CHAPTER III

SYNTAX AND SEMANTICS OF WORDS IN LEXICON
3.1. SYNTAX AND LEXICON

Grammar is a set of rules which describes how lexical units can be joined to give a meaningful sentence. The grammar tells if a sentence is formed according to the rules of a given language. Different grammatical formalisms are available for the analysis of a natural language with different perspectives. Grammatical theories are now becoming mathematically precise in their description. Introduction of the generative grammar, applied with mathematical representation to language, led to the computerization of natural languages. It was based on context-free grammar and was the first serious attempt in that direction. It was a break-through in linguistic theories and grammar formalisms. Since then, several formal theories of grammar were developed for analyzing natural language texts. Recent researches in linguistics and NLP led to the conclusion that lexicon plays a very important role in grammar. Properties associated with lexical items are important and play the central role in the grammar. This idea led to the lexicalization of grammar which is fundamental for the computational linguistics. Lexicalization brought a number of terms to the linguistic field like lexicalized grammar, lexical semantics, lexical conceptual structure, lexical functional grammar, etc. for describing different aspects.

The first grammar developed with the intention of computerizing natural language was Phrase Structure Grammar (PSG) which tried to generate sentences from a simple lexicon consisting of categories like nouns, verbs, etc. and a formal grammar rule. Transformational Grammar (TG) and generative grammars made use of a simple lexicon with more information like binary features and few formal grammar rules. Case Grammar, Generalized Phrase Structure Grammar (GPSG), Functional Grammar (FG), etc. seek more from the lexicon including semantic information. Lexical Functional Grammar (LFG), Tree Adjoining Grammar (TAG), etc. require most of the information from the lexicons.

---

3.2 GRAMMATICAL FORMALISMS AND THE PLACE OF LEXICON

The basic difference between the lexicon and the grammar lies in respect of their being open-ended and close-ended respectively. The grammatical rules of a language are internalized by an individual by the age of five or six years. Practically little is added to the grammatical structure afterwards. On the contrary, the acquisition of vocabulary is an ongoing and continuous process which lasts till the end of one’s life. Similarly, new lexical items are added to the lexicon, and it is constantly changing. In other words, grammar rules are static for a given time, but the lexicon is dynamic and get updated continuously. This is true with machine lexicon also as a system starts with limited lexical entries and addition of new entries goes on according to the need.

We shall briefly describe grammar formalisms, particularly three important ones, widely used in computational linguistics, viz. Lexical Functional Grammar (LFG) of Bresnson (1986), Government and Binding theory (GB) of Chomsky (1982&1991) and Paninian Grammar (PG) as presented by Akshar Bharati, et al.,(1994) with the following perspectives:

1. Does there exist a lexicon in a grammar formalism?
2. If yes, what is its structure, position in the organization of the grammar, and what is the component (morphological, syntactic or semantic) of which it is a part?
3. Purpose and organization (phonological and/ or morphological)?
4. Phonological, syntactic & semantic information required for these formalisms and how it works?

Before we go to the details let us have a general view of some important grammar formalisms, both Western and Indian. All grammar formalisms have the concept of lexicon, but lexical representation is different for each of them. Lexicon according to different grammatical theories is discussed in this section. And then we will see the structure of lexicon of the grammars of our choice.
3.2.1. Western Theories

a. Traditional (Descriptive) Grammars

Bloomfield, father of modern linguistics, considered grammar and lexicon as two parts of linguistic description and opined that 'lexicon is really an appendix of the grammar, a list of basic irregularities'.

His statements seem to be inspired by the fact that grammar takes care of all the regular and predictable forms of language whereas dictionary gives all the irregular and unpredictable forms, and also forms with irregular and unpredictable meaning. The dictionary gives irregular plurals, irregular forms of verbs and other unpredictable forms in the paradigm of each lexical unit. It does not enter regular inflected forms but gives derivative forms. It gives all the lexical units of a language because the relation between the form and the meaning is not predictable. This approach has been followed for a long time, till the introduction for generative grammars. Then the approach became reverse, as the lexicon became source of regularities and grammar a set of rules for irregularities.

b. Context Free Grammar (CFG): This is the first generation grammar rules used for the analysis of natural languages. As the title indicates it gives no importance to the context and any category can occur at any position without considering the context. A context free grammar consists of rules and the operation is that of rewriting or substitution.

3.2.1.1. Chomskyan Syntax

a. Phrase Structure Grammar (PSG) was the first attempt of presenting grammar rules with mathematical expressions. PSG of Noam Chomsky (1957) is an extension of IC analysis of Zelling Harris. PSG tried to generate sentences from a simple lexicon of grammatical categories like noun, verb, etc. with the help of a simple formal substitution rules involving Np followed by Vp. Vp consisting of verb, noun, etc., and

---

2 Bloomfield, 1933:74.
the like. The basic idea of PS is that a sentence of a language can be divided into constituent parts such that each word in a sentence is a constituent of some phrase that, in turn is constituted of lexical categories (e.g. N, V, Det., etc.) called terminal strings. The structure looks like:

\[ S \rightarrow NP + VP, \]
\[ \text{where } VP \rightarrow V + NP \text{ and } NP \rightarrow D + N, \text{ etc.} \]

b. **Standard Theory** is the grammar which operates using transformational rules, and thus termed transformational grammar. Chomsky introduced this as an illustration of a generative device more powerful than PS grammar. In this view very many sentence types can be economically derived by supplementing the constituent analysis rules of PS grammar with rules for transforming one sentence to another. The rule of passivisation or question formation, etc. can be generated from the active sentences. The arguments are persuasive, and as a result TGs became the most influential in the development of generative grammar theory; indeed the field as whole has since been variously known as generative grammar. The constituent relationships are put in tree diagram or expressed as rewrite-rules as:

1. \[ NP \rightarrow \text{innale pooya raaman ...} \]
   \[ 'Ram who came yesterday ...' \]

2. \[ NP \rightarrow \text{RPR + N} \]
   \[ \rightarrow \text{Adv + RP + N} \]

3. \[ \begin{array}{c}
\text{NP} \\
/ \backslash \\
\text{RPP} \rightarrow \text{N} \\
/ \backslash \\
\text{Temp RP} \rightarrow \text{N}
\end{array} \]

4. \[ \operatorname{np} \left[ \operatorname{Rpp} \left[ \text{Temp RP}_{\text{Rpp}} \right] \right] \operatorname{N}_{\text{np}} \]
The standard theory of TG (Chomsky, 1965) consists of three components. 1) a syntactic component, comprising a basic set of phrase structure rules, which together with lexical information provides the deep-structure information about sentences, and a set of transformation rules for generating surface structure. 2) a phonological component, which converts strings of syntactic elements into pronounceable utterance. 3) a semantic component, which provides a representation of meaning of the lexical items to be used in the sentence. Deep structure (DS) and Surface structure (SS) first appeared in standard theory to incorporate semantic aspect into the grammar. Later these terms were changed into L-structure and F-structure.3

c. X-bar Syntax and GB Theory

A detailed discussion of GB theory with respect to lexicon is done in section 3.3, as it is one of the formalism given importance.

d. Universal Grammar and Minimalist Approach

As an extension to GB grammar Chomsky proposed the principles and parameters (P&P) hypothesis and argued for a universal grammar. Universal grammar is a system of rules that are common to all human languages. The parameters are those by which languages vary, and individual languages illustrate different setting or values of the parameter. For example, at the word level or morpheme level the parameters can be prefixing, suffixing, infixing, etc. or head first, head last, etc.

The latest version of Chomskyan syntax (Chomsky 1994) is called minimalist theory, where grammar rules are minimum and the lexicon is the most important source of information. The lexicon bears maximum information leaving very less to the grammar component. It is said to be a highly computational grammar.

