CHAPTER II

PHONOLOGICAL AND MORPHOLOGICAL INFORMATION
2.1 INTRODUCTION

This chapter discusses the phonological and morphological aspects of a lexicon in general, and of a COLEX in particular. Phonological knowledge is concerned with the sound aspects and the structure of words including the internal structure. These information are needed in lexical systems which support speech and sound. This is also needed for spell-checking, morphological analysis and generation. Characteristics of sounds and the syllabic structure of both the source and target languages and the morphophonemic alternations called sandhi rules, which are morphologically important in these languages, are the issues discussed at the outset. Classifications of lexical items depending on phonological, morphological and word formation rules, etc. are the main issues discussed. Phonology became important after the introduction of generative phonology (Chomsky, 1971), which defines various forms of a lexical unit.\textsuperscript{1} Morphological information is needed to develop morphology based systems for word analysis and generation. The following discussion takes examples mostly from the source language, Malayalam.

Malayalam is an agglutinative language, and the formation of grammatical items is through concatenation of morphemes one after the other.\textsuperscript{2} So is the case with Hindi, where suffixation is the main word formation method in both inflection and derivation.

Lexicons, meant to develop speech based systems like speech recognition for speech-to-text, or speech generation for text-to-speech, etc.\textsuperscript{3} need the relevant phonetic and phonological information at different levels. In text-to-speech a text is taken as input, and the speech corresponding to the text is produced. In speech-to-text, the reverse takes place. The morphological rules and word formation theories of language are important to a number of applications, which set how grammatical units are formed.

\begin{itemize}
  \item \textsuperscript{1} Mohanan, 1984; 72-78
  \item \textsuperscript{2} Ezhuthachan, K.N., 1984
  \item \textsuperscript{3} Eric Laporte, 1989 & Sproat, R., 1992: 7-9.
\end{itemize}
from the primitive forms. Lexical systems which support morphological analysis and generation depend primarily on phonological and morphological knowledge.

2.2 PHONOLOGY AND LEXICON

Each language has its own sound system by which that language is pronounced or transcribed. The basic unit of a language is word, and a word is made of phonemes. A dictionary is a lexical data-base which provides information about the vocabulary of the language. It specifies for each word of the language its associated morpha-phonemic features too.

2.2.1. Phonology and the Syllabic Structure of Malayalam and Hindi

Malayalam and Hindi share many phonological and syllabic structural features. Both the languages have adopted most of the phonological features from Sanskrit.

Malayalam has 53 alphabet, in which 16 are vowels and the rest consonants. Among the sixteen, there are seven basic short vowels (a,i,u,e,o,r,l), seven equivalent long vowels (aa,ii,uu,ee,oo,rr,ll) and 2 diphthongs (ai, au)\(^4\).

### Phonemic Chart of Malayalam - Consonants

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Palato-Alveolar</th>
<th>Retroflex</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Voice less Stops</strong></td>
<td>p</td>
<td>t</td>
<td>c</td>
<td>t</td>
<td>s</td>
<td>k</td>
<td>g</td>
</tr>
<tr>
<td><strong>Voiceless Aspirated Stop</strong></td>
<td>ph</td>
<td>h</td>
<td>ch</td>
<td>th</td>
<td>k</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Voiced Stop</strong></td>
<td>b</td>
<td>d</td>
<td>j</td>
<td>d</td>
<td>g</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Voiced aspirated Stop</strong></td>
<td>bh</td>
<td>dh</td>
<td>jh</td>
<td>dh</td>
<td>gh</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nasal</strong></td>
<td>m</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td><strong>Fricative</strong></td>
<td></td>
<td>s</td>
<td>s</td>
<td>s</td>
<td>h</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lateral</strong></td>
<td>l</td>
<td>e</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tap</strong></td>
<td></td>
<td>r</td>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Friction less continuant</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Glide</strong></td>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>y</td>
<td>u</td>
</tr>
</tbody>
</table>
Vowels (Same for both the languages)

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>i ii</td>
<td>u uu</td>
<td></td>
</tr>
<tr>
<td>Mid</td>
<td>e ee</td>
<td>ô</td>
<td>o oo</td>
</tr>
<tr>
<td>Low</td>
<td>a aa</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Phonemic Chart of Hindi - Consonants

<table>
<thead>
<tr>
<th>Bilabial</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Palato-Alveolar</th>
<th>Retroflex</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice less</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stops</td>
<td>p u</td>
<td>c ē</td>
<td>t ŋ</td>
<td>k ŋ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voice less</td>
<td>ph</td>
<td>ch</td>
<td>th</td>
<td>kh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspirated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stop</td>
<td>b d</td>
<td>j d</td>
<td>g ŋ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voiced</td>
<td>bh</td>
<td>jh</td>
<td>dh</td>
<td>gh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>voiced</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aspirated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stop</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>m n</td>
<td>n ŋ</td>
<td>n ŋ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>s s</td>
<td>s ṭ</td>
<td>s ṭ</td>
<td>h ṭ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral</td>
<td>l r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap</td>
<td>r ñ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glide</td>
<td>w ṭ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>y ṭ</td>
</tr>
</tbody>
</table>
Hindi has a similar phonemic inventory. There are few exceptions in number. The exceptions in Malayalam are त,न,ल,म while those in Hindi are ज,ह. Both the languages have adopted a good number of phonemes from Sanskrit like the aspirated stops, spirants, etc.

Distribution: Vowels in both languages have short and long contrasts, and can, with the exception of final long vowel, occur at the initial, medial and final positions. The final long vowel may occur only in vocative case.

\begin{itemize}
\item e.g., innu 'today' (initial short)
\item aaNi 'nail' (initial long)
\item kuTTi 'child' (medial short)
\item keeLvi 'hearing' (medial long & final short)
\item raamaa 'oh Ram' (final long)
\end{itemize}

Consonants except the aspirated stops can occur at any place. And velar nasal occupy only at the word initial position. Heavy gemination and clustering is possible in Malayalam, and the same can be seen in Hindi too because of the loan words. Gemination, lengthening of consonants, is permitted except in word initial position. Geminated consonants never occupy the word initial position. In modern Malayalam consonant cluster is common and most of the consonants make the clustering. Consonant clusters are generally seen only in the words borrowed from Sanskrit or other languages.

