.

Apart from these there are category specific attributes.

I. Noun: As the number of nouns is endless compared to other categories, it should be treated carefully considering the following issues.

1. Different usages, meaning, origin, etc.
2. Criteria for categorizing compound nouns
3. Regular and irregular forms
II. Verb: Though verbs are limited in number, they play unlimited roles. Verb is the central unit of a sentence and has got significant role control over other categories. It needs particular attention in the following aspects.

1. Grammatical attributes like transitive-intransitive, active-passive, finite-infinite, tense/mood/aspect, etc.
2. Case identification with respect to subject and object, like theta-marker, etc.
3. Regular or irregular morphology identification.

III. Modifier: Modifier includes adjectives and adverbs and their semantic information such as quantitative/qualitative, temporal/spatial, etc.

IV. Avyayas (Items other than noun, verb and modifier) include post-positions, indeclinables, etc. Most of the forms undergo no structural and category change.

So keeping all these things in mind we propose a structure for a bilingual lexicon of Malayalam-Hindi with the following information.

**Head word in Source language (Malayalam)**

- **field1** category: (Noun, verb, modifier (adjective/adverb), postposition
  - **sub-cat.**: (Abstract/ concrete, transitive/ intransitive, quality/numeral, temporal/ spatial, etc.

- **field2** Semantic Attributes: (Argument structure(1/2/3 No., subject /object, animate/ inanimate, etc.)

- **field3** Inflectional forms: For verb (Infinitive/ tense forms - past/present/future, aspect/modal forms). for nouns (plural/case forms/participial forms etc.)

- **field4** Derivational forms: (verbal noun, adjective, verb, etc.)

- **field5** Compounds / Phrases

- **field6** Synonymous word(s)
field7  **Etymology or origin**

**Head word in Target** language (Hindi)

field8  **Inflectional** forms: (infinitive/ tense forms (with GNP), plural/ case forms/ participial forms etc.)

field9  **Derivational** forms: (verbal noun, verb, adjective, etc.)

field10  **Compounds / Phrases**

field11  **Synonymous** word

Example

Mal. Head word  L1  >  *paTi-kku*  'study'

Field1  verb (Tran., Animative)

Field2  Arg 2 (Sub/obj)

Field3  *paTi-kku-ka/paTi-ccu/paTi-kku-nnu/paTi-kk-um/

paTi-ppi-kku-nnu

Field4  *paTi-ttam/ PaT-anam/ paTippu, paTi-cc-a-van/ vaL,

Field5  *paTippumuRi

Field6  --

Hindi L2  >  *paTna*

Field7  *apaT-na/ paT-aa/ paT-ttac/ paT-um +GNP, paT-aan

Field8  *paT-yi, paTnevaala,

Field9  *paTnevaala_kamara

Field10  --

Source and target equivalents are given one after the other, and almost in all cases they have same category and argument structure/s. All the inflectional forms are given in a regular pattern, i.e., infinite comes first, and then the tense forms (past, present, future, causative, etc.). Option is provided for storing either by on-line input from keyboard or by a morphological analyzer. Since Malayalam language doesn’t show any
GNP ending, which is important in Hindi, gender specification should be provided for
the latter. The derivational forms and compound or phrasal forms etc. are the fields
that follow.

In this system, all the major lexical categories will have an assigned value. The phrase
or compound forms will also have the same value. As the system provides a
morphological analyzer, base or root morph could be used as the basic unit along with
pattern specification. The bound morpheme, say, suffix or prefix, will also have
specific identification. An entry is to have at least five or six fields such as base form,
category, argument structure, semantic and pragmatic tags, inflectional and derivational
forms, phrasal and compound forms and synonymous words.

The following are the main requirements and tasks in constructing a COLEX which
is discussed in detail in the fifth chapter.

1. Organizing entries from both the languages for a bilingual lexical data for a
given purpose. We follow transfer cum interlingua method.
2. The method of arranging and transforming the lexical formats in an on-line file
in such a way that the various types of lexical information become easily
accessible to the user.
3. Providing the bilingual lexical data base with lexical entries containing
sufficient grammatical and semantic information for various other purposes.
4. Producing the required lexical item from the basic lexicon through the
morphological generator, and parsing a construction by the morphological
analyzer.
5. Providing a dictionary editor (a menu driven) to facilitate coding, updating,
deleting and an interfacing facility.

1.13. ORGANIZATION OF THE THESIS

The thesis contains five chapters besides the conclusion. The first chapter is
Introduction which looks into the details of computational lexicography and the related
fields. It gives a glance at the various other types of lexical data and uses of COLEX. Different types of knowledge required in the lexicon, and their application, and the limitation of the available lexical data, etc. are briefed. The second chapter is about the phonological and morphological aspects in a lexicon. The concept of word and various word formation rules and techniques etc. are discussed. Inflectional and derivational morphology and compounding are also discussed, particularly, of Malayalam. The third chapter is on the syntactic and the related grammatical issues taking the source language as example and on the semantic aspects for a lexicon. Grammar and lexicon, lexical semantics, argument structure and thematic roles/theta-grid, case/karaka relations, semantics of compounds and phrases, and lastly, pragmatic and other deep level lexical information are the topics discussed. Chapter four is on the basic requirements of a multilingual COLEX, particularly of Malayalam-Hindi and information needed for different purposes, viz., syntactic or semantic parsing, machine translation, CALT/CALL, etc. Fifth chapter looks into the details of design and implementation of COLEX proposed and discusses the different modules used, viz. file handling search procedure, retrieval, morphological analysis/morphological generation, etc. We follow the "B-tree" search method and stores four category data. A dictionary editor (menu-driven) which facilitate coding, updating, deleting and checking of entries of the lexicons etc. are also discussed. Description about various techniques that can be used for the modification or conversion for different purposes is given, and the implementation issues are also discussed. Conclusion and Future Prospects conclude the discussion by assessing the plus and minus of the system developed. It gives hints on the prospects of the model too.