3 See Haegman L, 1991 & Cook. V. J. 1990;125
3.2.1.2. Generalized Phrase Structure Grammar (GPSG)

GPSG developed by G. Gazdar is an extension of context free grammar. It has no transformation and the syntactic structure of a sentence is a single phrase marker, and category has a set of feature specifications which the rule can access. It incorporates meta-rules, which determine the membership of a large set of fully specified phrase structure rules, and an algorithm for admitting structural description of sentence. In addition, the approach requires the statement of feature-co-occurrence restriction and feature-specific defaults. A lexical entry in GPSG can be represented as:

```
varu 'to come' [ cat [head verb (finite) ] ]
[ sub cat < NP [noun] ]
  content [ relation varu ]
  [ agent [] ]
```

GPSG extensively uses feature structures which are equivalent to arguments of X-bar theory.

3.2.1.3. Lexical Based Grammars

Lexicalization of grammar assigns a central role to the lexicon. Two desirable properties of grammar are lexicalization and locality. The definition of lexicalized grammar incorporates the following features.

1) Every finite grammar is associated with one or more lexical items, and the associated lexical item(s) form the anchor of the corresponding structure which must be realized in a sentence.

2) There are one or more operations for composing the structures and these operations on structures are used in deriving sentences.

---

4 Bharati Akshar et al. 1995:159
As recent works in linguistics and NLP have demonstrated, lexicon plays a very important role in grammar. Properties associated with lexical items are important and play a central role in the lexicalized grammar. Important lexicalized grammars are Lexical Functional Grammar (LFG) and Tree Adjoining Grammar (TAG).

**a. Lexical Functional Grammar**

Discussion of LFG is taken-up in section 3.3.

**b. Tree Adjoining Grammar**

TAG is an extension of LFG. It was introduced by Aravind. K. Joshy, Levy and Takahashi. It consists of initial trees and auxiliary trees and two operations, substitution and adjoining. It is a finite set of elementary trees and uses an operation called adjoining to compose bigger trees. Elementary trees correspond to universal linguistic structures and localize the dependencies such as subcategorization and filler-gap. An initial tree corresponds to the structure for a simple sentence, one in which no recursive derivation occurs. An auxiliary tree corresponds to recursion on non-terminals. Therefore, the first node of an auxiliary tree whose label is the same as its root node. The adjoining operation generates extended structure from the minimal linguistic structure. TAG assures that the dependencies between lexical items are encapsulated with the elementary tree.

**3.2.1.4. Other Theories**

A number of other grammatical theories are available which cannot be ignored, though some of them were rejected, such as case grammar of Fillmore(1968), etc. as they failed to tackle the different aspects of human languages. But some of the arguments they put forward are still valid. Here, we are attempting to discuss some of them.

---

5 See LFG in section 3.3.2

6 Aravind K Joshy, et al., 1990
a. Case Grammar

Fillmore (1968) introduced the concept of case grammar following the Chomsky's phrase structure grammar, where the latter failed to include semantics. Fillmore (1968;24-5) has defined six cases viz. agent, instrumental, dative, factitive, locative, objective.\(^7\) The case concept of Panini with modifications was the base of this.

b. Categorical Grammar (CG)

As the title says the category (grammatical) is the basis of GG. A large proportion of current work on CG is either directly computational or heavily influenced by computational correlation.\(^8\) The syntactic behavior of any item is directly encoded in its lexical category specification: categories, atomic or complex, replace phrase-structure rules and this makes a separate grammar rule component. It contains two or three fundamental categories of sentence, S and Nominal, N or S, NP and VP, and two functor symbols, "/" and "\" indicating what is expected to the right and the left respectively. A characteristic feature of CG is that the lexical entries of words encode virtually all the information about how words are combined into phrases; there is no separate component of syntactic rules as is found in most other grammatical frameworks. It is the categories of CG which carry the burdens of syntactic and semantic description. The functor categories or complex categories are most simply written with a vertical slash notation, with the value to the left and argument to the right of the slash connective.

i.e., the intransitive verb is S/N and adj. N/N and transitive verb (S/N)/N.

This is the way of describing the syntactic and semantic behavior assigned to a category in the lexicon. Any word with a number of different behavior patterns will correspondingly have a number of different categories, so that a verb like 'eat' which

\(^7\) See section 3.6 for more on case relations.

\(^8\) Mary McGee Wood, 1993.
can be either transitive or intransitive ('Mary eats mango', 'monkey eats') will have the two categories (S/N)/N, S/N. A categorical lexicon is larger than those of other linguistic theories, but that is rather a symptom of the centrality of the lexicon than of the greater overall complexity of the grammar as a whole. Given a lexicon full of active and complex category assignments, it is possible to identify the categories of the individual words in a sentence; but, to check the legality of that sentence and derive a semantic interpretation, there must be rules which can be applied to these categories.

c. Functional grammar (FG)

FG is a general theory of grammatical organization of NL, based on the functional view of the nature of language. It is a sentence grammar envisaged as part of a wider theory of verbal interaction and ultimately as a sub-component for a model of natural language using system. The major grammatical rules of FG fall into four types: predication formation, terms formation, argument and expression rules. The predication and terms formation rules apply in the 'fund' to extend the sets of predicate frames and terms that may serve as input to the formation of sentences. The fund consists of a lexicon containing the properties of lexical items that must necessarily be learnt and memorized and two rules components generating forms derivable by synchronically productive rules. The FG stance on the lexical representation is that the irregular and unproductive forms are stored in the lexicon while regular and productive ones are derived by word formation rules. Since regular and productive forms are excluded from the FG lexicon, the lexical entries are ridden off all inflected and derivation morphemes that can be predicted by productive rules.

d. Valency Grammar (VG)

VG discusses words according to the number of valencies. The number and nature of the complements which a given verb takes are given by its valency. For example, Verb has a valency value between 1 to 3.
e. Cognitive grammar

Cognitive grammar is based on the human cognitive aspects and it is modelling human mind. Cognitive grammar is concerned with the mental representation of the world and its relation to language. It assumes that syntax is totally determined by the structure of meaning. Cognitive rules and the lexicon based on mental lexicon are the main components here.

3.2.2. Indian Grammatical Theories

There are two basic grammatical formalisms usually followed in the Indian tradition. The first one is of Panini in Sanskrit. All later grammars of almost all Indian languages generally followed the Paninian system, even upto the extent of adopting the same terminologies. The second one is of Tholkapiyam in Tamil, which was restricted to the language itself and to a certain extent this also follows certain Paninian concepts.

a. Tholkapiyam: The known Tamil grammar begins with Tholkapiyam, said to be of 2nd Cent. B.C.9 It was a treatise on grammar, rhetoric and literary criticism. It demonstrates that Tamil had evolved obsolete forms and semantic shifts, necessitating explanation to contemporaries, and contains a glossary of Sanskrit literary terms with Tamil translations.10 Except Tamil no language follows this tradition.

b. Paninian grammar (PG) is associated with the verb-noun relation through specifying kaaraka relations.11 The kaaraka theory is discussed in detail in section 3.3

---


10 Gregory James, 1989

11 Kapoor, Kapil, 1985
3.2.3. Grammatical Theories in Malayalam

Malayalam grammatical tradition\(^{12}\) starts only from 18th century, though there was an earlier work of 14th century treatise called 'Leelathilakam', written in Sanskrit by an unknown author. It was not a real grammar, but is meant as an aid for writers and connoisseurs, and it discussed only certain aspects of the grammar. After the European invasion to the land, a number of grammars and dictionaries were produced. The most important among them are Gundart’s works (a grammar in 1868 and a dictionary in 1872), Garthwite’s, George Mathen’s, etc. At the end of 19th century a number of natives also came in the field. 'Keralapanini' by A.R. Rajaraja Varma (1895) is the most important one. Even today it is considered to be the best grammar in Malayalam. This was written following the rules and style of the Astadyaayi of Panini. Latest scientific grammar is K. Sukumara Pilla’s (1980) 'kairalisahdaanusaasanam'. A number of studies based on modern grammatical theories on morphology and syntax had been attempted in the second half of this century. A. P. Andrewskutty (1978, 1984), K.P Mohanan (1984, 1986), Jayaseelan (1991), V.R.Prabhodhachandran Nayar (1978, 1982) and others authored some of the noticeable works.