Malayalam and Hindi are syllabic languages, which is one of the basic pan-Indian characteristics. Vowels constitute the nucleus of syllable. So, there will be as many number of syllables in a word as there are vowels. Initial consonant or consonant cluster form the onset of the first syllable and word final consonant forms the code of the final syllable. Mediai consonants with or without length form the onset of the

\footnotesize


6 Joseph, P.M., 1986
following vowel. The first member of the medial consonant cluster goes with the proceeding vowel and the second member goes with the following vowel.⁷

Most of the roots are either mono-syllabic or di-syllabic, i.e., the root has got a single, or double syllable(s) as cited below.

\[\text{e.g., } \text{vaa-} (\text{Mal.}) \quad \text{aa-} (\text{Hin.}) \quad \text{to come} \]
\[\text{koTu-} (\text{Mal.}) \quad \text{deekh-} (\text{Hin.}) \quad \text{to see} \]

In both the languages, certain combinations of phonemes and syllables are permitted and certain are not.

\[\text{e.g., clusters like } *\text{Tk}, *\text{ns}, *\text{vcb}, *\text{sb} \text{ etc.} \]

Certain features can be noticed in both the languages such as a particular consonant as word ending, etc. For example "Raman" in Malayalam is "Ram" in Hindi, or the word final consonant is written along with a vowel in Hindi.

\[\text{e.g., lekshmaNan (Mal.) and lekshman (Hindi) } \text{Lakshman} \]

The primary methodological task of a lexicographer is identification of words. If the words are monomorphemic and meaningful, the lexicographer doesn't have difficulty with such items. But the lexicographer has to see whether the definition of meaningfulness should be restricted to the referential alone or should also include the grammatical, and in the latter case, grammatical morphemes should also be included in the dictionary. Poly-morphemic and poly-lexic units are the two items, which a lexicographer finds difficult, such as identification of the base component which holds the key of its meaning, etc.

Both the languages have considerable dialect variations with sound or pronunciation

⁷ For more detail see Mohanan, 1984 & Ezhuttachan, 1984
differences and these differences may be shown in the lexicon. Both standard and non-standard forms can appear in the lexicon. Stress, intonation, etc. are other related issues for a lexicon.

1. Standard and literary form (Mal.)
   e.g., pustakam 'book'

2. Dialectal form
   e.g., for a dialectal language speaker,
   puttakam for pustakam 'book'

3. Speech variations

   Rarely a spoken form may show a different representation with respect to the written mode.

   e.g., these Malayalam words are pronounced as
   bala (written form) /belam/ (spoken form) 'force'
   ravi /revi/ 'Ravi' (different vowels, same pronunciation)
   naaya /naaya/ 'dog' (same letters with different pronunciation)

Phonological issues of Hindi:

1. Standard
   e.g., pustak 'book'

2. Dialectal
   e.g., kitta for kitna 'how much'

---

8 Somasekharan Nair, 1982

9 See Kellog, 1897 (78-28 of 1987 edn.)
3. Speech variations
   e.g., puStaka 'book'

Phonologically words are classified into the following types according to:

i. the end phoneme:

   a. Consonant Ending
      e.g., maram 'tree'
            raam 'Ram'

   b. Vowel Ending
      e.g., kiLi 'bird'
            cidiya 'bird'

ii. number of syllables:

   a. Monosyllabic (CV) / (CV represents a syllable)
      e.g., nii 'you'

   b. Disyllabic (CVCV)
      e.g., mala 'mountain'
            tuu 'you'
            nadi 'river'

   c. Trisyllabic (CVCVCV)
      e.g., viRaku 'wood'
            saccaayi 'truth'

   d. Polysyllabic (CVCVCVCV)
      e.g., purushan 'man'

iii. root alternation:

   In few cases the roots of some words get changed or the words have a nominal
form and an oblique form. Hindi makes more oblique forms than Malayalam.

\[\text{e.g., njaan 'I' -> en- (Mal.)}\]
\[\text{mem 'I' -> me- /muche- (Hin.)}\]

iv. affixational morphemes:

Prefixes are hardly nil and suffixes are more in numbers in both the languages.

\[\text{e.g., /a-/, /-om/, /-ppan/, /-yaam/}\]

A word as the smallest free unit is usually identified by pause in spoken form and with a space in written medium. The term *adam* was used to refer this in Indian tradition. Though the terminological definition includes more, the concept is same.

2.2.2. Source of Vocabulary

A language has infinite number of words. They include native and borrowed items. Malayalam and Hindi have more borrowed or loan lexical items than native ones. Sanskrit and Procreate are two major sources, along with Persian, Arabic, English, Greek, Latin, etc. Malayalam also contains words of Portuguese, Chinese, Syrian and Hebrew origin.\(^{10}\) The information regarding origin of word is called etymology and it may also appear in lexicon. A number of words with same etymology can be seen in both Malayalam and Hindi. That is, a lot of Sanskrit, Persian and Arabic words are there in both the languages either in *thadbhava* or *thadsama* forms.\(^{11}\) And this information may be used to easily identify the words.

Languages have two main sources of new vocabulary. New words can be borrowed from other languages or coined from its own resources. Also by metaphorical extension of an existing word or by any one of the word formation processes new

\(^{10}\) Joseph, P.M., 1986.