Grammatical characteristics of the language discussed in the foresaid grammar works are mainly on morphological and sandhi rules\(^{13}\) and also on syntax of noun: co-ordination, attributive combinations, definite numeral attributes, etc. Cases, kaaraka relations, verb and particle to which each case is related, etc. are other issues that took some attention. Sandhi rules, case and kaaraka relations are mostly followed on the lines of Panini. And other issues are regarding syntax of verb, relative and verbal participles, conditions, concessives, auxiliaries, particles, etc. As 'keralapanini' is the most accepted grammar today, most of the concepts/ideas we follow are from it.

\(^{12}\) See Ezhuttachan, K. N. 1982.

\(^{13}\) See the section 2.3
3.3. STRUCTURE OF LEXICON IN GB, LFG AND PG

As pointed out, we are concentrating our discussion of grammar formalisms and lexical structures only on three prominent grammars. They are the basic and much in use for NL processing. Among them LFG parsers are being used more and are available in the market for research and commercial purposes. The lexical representation used by our system (discussed in chapter 5) can be converted to the structure for any of these required formalism with minor modifications.

a. GB Theory

GB is one of the dominant theories for NL processing, but is less used in computational application. This is because it has less to say on parsing and generation, which are very important for computation. GB has an interacting system of principle which places constraints to filter out ungrammatical sentences. Following are the components.\textsuperscript{14}

\begin{itemize}
  \item \textbf{X-bar theory} - gives structure of phrases
  \item \textbf{Thematic theory} - assigns syntactic positions and does not directly deal with arguments
  \item \textbf{Government} - defines the domain of influence for a head
  \item \textbf{Case theory} - case is assigned by category head
  \item \textbf{Bounding theory} - subjacency
  \item \textbf{Binding theory} - governing category
  \item \textbf{Empty category} - marked traces
  \item \textbf{Projection principle} - helps lexical properties to be represented by categorical structure at every level of syntactic representation
\end{itemize}

The combination of X-bar theory and the projection principle welds syntax and the lexicon together.

\textsuperscript{14} Haegeman, Liliane 1991
GB syntax recognizes two types of construction X and X' (read X-bar). X' contains an optional constituent called COMP (which is found in sentence). X-bar theory discusses about lexical categories based on features such as [+N, +V] yielding the categories noun [+N,-V], verb [-N, +V], adjective [+N,+V], pre/post position [-N,-V], etc. and also non-lexical categories such as complementizer and inflections (tenses, argument elements and models). A phrase in X-bar syntax contains at least a head as well as other constituents. Four types of phrases used in X-bar syntax are verb phrase (VP), noun phrase (NP), adjective phrase (AP) and prepositional phrase (PP). Each of these contains the appropriate V, N, A or P, i.e., lexical categories corresponding to major word classes in the lexicon. The phrases in the sentence are tied to the lexicon via their heads. For example:

raaman siitaye vanattileekku ayaccu

"Ram sent Seeta to the forest"

this sentence has the s-structure

```
S'
/ \       
comp  S    
/ \       
N' V'     
/ / \     
N  N  Adv  V' 
/ / / \     
raaman  siitaye  V 
\                             
\ kaaTTil    \ 
\ ayaccu
```

The first NP contain a head N with the lexical item *raaman*, the VP contain a V with the lexical item *ayaccu*. The G.B. theory incorporates a theory of the structure of phrases called X-bar syntax. Its aim is to express generalization about the phrase
structure of all human languages rather than features that are idiosyncratic to one part
of language or to a single language. Within the theory, phrase structure is
comparatively simple which is derived from a few principles and parameters. X-bar
syntax bases the syntax on lexical categories that link with entries in the lexicon.
Phrases in the syntax have heads that are lexical categories linked to lexical entries.
The basic organization of GB has three levels of representation for a sentence -
D-structure, S-structure and LF representation. This can be seen as below.

Lexicon

| government |

D-structure

| move-@ |

S-structure

/ \ / \ PF-representation LF-representation

A simple lexical entry in GB is shown below:

varuka 'to come' : V + [NP].
koTukku 'to give' : V + [NP] + [NP]

X-bar theory is a frame of phrase structure which is on common properties of phrases
like NP, VP, etc. The syntactic categories of phrasal constituents such as VP, NP, etc.
are determined lexically. The VP is a constituent where V is the head, NP is headed
by N, PP is headed by a pre/post position, and AP is headed by an adjective. Phrasal
representation of a Malayalam sentence can be shown as:

VP -> V- (PP)-(NP)
NP -> (Dem) (AP)-(PP)-N
AP -> (Adv) -A -NP
PP -> (Adv) -NP -PP
In GB the government is the structural property which is involved in syntactic process such as theta-marking and case marking. And binding is responsible for assigning an appropriate interpretation to the NPs in a sentence.\(^{15}\)

b. Lexical Functional Grammar (LFG)

LFG is a strong computational formalism, designed by Kaplan and Bresnan (1982). It addresses how to extract grammatical relation from a sentence in positional language such as English. It offers a suitable grammar formalism for expressing exactly the kind of feature structure buildings. Its weakness is that it does not offer any explanation regarding lexical ambiguity, adjuncts, optional theta-roles, and mapping from grammatical relation to theta-roles. LFG postulates two levels of representation: one based on constituent structure (C-structure) which closely resembles the standard phrase structure representation, and the other on grammatical functions called feature-structure (F-structure) such as subject and object. LFG indicates how the grammatical functions can be enumerated exhaustively in the lexicon.

The F-structure contains a set of attribute values, and it may be hierarchial. In the F-structure, subject, predicate, object, object2, tense and specifiers are attributes, and their values are given next to them. Subj., obj., obj2., etc. are called grammatical functions because they are specified by the grammar and not by the lexicon. Predicate is an attribute that maps F-structure which in turn attributes to a semantic representation. The terminal symbols in the grammar yield equations using the lexicon. Lexical ambiguity, adjuncts, optional theta-roles and mapping grammatical relations to theta-roles are left for the lexicon, without offering any linguistic insight as to how to handle them.

Examples of lexical representation for LFG are as follows:

atu 'that' : det ('^spec='atu'), kuTTi 'boy' : ('^pred='kuTTi')('^num=sing).

\(^{15}\) Haegeman, Liliane 1991;187
varuka 'to come': V, varu <(subj)>
uNTaakkuka 'to make' : V, uNTaakku <(subj.)>(obj)>

A number of attempts have been made to use LFG formalism for non-configurational languages.\(^\text{16}\)

c. Paninian Grammar

Karaka theory in Indian traditional grammar refers to the syntactico-semantic arrangement in a sentence. Paninian grammar (PG) which is the Karaka theory aims to construct a theory of human language communication. It combines syntax, semantic and pragmatic in an overall framework. Here lexicon is outside the grammar. Paninian grammar possesses four different levels of representation of a sentence - semantic, karaka, vibhakti and surface levels. Karaka is central to the model. It is case-based, consisting of six cases called karakas, which are in fact syntactico-semantic relations mediating mappings between grammatical relations and semantics. Karaka theory refers to the syntactical arrangements in a sentence. It is mainly identified with the karta (subject) to karma (object) relations. The verb denotes an activity or state while the noun denotes a participant in the activity. This considers verb as the head and noun as modifier.

Karaka is defined as the relation of nouns with verb in a sentence. Whatever things that complete or establish a verb are its karakas. The meaning of a verb is held to be the action directed towards an object or a result. Another important term used by Paninian grammarians is vibhakti. The karakas are represented by the vibhaktis at the morphological level. Malayalam also adopted the same case structure and it includes vocative as a separate case which later grammarians discarded.

In the following sentences the cases or karakas identified are different in each grammar formalism.