\(^{11}\) See Joseph ; 1986, Godavarma ; 1976, Zaidi ; 1991
words are formed.\textsuperscript{12}

2.3. LEXICAL PHONOLOGY AND MORPHOLOGY

Lexical phonology (Mohanan, 1984) is the field which discusses the structure of word and the word formation rules. Morphology deals with the structure of word and its generation from the root. The morphemic structure of a language differs from others at various levels. The early stages of generative linguistics did not account for morphology. The grammar was conceived of as a device that maps a set of morphemes directly onto a set of sentences. Except for the derivatives of a sentence, the notion of word played no role in linguistic theory. The lexicon was viewed as an instructed collection of whatever as idiosyncratic and unpredictable. The standard theory did not distinguish between morphology and syntax.

Lexical phonology and morphology (LPM) is a successor to the standard generative phonology. In this it has at its core the notion of a set of underlying representations of morphemes, which are converted to their surface forms by subjecting them to a list of ordered phonological rules. The main points of lexical phonology are in the organization. The lexicon has the main productive role here. And it seeks more interactions of morphology and phonology: the lexicon is seen as consisting of a set of ordered levels, each forming the domain for certain phonological and morphological process.

Chomsky (1970) proposed that certain regular relationship between words could be expressed in terms of lexical rules, and these rules were different in nature from the syntactic rules which determine sentence structure. This was the reverse to his earlier approach,\textsuperscript{13} where the output of the lexicon was a set of morphemes. This was against his theory of 1970, where the lexicon was a set of words. He argued that derivational morphology should be relocated in the lexicon, creating a new morphological

\textsuperscript{12} See section 2.5

\textsuperscript{13} Noam Chomsky, 1965
component in the lexicon. This initial move towards lexicalisation has been gradually extended, with the formalities of the strong lexicalist hypothesis, which holds that no syntactic rule may refer to word internal structure, and therefore that inflectional affixation and compounding should also operate lexically. In lexical phonology the lexicon is the domain of all morphological process. Another significant work in this area was that of Halle (1973), which added a new module to the lexicon called the word formation component. This attributed a richer structure to the lexicon, and the role of lexicon became more important. Logical form (LF) and phonetic form (PF) are proposed by lexical grammars of later age.

2.4. CLASSIFICATION OF MORPHEMES AND WORDS

Morphologically, words can be grouped into different patterns and types. These types of classification can be made by considering the way they make inflectional and derivational forms. Both the languages have allomorphic variations and hence the possibility of forming lexical units from different lexical forms with different suffixes. Morphemes are of different types or patterns depending upon its structure or form and its function in the word formation.

There are two types of morphemes. Free morphemes and bound morphemes. Free morphemes are always basic of the language and bound morphemes are those which are used to make lexical categories through lexicalisation process. Free morphemes called root, can be used independently and suffix morphemes or bound morphemes get concatenated with the root to generate new lexical items.

2.4.1 Root / Stem

Root of a word means the form left after deleting all bound morphemes attached to it. And the term stem means the remaining form after deleting one morpheme. *kuTTi* is the root and *kuTTikal* is the stem, as in the following.
e.g., kuTTi 'child'
kuTTi-kaL-uTe 'of children'

Compared to Hindi, affixational morphemes are found more in Malayalam. Some of the frequently used affixes are /a-/ /aaN-/, /-kaL/, /-maal/, /-al/, /-il/, /-ttam/, /-nRe/, etc. Prefixes (the first two above) are comparatively less and most of them are seen in the words of Sanskrit origin. Final and non-final suffixes cannot be differentiated since the language is agglutinative where an additional morpheme can change a final to a non-final morpheme.

e.g., 1. /aNam/ in paT-anam 'study'(n)
2. /tti/ and /nu/ in mara-tti-nu 'tree (dat.)'
   /tt/ (a link morph)
3. a-(negative marker) as in a-himsa 'non-violence'
   /a:N-/ (masculine) in aaN-kuTTi 'male child'
4. /apa- in apa-maanam 'insult'
5. /ttam/ in kuTTi-ttam 'childhood'

Hindi usually has mono-, di-, or tri-syllabic roots. Root remains after all affixes are removed. A stem is something left out after removing one or two affix/es.

e.g., becca 'child' (root) baccpan 'childhood'

2.4.2 Suffix/Prefix/Infix

Suffixes are attached after the root. It could be in final or non-final position.

1. Final
   e.g., paT-aaI 'study'
   buTaapan 'oldhood'
2. Non-final

\[ \text{ledki-yaan 'girl'} \]

**Prefix** is added prior to the root and only a few are available in both the languages.

\[ \text{a-(negative marker), } \rightarrow \text{ a-himsa 'non-violence'} \]

**Infixed** are added in between two morphemes and do not refer any meaning. They number less than half a dozen each in both the languages (-y-, -tt-).

\[ \text{e.g., } \text{kuTTi-y-uTe 'of child'} \text{ (link morpheme or euphonic morpheme)} \]

2.4.3. **Word structure**

Position of morphemes in a lexical form:

Different morphemes take different positions in word formation. But the position of functional morphemes is fixed. The basic structure of Malayalam words can be represented as below.

\[
\pm \text{Prefix} + \text{Root} + \text{Suff1} + \text{Suff2} + \text{Suff3}
\]

\[ \text{e.g., } \text{kuTTi(root)-tta(vn morp)-tti(link morph)-nRe(case-gen.) 'of childhood'} \]

Position of morphemes in Hindi is also same as in Malayalam.

\[
\pm \text{Prefix} + \text{Root} + \text{Suff1} + \text{Suff2} + \text{Suff3}
\]

\[ \text{e.g., } \text{becc-pan-ki 'of childhood'} \]
\[ \text{karvaana 'cause to do'} \]
2.4.4. Word Classification Based on Morphology

Morphologically, words in Malayalam are classified into different groups considering the way they make inflectional forms. After considering the grammatical classifications, the following classification is attempted.

2.4.4.1. Morphology of Noun

The sequence of suffixal morphemes in a nominal formation is gender, number and case.

a). Nouns are morphologically defined as those units which take case markers. The noun group includes pronoun, proper noun, verbal noun, etc.