\(^\text{16}\) See Akshar Bharathi et al. (1994), Lobiyal (1996), etc. on LFG and non-positional languages like Hindi.
In the above examples, the activities ooTi and uRangi have the noun modifier kuTTi; kuTTi and pazam are the modifiers of the head tinnu. For Paninian tradition all the above cases are karta-karaka relations. But different theories give different accounts of modifiers and kinds of relations between the noun and the verb. GB theory (1982) and Fillmore's Case grammar identify thematic (theta) relations between the subject and object, whereas Paninian theory analyses these in terms of karaka relation. Fillmore (1968) calls the subject of the first two examples as agent and that of the third as an experiencer. It is also noted by all the grammarians that different verbs have different karakas. For example, for sakarmaka kriya (transitive verb) karma (object) is important. For a verb like irikkuka location is important. Rajaraja Varma (1895) says that the words which satisfy the syntactic expectancy of a verb can be considered as its karakas. Following Sanskrit grammarians, he adopted karta, the agent of action, as the most important karaka. The others are karma (object), saaksi, svaami, karana, etc. Karaka marking in the language is mostly through post positions.

e.g., aTikku- 'to beat' has karta (agent) and karma (object) and karana (instrument) karaka.

The assumption made is that each situation to be expressed can be framed as an action or event (kriya) with associated factors (karaka).

3.4. SYNTACTIC INFORMATION

This section looks for the details of syntactic information for a lexicon. Grammar is considered important which defines, with certain rules, how lexical units can be joined
giving a meaning. Different grammatical formalisms, as we have seen, are available for language analysis. Traditional descriptive grammars [Paninian, Bloomfield (1933), Hocket (1946), Harries (1954)], transformational and generative grammars [Chomsky (1965 & 1981), Generalized Phrase Structure Grammar (Gazdar, 1985)], functional grammars [Lexical Functional Grammar (Bresnan, 1982), categorical grammar, Tree Adjoining Grammar (Aravind Joshy, 1983)], etc. give importance to various aspects of grammar and each one has its own treatment to the lexicon.

The concept of word in Indian tradition has been classified by considering the morphological or structural aspects into two divisions - verbal root and nominal root. The former is again divided into vyayaapara (activity) and phala (result), and the latter into vyakti (individual) and jaati (universal). The western concept of word is based on semantic functions, and the words are grouped into various parts of speech. Modern grammarians consider both the structural and semantic aspects. The words are grouped into different categories and sub-categories according to the structure and functions.

3.4.1. Syntactic Properties of words.

Different syntactic categories like noun, verb, adjective, adverb, etc. and various sub-categories are identified by considering the function and usages in the traditional grammars. These groupings appear in modern grammars too, with some modifications. Following are the areas of knowledge at the syntactic level to be incorporated in a lexicon, particularly in COLEX.

Nouns are indefinite in number and belong to different sub-groups. They behave differently at different grammatical environments. Definite number of verbs operate on indefinite number of grammatical rules. Nouns and verbs create maximum number of difficulties while developing a lexicon for NLP application. The unlimited number of vocabulary items, morpheme structure, change in basic form, etc. are some of the issues that need careful attention. The number of items for other parts of speech is

17 See the morphological and lexical classification of words (section 2.4.4).
limited and for this reason research and development has been taking place on problems related to the interpretation of the former classes, i.e., noun and verb.

I. **Noun** has to be dealt by considering the following syntactic issues.

1. Number, gender and case inflections (noun paradigms)
2. Derivative formation.
3. Derivation of a single item from different roots
4. Sub-category
5. Criteria for compound formations i.e., Samasas
6. Syntactic category and meaning after compounding.

Following are the functional subclassifications of nouns.

A. Nouns are sub-classified, considering the properties, into pronoun, proper noun, verbal noun, numeral, etc.

    e.g., avan 'he', mohan 'Mohan', tavaLa 'frog', maraNam 'death'

i. **Number**: Singular and plural are the two class in both the languages. Number is related to countable/non-countable. And countables are classified into singular and plural. The following are some other issues related to the category.

    a, conflict between form and meaning: collective and plural used as singular.
    b, material noun and abstract noun.
    c, nouns without a singular.

Plural forms in Malayalam are with the morphemes: /-kal/, /-maar/, /-ngal/, /-ar/, etc.

    e.g., kuTTi-kal 'boys'
    amma-maar 'mothers'
peNNu-ngaL 'ladies'
av-ar 'they'

Plural forms in Hindi are with the morphemes /-yaam/, /-eel, /-oom/ (oblique), etc.

e.g., ladki-yaam 'girls'
ledk(a)-ee 'boys'

ii. Gender: This differentiates between male female. The following are the three genders in the source language.

1. Masculine
2. Feminine
3. Neuter

There no concept of neuter gender in Hindi and every noun gets assigned a gender (masc. or fem.). All the nouns are grouped either as masculine or feminine and mostly this information is mentioned in the lexicon. Masculine, feminine and non-gender are available in Malayalam. Distinctive morphemes are there to indicate the first two genders. Masculine form in Malayalam can be identified with the morpheme /-an/ and feminine forms with /-i/, /-tti/ and /-al/. Few cases in Hindi can also be identified as feminine with /-i/.

e.g., av-an 'he' wah (H) [no gender distinction]
av-aL 'she'
velakkaar-an 'male servant' naukar [masc]
velakkaar-i 'female servant' naukarani [fem]
ceeTTan 'elder brother' bhaayi [masc]
ceecci 'elder sister' behan [fem]

Pronoun includes personal, reflexive, relative, indefinite, and interrogative. A single form is used for proximate and distant third person in Hindi, whereas in Malayalam
three separate forms are in use.

- vah (‘he’ avan, ‘she’ avaL, ‘that’ atu)
- yah (‘he’ ivan, ‘she’ ival, ‘that’ itu)
- tuu ‘you’ nii
- vaha ‘they’ avar

**Numeral** refers the number system and ordinal represents the adjectival function.

- onnu ‘one’ eeku
- orra ‘single’ eeki
- onnaam ‘first’ pahlaa

**Compound** noun form has more than one free word. The following are the possibilities:

i. **Noun + Noun**
   - accan-amma ‘Father-Mother’ maam-baap

ii. **Adjective + Noun**
   - van-paTa ‘big Army’

iii. **Verb (non-finite) + Noun**
   - paRakkum-kiLi ‘flying bird’

**II Verb** - A verb refers to an action/state/process. The basic syntactic classifications are transitive, intransitive, active, passive, etc. The issues related to the verb category are:

1. Active / passive voice forms and their functions
2. Causative formation and surroundings
3. Transitive-intransitive
4. Tense / Aspect / Mood
5. Infinite / Participle
6. Verbal noun / Gerund
7. Compounds and phrases

A. Simple Verb is a unimorphic verb root, which contains only a single morpheme representing the basic meaning. Simple verbs are more in number compared to compound verbs. The base form of a verb is represented with an infinitive marker -ka. A clear auxiliary is absent in Malayalam, but few verbs like petu, irikku, konTirikku, etc. are added to the main verb to refer the aspect and mood.

\[
\begin{align*}
\text{e.g.,} & \quad \text{pookuka 'to go' (simple) jaana} \\
& \quad V + -konTirikku- \quad \text{(continuous aspect)} \\
& \quad \text{poyikoNTirikkunnu 'keep on going' jaan \ rehii \ hen.}
\end{align*}
\]

B. Compound Verb is a multimorphemic verb with more than two verb roots. It is formed either from two verbs, of which one is the basic, or from one verb and a noun. The first group is called explicator compound verb where there are two finite verbs V1 and V2, and V1 is the main and bears the meaning and V2 is lexically empty and serves as modifier or explicator element of V1. poo, petu, irikku, konTirikku, etc. appear as explicator to refer the aspect and mood.

i. Verb + Verb
   \[
   \begin{align*}
   \text{e.g.,} & \quad \text{vannu-kaNTu 'come-see' aakar_deekkaa} \\
   & \quad \text{kaLanjnu-pooyi 'loss (went)' - 'lost' nastu_huvaa}
   \end{align*}
   \]

ii. Verb + Auxiliary
   \[
   \begin{align*}
   \text{e.g.,} & \quad \text{vannirunnu 'had come' aa gayaa dhaa}
   \end{align*}
   \]

iii. Noun + Verb
   \[
   \begin{align*}
   \text{e.g.,} & \quad \text{mayaatayi 'become corpse' mar_geyaa}
   \end{align*}
   \]
The inflectional forms are formed on the basis of tense, aspect and mood. Tense forms are identified with the tense markers\(^{18}\) - the future marker is \(-/uml\), present marker is \(-/unnu\) and the common past markers are \(-/u\) & \(-/i\).