Noun declensions are according to gender, number, and case. There are three genders viz. masculine, feminine and neuter in Malayalam, whereas Hindi has only two genders, incorporating inanimate (neutral in Malayalam) into either masculine or feminine.

\[\begin{align*}
\text{e.g.,} & \quad \text{avan 'he',} & \quad \text{ledka 'boy' (masc)} \\
& \quad \text{mohan 'Mohan',} & \quad \text{bussa 'bus' (fern)} \\
& \quad \text{tavaLa 'frog',} & \\
& \quad \text{maranam 'death'} & \\
\end{align*}\]

i. Number is related to countable/un-countable. Singular and plural are the two classes in both the languages.

Singular in both the languages are with zero morpheme, ie. the basic form itself is singular. Plural markers in Malayalam are /-kal/, /-arl/ and they have got some phonologically conditioned allophones like /-maar/, /-ngal/, /-yaar/, etc.
Plural morphemes in Hindi are /-yaam/, /-eel/, /-oom/(oblique), etc. The final vowels get changed at the plural formation as seen below.

-aa > -ee ; ledkaa > ledkee 'boys'
-aam > -oom ; ladki-yaam > ledkiyoom 'of girls'
-i > -yaam ; 'ledki > ledkiyaam 'girls'
-ee > -oom ; ledkee > ledk-oom-ke 'of boys'
-aa + -yeem ; sabha > sabhaayeem

ii. Gender differentiates with male, female and non-gender. In Malayalam all the three types are present, while in Hindi only two are (masculine & feminine). Since neuter is absent in Hindi all the nouns are grouped either as masculine or feminine, often arbitrarily.

Masculine markers in Malayalam are /-an/ and feminine are /-i/, /-tti/, /-all

- e.g., av-an 'he' wah (H)
  av-aL 'she'
  veelakkaar-an 'male servant' naukar
  veelakkaar-i 'female servant' naukaarani

Non-gender: Non-human beings and plurals don't have any gender differences and are always treated as non-gender.

- e.g., kuTTi-kal 'boys', (pl)
  veelakkaar-i 'servant'(fem)
  ava 'they'
Gender in Hindi is arbitrary in case of inanimate items. The gender of a few nouns can be known by certain rules. Apart from the arbitrary words, most of the forms with the endings -aa, -am, -tva, -ya, -sha, -u, -uu, -pan, -na, etc. are masculine, and those with -na, -ta, -i,-ii, -hat, -vat, etc. are feminine. But usually the dictionary provides the masculine and feminine identification.

\[\text{e.g., kalam (mas) 'pen'}\]
\[\text{bijli (fem) 'current'}\]

**Feminine** formation in Hindi

-aa > -ii ; ledkaa > ledkii 'girl'
-a > -in ; maalik > maalikin 'lord'
-aa > -iyaa ; kuttaa > kuttya 'bitch'

In a few cases in Hindi the feminine to masculine formation is possible as in the following.

-an > -anooi : behan 'sister' > behanooi 'sister husband'

iii. **Case** relates nouns in a sentence. There are eight cases formed, mostly by adding case marker or post position in both the languages. The following are the case markers.

<table>
<thead>
<tr>
<th>case</th>
<th>Hindi</th>
<th>Malayalam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominative (karta)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Accusative (karma)</td>
<td>-koo</td>
<td>-ee</td>
</tr>
<tr>
<td>Dative (sampradana)</td>
<td>-koo</td>
<td>-u</td>
</tr>
<tr>
<td>Ablative (apaadaana)</td>
<td>-see</td>
<td>-il ninnu</td>
</tr>
<tr>
<td>Agentive (karanu)</td>
<td>-ne</td>
<td>-aal</td>
</tr>
<tr>
<td>Locative (adhiarkanu)</td>
<td>-mee, par, taku</td>
<td>-il,-meel,varee</td>
</tr>
<tr>
<td>genitive (sambandh)</td>
<td>-kaa,-kee, -ki</td>
<td>-nre, -uTe, -uLLa</td>
</tr>
</tbody>
</table>
Nouns and pronouns in Malayalam are grouped into following patterns by considering case form or plural formations:

1. aaraNTu 'somebody' -u > -e aaraNTe (Dative case)
2. aaTu 'goat' -u > -ine aaTine
3. aaru 'who' -u > -e aare
4. neetaavu 'leader' -u > -ine neetaavine
5. accan 'father' -n + -e accane
6. amma 'mother' -a + -ye ammaye
7. appu 'Appu' -a + -vine appuvine
8. ayyar 'Aiyar' -r + -e ayyare
9. cennaaya 'wolf' -ya + -ye cennaayaye
10. ceecci 'sister' -i + -ye ceecciye
11. kaaTu 'forest' -tu > -tine kaaTine
12. kaNaNu 'eye' -u > -i kaNNine
13. kaTal 'sea' -l + -ine kaTaline
14. kutira 'horse' -ra + -ye kutiraye
15. makaL 'daughter' -L + -e makaLe
16. maram 'tree' -m > -ttee maratte
17. paSu 'cow' -su + -vine paSuvine
18. peNNu 'girl' -Nu + -nna/ peNNunnal (pl.) Nu > Nine peNNine (dat.)
19. avaL 'she' -L + -e avaLe
20. avar 'they' -r + -e avare
21. adheeham 'he (hon.)' -m > -ttine adeehattine
22. njaan 'I' enne
23. naam 'we (incl.)' -- namme
24. nii 'you' -- ninne
25. taan 'you (non hon.)' -- tanne
Nominal sub groups

a. Pronoun: There are a few distinctive forms available in Malayalam like the first person plural, which can be of inclusive and exclusive types. Unlike in Hindi, there is no pronominal marker to specify person-number-gender agreement in the finite verb.\textsuperscript{14}

\begin{itemize}
\item e.g., avar 'they', avan 'he', ival 'she(prox.)'
\item njangngal 'we' (excl)
\item nammal 'we' (incl)
\end{itemize}

b. Numeral: Of the cardinal and ordinal types of numerals, ordinal is post positional in Malayalam, whereas it is pre-positional in Hindi

\begin{itemize}
\item e.g., onnu 'one', raNTu 'two', pant-raNTu 'twelve'
\item eeka 'one', ikk-iisa 'twenty one'
\end{itemize}

c. Simple and Compound: Unlike the above functional classifications simple and compound are the structural classification of nouns. These could be any type of noun like personal, pronoun, abstract, etc. Compound forms have more than one free unit. There are at least three types compounds with respect to its constituents.