a. Tense

i. Past Tense markers \([-i, -ccu, -nnu, -u, -nnu]\)
   
e.g., ravi kuTtiye aTicz 'Ravi beat the child'
   
   Ravi boy-Acc. beat-past.
   
   vanTi roodil ninnu 'bus stopped at the road'

ii. Present tense \([-/unnu]\)
   
e.g., rema ammakku paisa koTukkunnu 'Rema is giving money
   
   Rema mother-Dat. money give-Past to mother'
   
   revikku panikkunnu 'Ravi is feeling fever'

iii. Future Tense \([-/um]\)
   
e.g., ayal varum 'he will come'

Transitive, intransitive and causative are the basic syntactic sub-categories. Transitive verb always need an object whereas an intransitive does not. Causative needs an additional agent to perform. Causative verb is a form of verb, the action of which was performed by an agent because of the inspiration by another agent.

   e.g., raaman goopiye kaapi kuTippiccu 'Ram made Gopi to drink coffee'
   
   devi ko dantu dilvaya 'Devi is caused to feel pain'

Voice indicates relation of the subject to the action. Active and passive are the two voices, where active indicates subject to do an action or to become something. Passive

\(^{18}\) See section 2.4.4. about tense formations in Malayalam.
represents the object as to act as the subject. For example the grammatical subject which is not the logical subject but logical object as raamun in the following sentence.

eg., raaman kollapettu. 'Ram is killed'

**Participle** (Relative): This is an adjectival form derived from any verb. It keeps all inflectional qualifications, and instead of the finite verbal end /-u/ adjectival marker /-a/ is added.

eg., raaman pooya viiTu 'the house where Ram did go'
joo ghar raam ne geya

**Verbal Participle** is a non-finite verbal form to modify a finite verb.

e.g., raamanu paisa kiTTi tuTangngi 'Ram starts getting his money'
raam ko paisa milna suru kii.

b. **Aspect and Mood**

**Aspect** is the form of verb indicating the type of character to the action, like perfect, imperfect, continuous, etc. Grammatical components of all aspects and moods in Malayalam involve verbal compounds in which at least one of the constituents functions as auxiliary to what is often distinguished as a principal verbal function. Auxiliary verbal forms in Malayalam can occur after verbal forms belonging to the categories called verbal participle, purposive, infinitive and verbal noun after finite verbal forms, ending in present or future tense suffixes, etc.

Following are the semantic features concerning the aspectual construction in Malayalam. The verbal roots shown in the brackets are added to the main verb to denote the given meanings.
1. Perfect or completion refers action as a whole: (kazi-'to be over', iTTu-'after', vazi-'to place')

   eg., avan pooyikazinju 'he has already left'
   avan iviTe vanniTTu pooyi 'he went having came here'

2. Progressive or continuance refers action in progress: (kol- 'to continue', poor- 'to follow')

   eg., aval ciriccukoNTu naTannu
        'she walked laughing continuously'

3. Probability mentions uncertain action: (kaaNu 'to see', irikkuu 'to sit', eekku 'to take charge')

   eg., tiivaNTi samayattu pooyi kaaNum
        'train might have gone on time'

4. Manner shows the mode of action: (kalay- 'to give up', poo- 'to go', iri- 'to sit', nil- 'to continue', tulakku- 'to finish', aruL- 'to order')

   eg., raama siRRikku pooyikaLanju
        'Ram went to city intentionally'

5. Benefit is advantageous for the actor: (etu- 'to take', kol- 'to held', tar- 'to give', koTu-'to give away')

   eg., ravi pustakam valiccetuttu
        'revi pulled and took out the book'

6. Attempt is drive by the actor: (nook- 'to see', poo- 'to keep go')
eg., siita pookan nooki 'Sita tried to go'
   siita raamane talli nooki 'Sita tried pushing Ram'

**Mood** indicates the manner of action. Both the languages have some important moods like indicative, imperative, and subjunctive, which show the states of action and can be identified by the presence of the respective markers.

i. **Imperative** has the marker $+\text{fin} (0) = V \text{stem}$
   e.g., poo '(you) go' jaa

ii. **Imperative-polite** has the marker $+\text{fin},-\text{indi},+\text{pol} (u) = V \text{stem}+u$
    e.g., pook-u '(you) please go' jaavoo

The polite form is with $+\text{fin},-\text{indi},+\text{pol} (u+in) = V \text{stem}+u+in$
    e.g., pook-u-v-in ' (you-hon.) please go' jaayiye

iii. **Optative** is added with $+\text{opt} (aTTe) = V \text{stem} + aTTe$
    e.g., pook-aTTe 'let (him) go' jaane_do

iv. **Obligatory** has the final form with $+\text{obl} (aNam) = V \text{stem} + aNam$
    e.g., pook-aNam ' (he) wants to go' jaana_hen

v. **Permissive** behaves with the marker $+\text{perm} (aam) = V \text{stem} + aam$
    e.g., pookaam ' (shall we) go' jaa_saktaa_hen

vi. **Infinitive** has the three forms

   a. with $+\text{purposive} (aan) = V \text{stem} + aan$
      e.g., pookaan 'to go' jaane_keliye
b. Habitual and continuous action is with the form +imminent (aar) = V stem + aar
   e.g., pookaar 'used to go' jaatte_rehta

c. Simultaneous action is with +concur (ee) = V stem +ee
   e.g., pookave 'while going' jaattee_vaktu

vii. Negation in Malayalam is generally lexical, and behaves syntactically as a full verb.
   e.g., raman paisa koTukkunnilla raam paisa nahiim detta
        'Ram is not giving money'

In Hindi a verb is inflected according to the gender, number and person of either subject or object, or it has reference only to the action. Hindi verb has three constructions.

a. Subject construction: The verb has same number, gender and person as a logical subject.
   e.g., raam jaa rahaa he 'Ram is going'
        siita jaa rahii hen 'Seetha is going'

b. Object construction: the verb has same gender, number and person as its logical object (the person or thing to which the action is directed).

c. Neutral: the verb agrees neither with the subject nor with the object person, gender and number, but is always placed in the third person singular masculine.

III. Adjective always appears before a noun and qualifies it. Functional classes of adjectives are:
1. Simple
e.g., nalla 'good'

2. Ordinal
e.g., oRRa 'single'

3. Relative participial is derived from a finite form of a verb, and it retains the tense, aspect, mood qualities.
e.g., oTunna 'running'
ciriccucoNTirikkuna '(one) who keeps on laughing' joo hastee rehtee hen.

VI. Adverbs qualify verbs, and they are functionally temporal, spatial, directional, aspectual, etc.

1. Temporal
e.g., vaLare 'very'

2. Aspectual
e.g., patukke 'slowly'

3. Directional
e.g., niiNTu 'lengthy'

4. Spatial
e.g., uyaratil 'at high'

V. Post positions are bound forms which never take any inflectional changes and come always after a noun to refer cases. Post positions are added to the case affixes to indicate special meaning too. Different grammarians in Malayalam have listed them from 80 to 150.19

   e.g., koNTu 'by' see (instrumental)
ninnu 'from' see (ablative)
opole 'like' taraha (comparison)

19 See Radhakrishan, S. (1994) for a detailed analysis of post position in Malayalam.
f). **Coordinators** and other indeclinables are used to coordinate two nouns, or two phrases, or two clauses or even two sentences. Following are the important items of the group.

- um 'and'
- allenkil 'or'
- atinaal 'so'
- enkilum 'but'
- atukoNTu 'because'
- entennaal '..that..'
- ennaal 'but'
- enkil 'if'
- allenkil 'if not'
- enniTTum 'even though'
- tanneyalla 'still then'
- -urn 'and'
- yaa (option - sentence)
- isliyee (reason)
- too_bhii (restriction)
- kyoom_ki (reason)
- ki (clause)
- leekin (restriction)
- too (option)
- nahiin_too (option)
- phir_bhi (addition)
- balki (addition)

**Number-Gender agreement**: There is no gender and number agreement in Malayalam, while a strong agreement system is available in Hindi. Gender distinction, and the GNP marker are the two most important issues for Hindi.