\begin{itemize}
\item e.g., maram 'tree' (noun-simple)
\item raama-naamam 'name of Raam' (Noun + Noun)
\item van-paTa 'big Army' (Adjective + Noun)
\item paRakkum-kiLi 'flying bird' ( Verb + Noun)
\end{itemize}

\textsuperscript{14} See section 3.4 for GNP ending.
2.4.4.2. Morphology of Verb

A verbal stem in Malayalam may contain one or more root/s. A single morpheme stem will necessarily be the root of the verb. The inflectional suffixes are added to the root. Transitive marker, causative marker, tense markers, participial markers, etc. are the sequence in a construction. Auxiliary verbs are suffixed to the main verb to express aspects and moods.

Finite verb forms in Malayalam show only the tense markers and nothing to indicate gender-number agreements with respect to the subject or object, as in Hindi. That is, for any person or number or gender the verb form is same.

Verb declensions are tense, aspect, mood. There are three main tenses in Malayalam, and each has further divisions. Verbal forms for aspects and moods are also inflectional variations along with participles. Verbal noun, adjective derived from verb, etc. can be included in the derived forms.

The important classifications of verbs at the morphological level are:

A. Auxiliary: A clear auxiliary is absent, but some verbal forms are added to the main verb to refer the aspect and mood.15

For example, continuous aspect is represented by /koN Tirikku-

e.g., avan paaTikoNTirikkunnu 'he is singing'.

B. Simple: This is a unilexical form, which has only one morpheme. Simple verbs are more in number compared to compound verbs. The base form of a verb is represented with an infinitive marker /-ka/.

 e.g., pookuka 'to go'

Morphologically verbs are defined as those units which take tense markers. Tense forms, transitives and causatives, etc are some morphological sub-categories derived from the verb root.

**a. Tense Formation**

Basically there are three tense forms. Except the past, others have only one marker. The future marker is /-um/, present marker is /-unnu/ and the basic past markers are /-u/ and /-i/.

The following are the three main tense markers.

1. **Past:** [markers: /-i, -ccu, -nnu, -u, -nnu/]
   - e.g., 1. ravi kuTTiye aTiccu 'Ravi beat the child'
     
     Ravi boy-Acc. beat-past.

2. vanTi ninnu 'bus stopped'

Verbs in Malayalam are morphologically grouped into sixteen classes depending upon the past tense formation\(^{16}\), as follows:

<table>
<thead>
<tr>
<th>Base form</th>
<th>Past form</th>
<th>Morphemes identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. aru-ka</td>
<td>arutu</td>
<td>u- tu &gt; tu</td>
</tr>
<tr>
<td>2. uNNu-ka</td>
<td>uNTu</td>
<td>n-tu &gt; ntu</td>
</tr>
<tr>
<td>3. aRu-ka</td>
<td>aRRu</td>
<td>t-Ru &gt; RRu</td>
</tr>
<tr>
<td>4. aTayu-ka</td>
<td>aTanjnj</td>
<td>y-ntu &gt; nnu</td>
</tr>
<tr>
<td>5. araLu-ka</td>
<td>araNTu</td>
<td>l-n-ntu &gt; ntu</td>
</tr>
<tr>
<td>6. akaru-ka</td>
<td>akannu</td>
<td>r-nru &gt; nnu</td>
</tr>
<tr>
<td>7. akalu-ka</td>
<td>akannu</td>
<td>l-n-nt &gt; nnu</td>
</tr>
</tbody>
</table>

---

\(^{16}\) Suranad Kunjan Pilla, 1972 & Valantine, 1986
ii. Present marker in Malayalam is [-unnu].

e.g., 1. rema ammakku paisa koTukkunnu
Rema mother-Dat. money give-Past
'Rema is giving money to her mother'

2. rcvikku panikkunnu 'Ravi is feeling fever'

iii. Future tense also has only one marker [-um]

e.g., avan var-um 'he will come'

Hindi has three main tense forms, and fifteen sub tense forms based on aspects and moods. It is a must for all tense forms to have a number-gender agreement. This agreement may be with subject or object.

Past tense in Hindi has seven sub-classes and some of them are simple, perfect, imperfect, continuous, etc. Different markers are used to differentiate various agreements like first person-singular and plural (masculine and feminine), second person-singular and plural (masc. & fem.), and third person singular and plural (masc. and fem.). In the following way the final vowels get changed during the past tense

---

1. simple (VR[change of vowle]+/-
   -a + -a ; paTha 'study' > paThaa (mas.sg)
   -aa + -ee ; aayaa 'come' > aayee (mas.pl)
   -aa + -ii ; khaayaa 'eat' > khaayii (fem.sg)
   -a + -iim ; dauTha 'run' > dauThiim (fem.pl)

2. Immediate: simple + hee/heem/huum/hoo
   paThaa 'to know' + hee/heen/huum/hoo

3. Perfect: simple + dhaa/dhee/dhii/dhiim
   aayaa 'come' + dhaa/dhee/dhii/dhiim

Present forms: VR + (taa/tec/tii) + (hee/heem/huum/hoo)
   roo 'cry' + (taa/tec/tii) + (hee/heem/huum/hoo)

Future forms: VR + uum/eem + gaa/gee/gii/giim
   cal 'move' + uum/eem + gaa/gee/gii/giim

Other verbal inflectional forms like active-passive, transitives, causative1 and
causative2, participial forms, and other modal forms, are as follows.

b. Transitive Formation:

I.i. verbs ending in a stop (L/T and N/R): (The meanings given in the following cases
are of intransitives.)

ii. velar (k) to geminated (kk)
   e.g., Laku > iLakku 'to move'
aaTu > aaTTu 'to oscillate'

iii. veminated nasal (nn) to geminated (cc)
   e.g., aNannu > aNaccu 'to keep by side'

II. geminated nasal (nn) to geminated -kk.
   e.g., tuunnu > tuukku 'to hung'
   curunnu > curukku 'to shrink'

III. geminated nasal (nn) to geminated -cc
   e.g., iTunnu > iTiccu 'to smash'

IV. tril R added with -kk
   e.g., tiiRu > tiiRkku 'to finish'
   ceeru > ceeRkku 'to mix'

V. consonant cluster -lkk into -rtt
   e.g., nilkku > nirttu 'to stand'

VI. velar to dental (-kk to -tt)
   e.g., irikku > iruttu 'to sit'

c. Causative Formation: Generally causative 1 is formed by adding allomorphs like [-ippi-, -ccu-, -kku-] and causative 2 is generated with [-ippi-].

   e.g., deeviye veedanippiccu 'Devi is caused to felt pain'

Causative 1

1. iTu > iTiikkku 'to drop'
2. taru > tariikkku 'to give'
3. paaRu > paaRikkku 'to fly'
Causative2

1. eelkku > eellppikku 'to take over'
2. ormikku > ormippikku 'to remember'
3. eTukku > eTuppikku 'to take'
4. aTukku > aTuppikku 'to come close'
5. paRayu > paRayippikku 'to say'
6. cooru > coorttu 'to loose'

d. **Passive formation** is by adding peTu to the verb in Malayalam and (ko+....+ past form of the verb) in Hindi.

   raaman kolla peTTu 'Ram is killed'
   raam maara geyaa 'Ram is killed'

e. Moods

Following are the modal forms generated from the base of verb by adding the respective markers.

i. **Denotative mood**

   (VR. + -aan)

   e.g, pooku > pookaan 'for/to going'

   VR. + -na_caahiye > aana_chahiye

ii. **Imperative**

   Zero or lengthening

   e.g, varu > varu/varuu 'please come'
VR. + -viin
  e.g., varuviiin (polite)
VR + -iye
  e.g., aayiye

iii. Purposive / Request / Benefaction

VR. + -aTTe
  e.g., varaTTee 'let (him) come'
VR. + -ne_do
  e.g., aane_do

iv. Potential

VR. + -aam
  e.g., pookaam 'can go'
VR + -sakta
  e.g., cal_sakta 'can go'

v. Optative / infinitive

VR + -ka
  e.g., pookuka
VR + -oo
  e.g., caloo

vi. Expressive

VR + -uu
  e.g., pookuu
VR + -oo
  e.g., caloo
vii. Conditional

VR + -aal

  e.g., vannaal 'if come'

VR+ -eem_to

  e.g., jaayeem_to 'if go'

viii. Optative/purposive

VR + -aal/-aan

  e.g., 1. sasi Trainil madraasil pooyaal/ 'if Sasi go to Madras by train ...'
        2. sasi Trainil madraasil pookaan/ 'Sasi to want to go Madras by train ...'

f. Verbal participle

For verbal participle there is no particular marker which gives emphasis. But iTTu can
be added to give emphasis.

VR + -0, -u ; vannū 'come ...'
VR + -kar ; jaakar 'having gone'
VR + -iTTu ; vanniTTu 'having come'

e.g., ceecci vanniTTu pooyi.
  'Sister has gone having came'
  diidi aakar cali gayi (Hi)

g. Relative participle marker is always -a, which is also the adjectival marker, and
it is added to the finite form of the verb by simply replacing the final vowel 'u'.
-a; vannu > vanna 'who came'
jo_aaya 'who came'

\[ \text{e.g., mani ceyta pani 'the work Mony did'} \]

**h. Negation**

Morphological marker of negation is \textit{illa} or \textit{alla} which is added to the verb just after the tense marker. Negation is added to the finite form of the verb.

\[ \text{e.g., 1. raamanu paNam uNTu (pNamuNTu)} \]
'Ram is having money'
\[ \text{2. raamanu paNam illa (pNamilla) [syntactic negation]} \]
'Ram is not having money'
\[ \text{3. raaman paNam tarunnilla [morphological negation]} \]
'Ram is not giving money'

Negation can be represented in the following way.

1. +alla (nominal negation), +illa (nominal\verb negation) :
\[ \text{pooyi +illa > pooyilla 'didn't go'} \]
\[ \text{vara + alla > varayalla 'not a line'} \]

2. -aatta, -aa, -a (relative participle) ;
poorkaatta 'not gone'

3. -aate (verbal participle) ;
poorkaate 'having not gone'

\textbf{na} and \textbf{nahin} are the equivalents for Hindi

\[ \text{e.g., na jaana / nahiin jaana 'not go'} \]
C. Compound verbs are formed either with two verbs, or with a noun and verb.

i. Verb + Verb
   e.g., vannu-kaNTu 'come-see' (came and saw)
         kalNanjnju-pooyi 'loss went' (lost)

ii. Verb + Auxiliary
    e.g., vannirunnu 'had come'

iii. Noun + Verb
    e.g., aahaaram-aayi 'become food'

2.4.4.3. Lexical units other than noun and verb

Adjectives, adverbs, coordinators, post-positions, etc. come under this group. Except adjective, other forms hardly undergo any change of structure or category. Adjectives may change their category and become a noun or a verb.

1). Adjective: The particular marker for adjective is /-a/; same is relative participial marker. Relative participle behaves always as an adjective. Simple, participle and ordinal are important adjectival categories.