### 3.5. COMPOUNDS AND PHRASES

Compounds formation is more in Malayalam due to the agglutinative nature of the language, whereas in Hindi it is comparatively less. Samasas, nominal compounds, compounding of nouns with cases are more in Malayalam, but less in the target language.
Phrases are the following with respect to the main category of which it is made.

1. Noun phrase with noun as the head
2. Verb phrase with verb as the head
3. Adjectival phrase with an adjective as the head
4. Adverbial phrase with an adverb as the head.

Each one takes a particular position in a sentence and functions independently. Usually postposition does not make itself as head of a phrase.

Not much differences are there at the phrasal level between the two languages. Post position is found in both the languages and always after the noun. Adverb appears before the main verb.

A phrase has at least two components, a lexical category and the projection. The phrasal category is headed by the lexical category (Chomsky-1982).

A phrase for X-bar syntax can be represented by a general schema:

```
X''
\ /   \ /
./    \
X'
\ /   \ /
./    \
X
```
Here X represents any lexical category (N,A,V,P) and the phrasal projection is represented by means of the layered representation.

3.6. SEMANTICS AND LEXICON

Semantics deals with the meanings and the study of meaning can be of a word or a construction, i.e., it could be at the lexical level or phrasal level or sentence level. A lexicon deals with semantics of words and expressions like compounds and phrases. For recent grammatical theories semantic issue is the most important one particularly for the NLP applications. In computer understanding of language, semantic interpretation is the process of determining the meaning of the input. Some of the semantic issues that are going to be addressed here are the identification of the meaning of a word, or a sense of a construction (including pronoun, reflexive, definite reference expression, etc.), identifying meaning of forms to temporal objects, etc.

Semantic representations are structured configurations of semantic units which are grounded by the grammatical and conceptual knowledge. Semantic representations are abstract representations of meaning in that they are not identical with specific conceptual interpretations which an expression may have in particular context of use. They are the unique level on which the principles of compositionality of meaning are operative. Representation of knowledge can be done by a pattern of interconnected nodes and arcs. The earliest semantic networks were designed as intermediate languages for machine translation and most of the versions are still strongly oriented towards the feature of natural languages. But the recent approaches have grown in power and flexibility to compete with frame systems and logic programming systems as general knowledge representation of languages.

---


The simplest networks used in NLP are relational graphs. These graphs consist of nodes connected by arcs. Each node represents a concept and each arc represents some relationship between the corresponding concepts. For example the sentence:

raaman veLLam kuTicuu 'Ram drank water'

the verb kuTiccu has two nodes, one the subject raaman and object veLLam connected by arcs.

kuTiccu -- raaman(subj.)

| veLLam (obj.)

A semantic network is a mathematical frame, a directed graph composed of nodes and links, each link running from one node to a second node. Semantic networks vary widely in expressive power. Semantic networks may simplify the search for information relevant to a given entity because the links between nodes are directly accessible from each connecting node.

In the process of connecting a text with its meaning, the primary role of semantics is to decode the knowledge each lexical item possesses. 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 lexical entries in terms of arguments taken by 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. The theory of lexical semantics includes a component of lexical knowledge that describes the syntactic encoding of arguments. The lexicon extends the system's abilities for semantic interpretation too.

---

22 Martha, W. Evens; 1988
The purpose here is not to provide a thorough-going description of semantics of lexical items. But, we will focus on what we consider to be essential for a common purpose COLEX and the representation in the COLEX. Following are some terminologies related with semantics used by important grammatical formalisms. We will see how they are related or differed, and how they help in solving the problems of semantic parsing and resolving ambiguity.

3.6.1. aakaanksa/ yoogyata/ sanamidhi

These three grammatical terms of Sanskrit in Indian theories of meaning are widely used by other Indian languages as well. The terms are translated as **syntactic expectancy, logical consistency, and phonetic contiguity**. They refer to the expectation of what a verb or a noun should take and the constraints. All these are part of Karaka theory which is mainly identified with the *karta* (subject) to *karma* (object) relations. The meaning of a verb is held to be the action directed towards an object or result.

*Aakaanska* is the term which explains how syntactic unity is effected among various isolated words which compose the sentence. It is the most important among the three factors which constitute the basis of syntactic unity.$^{24}$ *Aakaanska* consists in a word not being able to convey a complete sense in the absence of another word. Or we can say it is the desire on the part of the listener to know the other words to complete the sense. A word is said to have aakaankasa for another, if it cannot, without the latter, produce knowledge of its inter-connection in an utterance.

For example the verb *kuTiccu* 'drank' in the following sentence expects an object and an animate noun to experience.

i. ravi caaya kuTiccu 'Ravi drank tea'

---


24 Kunjunni Raja, 1942; 151
Certain words necessarily require certain other words to complete the sense. Thus a noun in nominative case requires a verb to convey a complete meaning. For example, a verb koTukkuka 'to give' has expectancy for a word denoting some object. Kunjunni Raja (1942) opines that the aakaanksa that holds between words in a sentence is a grammatical one. For Rajaraja Varma (1895) aakaanska is the relation between the words in a sentence like kinship relation between a father and son.

Yogyata is the logical compatibility or consistency of the words in a sentence for mutual association. When the meaning of a sentence is not contradicted by experience, there is yogyata or consistency between its constituent words.

Samnidhi is explained as the condition that words in a sentence should be contiguous in time. This contiguity or proximity is the uninterrupted utterance or unbroken apprehension of words when they are in juxtaposition.

3.6.2. Argument Structure

This term is almost equivalent to the aakaanksa of Panini. The term argument was first used by Jakendoff (1964), and later adopted by various grammarians. It is the lexical representation of grammatical information about a predicate, and it is the part of lexical entry.25 Argument is the expectancy of number of items a word can take. The argument structure of a verb determines which elements of the sentence are obligatory. For example a transitive verb takes two arguments of which one as the subject and the other as the object. Likewise a di-transitive verb takes two objects and a subject, whereas an intransitive verb takes only one argument. If a verb expresses an activity involving two arguments, there will have to be at least two constituents in the sentence to enable these arguments to be expressed. Even nouns, adjectives, adverbs, postpositions, etc. can also be identified as having one or more arguments.

---

25 Grimshaw, J.B. 1990
We can represent the argument structures of verbs and adjectives as:

- Verb
  - { one place argument structure.
  - { two place argument structure.
  - { three place argument structure.

Adj { one place & two place argument structure

These conceptually defined argument structures can partly replace the classification of verbs in terms of either transitivity levels or subcategorization like transitive-intransitive frames.

The representation of the arguments of some Malayalam verbs can be as follows:

<table>
<thead>
<tr>
<th>Malayalam</th>
<th>English</th>
<th>Argument Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>kaNTu</td>
<td>'saw'</td>
<td>1(NP) 2(NP)</td>
</tr>
<tr>
<td>uNTaakki</td>
<td>'made'</td>
<td>1(NP) 2(NP)</td>
</tr>
<tr>
<td>koTuttu</td>
<td>'gave'</td>
<td>1(NP) 2(PP) 3(NP)</td>
</tr>
<tr>
<td>ciriceu</td>
<td>'laughed'</td>
<td>1(NP)</td>
</tr>
</tbody>
</table>

Argument structure of adjective and postposition can be seen as follows:

<table>
<thead>
<tr>
<th>Malayalam</th>
<th>English</th>
<th>Argument Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>nalla</td>
<td>'good'</td>
<td>Adj 1(NP)</td>
</tr>
<tr>
<td>kuSumbuLLa</td>
<td>'envious'</td>
<td>Adj 1(NP) 2(NP)</td>
</tr>
<tr>
<td>ninnu</td>
<td>'from'</td>
<td>PP 1(NP) 2(NP)</td>
</tr>
<tr>
<td>naTukku</td>
<td>'between'</td>
<td>PP 1(NP) 2(NP) 3(NP)</td>
</tr>
</tbody>
</table>

3.6.3. Thematic Representation and Theta Role

Theta-theory is the part of $X'$-theory that handles the relationships such as grammatical functions (GF) of subject, object, etc. In the following sentences the role of subject changes, though the actor is the same; the grammatical subject is not the 'mother' in the second sentence, but somebody else.
amma makaLe aTiccu 'Mother hit her daughter'
amma makaLe aTippiccu 'Mother caused to hit her daughter'

Theta role or thematic representation also is about the expectation of a particular verb from nouns. It raises the following questions:

- what roles are there and how can they be identified?
- where do 0-roles fit in a formal approach to lexical semantics?

Thematic roles can be explicitly linked with functions. Theta-theory concerns the fundamental logical notion 'argument of' (such as case) that any theory of grammar must account for. Within GB theory theta-theory takes a specific form that could not be anticipated on the basis of the logical notion 'argument of' alone. It has specific empirical content and exists in its interaction with units.

A lexical entry for a verb needs to specify the theta roles that go with it. So they are part of the contents of the lexical entry for an item, which gets assigned to a relevant NP in the sentence.

3.6.4. Projection Principles

Projection principles are the properties of lexical entries projected onto the syntax of the sentence. GB theory emphasizes the lexicon: speakers know what each word in the language means and how it is said and how it behaves syntactically. The theory integrates the syntactic description of the sentence with the properties of lexical items. The projection principle requires the syntax to accommodate the characteristics of each lexical item. Each lexical item in the language has idiosyncratic properties of its own recorded in its lexical entry.

3.6.5. Subject And Object

Subject is the term to refer the agent of an action, whereas the term object means the
recipient of the action. These concepts are very old and still continue to appear in different forms in present-day grammatical discussions. They are equivalent to the karta and karma of Paninian tradition. In the following sentence raainan is the subject and the object is poottu.

   e.g., raaman pottine talli 'Ram beat (past) buffalo'

Since Malayalam and Hindi are non-positional languages subject and object can appear any position unlike in English. Not the position but the case marker indicates the function.

3.6.6. Selectional Restriction

When a verb subcategorizes a complement belonging to a particular category, it is not the case that any item belonging to the relevant category can function as a complement. Instead, there are restrictions in the choices of complements, as seen in the example:

   kumaar pad konnu 'Kumar killed a dog'
   *maram paTTiye konnu 'tree killed a dog'

Here konnu 'killed' needs a NP, which must be an animate one like kumaar. Selectional restriction remains in other situations as well.

   V + Subj ; *paSu paaTi 'cow sung'
   A + N ; *karutta pakal 'black day'
   Adv + V ; *desyattil ciriccu 'laughed angrily'
   Subj + Adv ; *niLattil curukki 'lengthly shortened'

All these constructions are grammatically well-formed but semantically unacceptable. Animate/ inanimate, transitive/ intransitive, male/ female, abstract/ concrete, etc. are some other areas where selectional restrictions can be found.
3.6.7. Case Relations

Case is the grammatical or syntactic marker attached to a noun to show the relationship with the verb. Many languages use surface case inflections to indicate which noun stands in which graded relation to the verb. Case is most easily observed and studied in languages that have rich case morphology.

Fillmore (1968) in his classic case theory has defined six cases viz, agent, instrumental, dative, factitive, locative and objective. Sentential examples which don't fit into the set of cases provided by him suggest that the range of semantic concepts captured by the six cases is too limited. The problem seems to be not with the number of cases, but rather with the absence of criteria to determine what qualifies as a case. Case relations are more semantic and explicit in Malayalam.

Within the LFG framework, case is not invoked to account for the distribution of lexically filled NPs, as it is in recent Chomskyan frameworks. The distribution of arguments that may be subjected to grammatical or anaphoric control is handled in LFG by the theory of control, in terms of grammatical functions rather than syntactic positions. Case is used in the Chomskyan framework to define the distribution of NPs and to trigger movement. In LFG such alterations are determined by the mapping from argument structure to grammatical functions.

Lexical Conceptual Structures of Jackendoff are closely tied to syntactic structure and are nonetheless deep semantic representations and they are well suited to provide an interlingua. Independently needed, language-specific mapping mechanisms can then be exploited to account for different divergent syntactic expressions of common interlingua. The essential units of conceptual structure are conceptual constituents, each of which belongs to one of a small set of major ontological categories such as thing, event, state, action, place, path, property and amount. The conceptual structure of lexical item is an entity with zero or more open argument places. The meanings of the

---

26 Fillmore 1968;24-5
syntactic complements of the lexical items fill in the values of the item's argument places in the meaning of the sentence.

The generalization of lexical items across semantic fields is by no means totally free. Each word is quite particular about what fields it appears in. The conceptual function of a lexical entry can be shown as:

\[
[ \text{pooku } ([], \text{from}([]))]
\]

\[
\text{to}([]))]
\]

Dorr (1993) uses Jakendoff's LCS representation for lexical organization with two levels of description. Language-independent and language-specific parametric specification generates syntactic realization of the word and its arguments.

3.6.8. Ambiguity: Lexical and Semantic

Some semantic problems involve ambiguity, that is to say, one word or a construction can be interpreted in more than one ways. It is of three kinds - categorical ambiguity, transfer or translation ambiguity and homophonous and polysemous ambiguity.

A lexicon must be a resource to store information for disambiguating word sense. In a bilingual lexicon various information of the word in same language and in the target language, such as semantic tags to identify various senses to the word, sense disambiguation rules for eliminating ambiguity, semantic relationship between words, etc. should be stored.

a. Categorical Ambiguity: A given word may be assigned to more than one grammatical or syntactic category (e.g. N, V, Adj.)

   e.g., kaLi 'to play'(V), 'play/drama'(N)

b. Transfer Ambiguity: The transfer ambiguities occur in bi/multi-lingual context, where a single source language word can potentially be translated by a number of
different target language words or expressions. Source language word itself is not ambiguous, and it is ambiguous only from the perspective of another language. Most of the differences in the lexical system of languages arise from the conceptual differences. But, these can come also from the stylistic or grammatical differences.

e.g., ceecci 'elder sister' behan
aniyatti 'younger sister' behan

c. Lexical Gap is another issue in the development of bilingual lexicon. Lexical gap is the absence of one single word in a language corresponding to a lexical item in the other language. There are no single specific words to a particular cultural item such as iddaly/ vada/ kayar/ kanji.

e.g., aLiyan 'cousin' baanja/saala/mameera_bhaai

In these cases the lexical items are often borrowed untranslated, or words which carry a range of meanings in one language are rendered in another by a variety of words or a group of explanatory words so that the meaning is conveyed properly.

Some lexical gaps are the result of productive derivational morphology in one language which is not parallel in another.

e.g., njangngal ham 'we' (excl)
nammal ,, 'we' (incl.)
Both these gives only one word in Hindi
hindikkaaran hindivaala 'a man of Hindi language'

3.6.9. Anaphora

Anaphora is a term used to refer to an oblique reference being made to an entity mentioned explicitly elsewhere in a text. The most frequent linguistic device for this is the use of a pronoun like atu 'it', avan 'he', avar 'they', etc. or a demonstrative
pronoun like _itu_ 'this', _atu_ 'that' or phrases like _munpaRanja_ 'aforesaid'. There are three types of anaphora pointed out.

1. A full noun phrase, not linked to an NP in an argument position.
2. Pronoun such as _avan_ 'he' and _avane_ 'him', not to be linked to an NP in an argument position in the same domain.
3. Reflexive element such as _tanne_ 'himself' picks up its reference from the subject NP or the NP on which a reflexive is dependent for its interpretation. This usually appear antecedent to the reference.

   eg., ravi tannattaan paziccu 'Ravi cursed himself'.

3.7. SEMANTIC INFORMATION IN A LEXICON

The following are the semantic informational details needed as essential components which the structure of lexicon should have.27

3.7.1. Noun attributes

1. **Concrete and Abstract concepts.**
   a. Common noun represents a collective group.
      e.g., amma 'mother', mrgam 'animal'.
   b. Abstract refers an abstract entity.
      e.g., kaLLam 'theft'
   c. Concrete is an object of existence
      e.g., kallu 'stone'

27 See section 1.10 and section 4.3.
2. **Animate & inanimate** represent living or non-living

- e.g., inrgam 'animal', kiLi 'bird', miin 'fish',
  - wild-domestic: aaTu 'goat' pucca 'cat' - puli 'tiger' simham 'lion'
  - male - female: kaaLa 'bile', paSu 'cow' etc.
  - maram 'tree', mala 'mountain', paaRa 'rock',
  - veLLam 'water',
  - looham 'metal', uppu 'salt', etc.

a. **Human & non-human** refer human being in contrast to every other animate beings.