   1. Simple
      e.g., nalla 'good'

   2. Participial
      e.g., oTunna 'running'

   3. Ordinal
      e.g., oRRa 'single'
2). **Adverb**: There is no discrete marker for adverb.

   e.g., vaLare 'too much', zaada
   patukke 'slowly', saavdaan

3). **Post positions and other indeclinables**: These include bound forms as well as free forms, and they never undergo any inflectional changes. Post positions always come after a noun and refer cases. Post position either behaves as a case marker or when added to the case affix indicates special or added meaning. Indeclinables have functions of either coordination or exclamation, and the like.

   e.g., katti_koNTu 'with the knife' (instrumental)
   baank-il_ninnu 'from the bank' (ablative)
   avan-um (co-ordinative) 'he too'

   ke-saamne 'in front of'
   isliye 'so that'
   aur 'and', magar 'or'

2.5. **LEXICAL/WORD FORMATION**

Word formation theories deal with the word generation techniques. Much work has been done after the introduction of generative phonology, a theory developed parallel to generative grammar.\(^{18}\) Morphological structure of both languages are by and large suffixation.\(^{19}\)


\(^{19}\) Morphology of Hindi words with respect to computational approach has been discussed by Varshney et al., 1991. and Sinha, et al., 1993.
2.5.1. Word Formation in Malayalam.

Normally Dravidian morphology is characterized by the general pattern of

\[ \text{Stem (root + augments) + medial(s) + ending(s)} \]

Suffix structures of Malayalam are statable with regard to certain grammatical categories. Causative, tense, negative, participle, derivational suffixes, etc. can be stated with respect to their positions. New morphemes are added at the end of the word. With minor exceptions, one can rightly state that suffixation is the only type of affixation which creates derived or inflected forms in Malayalam and other Dravidian languages.\(^{20}\)

\[ \text{Stem + formative suffix(s)/free form.} \]

\[ \text{or} \]

\[ \text{prefix + Root + suffix(s)} \]

The formative suffixes include bound and free morphemes.

The root morpheme stands always as the first morpheme of a morphemic complex, and the root does not generally change, and it takes no part in morphology in Dravidian languages. But in some derived forms prefixing is permitted. Malayalam uses concatenation of morphemes one after the other for the word formation. The root, in a small number of cases, may be slightly affected in morphological process.

The structural positions in which suffixes occur can be represented as follows.

1. For verb;

\[ \text{Stem + Tr + Caus1 + Caus2 + LM + Ten/Aspect + VP/RP/VN/Cond} \]

\(^{20}\) Kamil Zvelabil, 1990.
2. For noun;
Stem + ordinal/Numb + Gend + Per + Case

e.g., raman + u (dative) > ramanu 'to Ram'

pen + KuTTi + kaL + kku > penkuttikaLkku
fem. child plr. accu. girls (Acc).

paRa + y + iec + u + kaLann + u > paRayiccukaLannu
tell LM. caus. past accru. past. "caused to tell"

2.5.2. Word Formation in Hindi

Though word formation in Hindi is mainly through suffixation, it takes place also through prefixation. The basic structure of a word in Hindi is almost same as in Malayalam.

1. For verb;
   Stem + Tr + Caus1 + Caus2 + LM + Ten/Aspect + VP/RP/VN/Cond

2. For noun;
   Stem + ordinal/Numb + Gend + Per + Case

The three morphological processes by which new lexical items are being produced from basic lexemes in the language are inflection, derivation and compounding.\textsuperscript{21} Inflection does not change part of speech, whereas derivation does change the part of speech and compounding may also change the part of speech.

Inflection is often required in particular syntactic context whereas derivation is never required by the syntax. For example the words cited below have different category and

\textsuperscript{21} Mohanan, 1986, Laurie Bauer, 1983:30.
function in the sentences:

kuTTi pooyi ‘child went’
kuTTikal pooyi ‘children went’ (infl., pl)
kuTTittam pooyi ‘childhood went / vanished’ (deriv., N)
aankuTTi pooyi ‘male child went’ (compd., N)

Inflectional morphology, because of its strong interaction with the syntax and semantics of sentences, seems particularly crucial for computational morphology whereas derivation is not. A great deal of work done on computational morphology centers largely on inflectional morphology simply for the reason that inflectional morphology is mostly unavoidable for many constructions.

i. **Inflection** is the process which generates grammatical forms of single lexeme which is semantically regular and has the same grammatical category. Nominal inflections are based on gender, number and case. Three genders - masculine, feminine and neuter are there in Malayalam and only the first two in Hindi. Tense, aspect, mood, voice, transitive, causative, etc. are the verbal inflections.22

ii. **Derivation** produces new lexical items from the root which will have different grammatical category. It makes no agreement and is semantically irregular. Derivation produces complex forms. A noun from a verb, or verb from an adjective, etc. are common derivational formations.

    e.g.,  
    kuTTi (n) > kuTTittam (n) ‘becpan’
    valutu (adj.) > valutaakuka (v) ‘bera hoona’

a. **Verbal Noun formation**: Various morphemes are used as allomorphs for the formation of verbal nouns.

---

22 See section 2.4.4 for more details on morphology of verbs and nouns.
-nu > -al ; toonnu 'to feel' > toonnal
-u > -TTam ; ooTu 'to run' > ooTTam
-yu > -ccil ; mushiyu 'to get dirt' > mushiccil
-kku > -ti ; porukku 'to reside' > poruti
-yu > 0 ; paNiyu 'to do' > paNi
-ngu > -kkam ; uRangu 'to sleep' > uRakkam
-ikku> -ana ; aloocikkku 'to remember' > aloocana
-ikku> -anam ; jenikkku 'to born' > jenanam
-kku > -am ; tuukku 'to hung' > tuukkam
-ru > -rca ; neeru 'to offer' > neerca
-yyu > -tu ; neyyu 'to weave' > neyttu
-Nu > -ca ; kaaNu 'to view' > kaazca

iii. **Reduplication** is another process in which same lexical item is used twice to refer a meaning of 'etc.' The second unit is formed in the same rhyme with a minor alteration.\(^{23}\) Echo word formation is another issue. Almost all categories make reduplicated forms. Due to morphophonemic changes some phonemes may get altered or deleted.\(^{24}\)

e.g., aTicca-Ticcu (v-v) 'beating-beating' (from aTiccu 'beat')
koccu-koccu (adj-adj) 'small-small' (from koccu 'small')
paisa-kisa (n-n) 'Paise and like that' (from paisa 'Paise')
melle-melle (adv-adv) 'slowly-slowly'

khaatee-khaatee 'eating-eating' (from khaataa 'eating')
idhar-udhar 'here and there' (from idhar 'here')

\(^{23}\) See Anvita Abbi, 1992 for more on reduplication in Indian languages.