- aaNu-peNNu 'Male-female', veluppan-karuppan 'white-black',
  - Asian-African, iLam-murru 'young-old', working- nonworking, etc.

b. **Animal** represents all fauna except human being.

- sastani 'mammal', 'paksi 'bird', kiiTam 'insect', iicca 'flee', urumpu 'ants', kutira 'horse' aana 'elephant', etc.

c. **Mobile** (self & non-self) & immobile differentiate between mobility including that by oneself or by others, and immobility.

- manusyan 'human', mrgam 'animal', vaNTi 'vehicle', nadi 'river',
  - vaayu 'air', plane, mastyam 'fish', etc. viiTu 'house'

3. **Range or part** refers to the position of part in the whole, or grade in the scale.

- tala 'head', kai 'hand', hrdayam 'heart', manasu 'mind', viral 'finger', tooL 'shoulder', etc.
4. **Measurement** gives quantity and assesses quality.

   meter, naazika 'furlong', Kilograam, tooila 'Pound', liter, uria 'ounce', byte, etc.

3.7.2. **Verb attributes**

a. **Semantic types** of verbs are grouped after considering the deep semantic aspects.\(^{28}\)

The following are some of them.

1. **Action verb** refers an action, which again can be grouped into three.

   a. Movement: pooyi 'went', niintuka 'to swim'
   b. Transfer: koTuttu 'gave'
   c. Expression: nookki 'looked'

2. **Emotion** verb expresses a feeling or an emotion.

   e.g., revikku **panikkunnu** 'Ravi is feeling fever'

3. **Static** verb represents a state of being.

   e.g., vaNTi **nilkkunnu** 'bus is standing'

4. **Experience** shows an experience to a subject.

   e.g., raaman paNam koNTu **sukhiccu**

   'Ram felt comfort with money'

b. **Case / Karaka Relations** of Verbs

---

\(^{28}\) See Mohan Bhas (1989) and Kapil Kapoor (1985).
i. **Accusative verb**

   e.g., *revi kuTTiyen aTiccu* 'Revi beat the child'
   Revi boy-Acc. beat-past.

ii. **Dative verb**

   e.g., *rema ammakku paisa koTukkunnu*
   'Rema is giving money to her mother'
   Rema mother-Dat. money give-Past

iii. **Agentive verb**

   e.g., *ayal varum* 'he (hon.)will come'

iv. **Experiential verb** takes a dative subject

   e.g., *deevikku veedaniceu* 'Devi felt pain'
   enikku daahikkunnu 'I am feeling thirsty'

v. **Theme**

   e.g., *maNi paNi ceytu* 'Mony did the work'

vi. **Goal**

   e.g., *ceecckku sambalam kiTTi.* 'Sister got her salary'

vii. **Source/Location/Manner/Time**

   e.g., *ravi innale iviTe ninnu bassil madraasinu pooyi*
   'Ravi went to Madras by bus yesterday from here'

c. **Argument structure.**

   a. One place argument (intransitive verbs)
   b. Two place argument (all intransitive verbs and transitive verbs.)
   c. Three place argument (transitive verbs)
3.7.3. Adjectival attributes

1. Quality
   - nalla 'good' acca
   - veLutta 'white' sapheedu

2. Quantity
   - valiya 'big' beRaa
   - niranjna 'filled' phara

3. Participial
   - pazutta 'ripe' pakkaa
   - koTutta 'given' diya_hua

4. Ordinal
   - aadya 'first' pahalaa

3.7.4. Adverbial attributes

1. Spatial
   - uyarattil 'at high' uunccayi_par
   - puRattu 'out' baahar

2. Temporal
   - neeratte 'early' deer_pahale
   - kaalattu 'on time/ in the morning' samaya_par/subaha

3. Manner
   - patukke 'slowly' diiree_se

4. Type
   - poole 'like' taraha

3.7.5. Attributes of Post Positions

1. Postposition representing case relations
   - meel 'above' uupar, koNTu 'by/with' see
2. Coordinatives
   um 'and' aur, -ee 'while' -ee

3. Comparatives
   poole 'like' taraha, kaal 'than' see

3.8. PRAGMATIC AND OTHER INFORMATION

The grammatical marker each category takes in a language is conveyed by the grammar rules. Grammar tells whether a construction is grammatically right or wrong. But it fails to specify whether it is semantically acceptable or not. By simply following grammar rules, semantically ill-formed sentences may also be generated. To avoid this, more non-linguistic knowledge should be incorporated. This may be pragmatic or other levels of knowledge.

Pragmatic information includes world knowledge which are non-linguistic like relational, ambiguity, domain specific, etymological, etc. These information also need to find a place in a complete lexicon. If a lexicon has all these pragmatic information, then solving the ambiguity and understanding natural language is very easy. But no lexicon can possess all such information, like a human being can. Because human beings have intelligence and their memory is indefinite while computer has no reasoning power and its memory is limited. Following are some of the extra-linguistic information relevant to the lexical construction.

a. Relations: Lexical items are related with one another in some way or other.

   e.g., aaNkuTTi 'boy' related to penkuTTi 'girl',
   pazaya 'old' to putiya 'new', etc.

b. Homophones and Homonyms: A lexical unit having same phonemes or spelling but different etymology and meaning is homophone. And words found to have various senses are called homonyms (naanaarthai).
c. **Synonym:** Several words used for the same sense are called synonyms, *(paryaaya sabda)*.

   e.g., aana, kari, gajam, hastni, etc. 'elephant'

d. **Antonyms:** Words having opposite meanings are antonyms.

   e.g., nalla 'acca', ciitta 'bura'

e. **Semantic domain:** Lexical units with domain specific and domain free specifications.

   e.g., the word *paTam* means 'movie' in film world and 'picture' in art field and 'photo' in photographic field.

f. **General domain:** Lexical unit having same meaning in different domains.

   e.g., the above word *paTam* in general sense 'picture'.

### 3.9. REPRESENTATION OF SYNTACTICO-SEMANTIC INFORMATION IN THE SYSTEM

Representing the syntactic and semantic information in a computational lexicon is one of the main tasks to be discussed while constructing a COLEX. This area received much attention and various proposals have come out but no consensus is in sight.

We have seen a number of syntactically related lexical information like category, subcategory information like transitive intransitive forms, syntactic transformations like causative formations, passivisation, nominalisation, verbal noun formation, and, of course, tense specification. Information such as number gender agreement, projection principles, subject-object position, etc. have to be placed in the lexicon in certain ways. As seen earlier most of the lexical information are kept in the lexicon by means of binary markers (+/-).
Different grammar formalisms expect the lexical information in different ways. One cannot follow all formalisms in single system. But a common procedure can be used which can be modified according the needs of formalisms.

Semantically related lexical information to be incorporated in a lexical system are argument structure, case relations, animate-inanimate, etc. Transitive-intransitive specification and the argument structures are central to lexicalized grammars. Number-gender and case specifiers are required for LFG in the binary terms.

As mentioned above the purpose for which and the way a dictionary is used affects its contents. A number of views have been expressed, by different groups, regarding the type of entries and information required for a lexicon.

Organizing the entries and their related information of both the languages in a bilingual lexical data for an easy access is the main task. The method of arranging and transforming the lexical data depend mainly on the purpose it is going to be used.

All the discussions are done based on the source language as the model. Minor modifications and changes are needed to the target language. So the source language and the related attributes and its equivalents in the target is the pattern we have adopted.

The next chapter discusses the lexical design of a COLEX with reference to Malayalam-Hindi and the information required for the purposes like syntactic and semantic parsing for MT or natural language understanding, and also for CALL/CALT.