\(^{24}\) See more detail in sandhi.
iv. **Compounding** produces lexical items from two or more roots. Compounding produces compounds, which are of four types according to the semantic criteria. They are endocentric, exocentric, appositional and copulative.\(^{25}\)

Compound forms are made with a combination of other words and the constituents are words in their own right. Compounds of nouns, adjectives, verb and their derivatives are possible. Compound words are formed mostly from two or more noun stems, or from a noun, a verb and a verb stem, and for a relatively small number of cases from other combinations as well.

2.6. MORPHO-SYNTAX OF WORD AND PHRASE

A word and a compound are syntactically the same. But morphologically the latter one is more complex than the former one.

Indian languages have comparatively free word order in a sentence. But morphologically or in phrasal level, word order has a sequence such as root followed by suffix/es, or main verb followed by auxiliary verb, etc. or noun followed by post position, etc. A verb can be nominalised or a noun can be turned into adjective, etc. by adding some forms such as aaNu, uNTu, etc. in Malayalam.

avan pooyilla 'he go not (past)'

*avan illa pooyi
he  not  go(past).

*poole siita atu paranjnja
like Sita  that what said  'the way siita told'

\(^{25}\) Laurie Bauer,1983-30.
2.7. COMPOUNDS AND SAMAS

Keralapanini (1897) defines compounds as the joining of words without case markers. Large compounds are rare in Malayalam and two to three words as constituents are common. Following are the types of compound words under categorical classification as opposed to the functional classification which is seen earlier.

1. Adjectival compound
2. Adverbial compound
3. Nominal compound
4. Verbal compound

2.7.1 Head and Modifier

Categories as head and modifier combine to yield a resultant category called phrase. A head is present in a construction where it has got more than one free lexical units. So head is possible in phrases and in compounds. The following forms have head at their right hand side.

- e.g., manusyakurangu 'man monkey'
- puukuuTa 'flower basket'
- vaziampalam 'street temple'

These forms are endocentric. Most of the compounds are endocentric. There is a small number of compounds that are said to be exocentric, having no head at all.

- e.g., accanammamaar 'father and mother'

According to the modern grammatical theories, a constituent $C_i$ is said to be the head of a constituent $C_j$ if it satisfies two conditions. First, it must bear the same syntactic category features of $C_j$. Second, its level must be 'one lower' in the $X'$ hierarchy than that of $C_j$. 
In a structure like the one given below the main verb is the head or the noun phrase.

\[
\begin{align*}
\text{2a.} & \quad \text{VP} \\
\text{2b.} & \quad \text{NP} \\
/ & \quad / \\
V & \quad \text{Aux.} \\
\backslash & \quad \backslash \\
vannu pooyi & \quad \text{vazhi ampalam} \\
came & \quad \text{went} \\
'\text{happened to come}' & \quad '\text{street temple}'
\end{align*}
\]

Head is the crucial item in characterizing the semantics of compounds. This emerges quite clearly when we consider n[N N]n compounds, for example. The compound \textit{sisupaadam} 'child’s foot' designates a string that is related to a \textit{sisu} 'child'. The compound 2b (see the fig. above) is having the head temple which is located in a street. But, in \textit{naayarambalam} 'Temple of Nair' the temple is 'related to the Nair community'. The compounds derived, except from verb, displays a specific and grammatically characterizable range of meaning interpretations. The semantic relation existing between the head and the constituent can vary considerably, and is discussed in the section 3.4

2.8. SANDHI

Since Malayalam can be written continuously within a phrase with out a pause or any space, it may undergo sandhi modification. Sandhi is defined as the intimate conjugation of letters. When letters join according to grammatical rules new changes occur. The sandhi rules in Malayalam are not regularly applicable to all forms, even though they may have similar phonological characteristics. Sandhi is divided into three kinds according to their places in the word.\textsuperscript{26}

\textsuperscript{26} Varma A R Rajaraja, 1897 & Nair A Sukumaran. 1984.
1. padamadhya sandhi (sandhi in the middle of a word)
2. padanta sandhi (sandhi at the end of a word)
3. ubhaya sandhi (containing both i.e., in the middle and end).

   e.g., maram (tree) + il (loc.) > marattil (in the tree)
   pon (gold) + pu (flower) > polppu (golden flower)
   maNi (pearl) + aRa (room) > maNiyaRa (grand room).

Another kind of division of sandhi is with reference to the vowels and consonants involved in it.

1. svara sandhi (vowel + vowel)
   e.g., caaya (tea) + illa (no) > caayayilla ('there is no tea', -y appears in the middle)

2. svara vyanjana sandhi (vowel + consonant)
   e.g., taamara + kuLam > taamarakkuLam ('the lotus pond', k doubles)

3. vyanjana svara sandhi (consonant + vowel)
   e.g., kaN + illa > kaNNilla ('there is no eye', n doubles)

4. vyanjana sandhi (consonant + consonant)
   e.g., nel + maNi > nenmaNi ('grain of paddy', l > n)

Third kind of division is according to the various changes that occur to the phonemes involved.

1. loopam (elision)
   e.g., atu + alla > atalla ('it is not', final u is elided)
2. aagamam (appearance of new letter)
   e.g.,  maNi + uNTu > maNiyuNTu ('there is money', y appears)

3. dvittam (gemination)
   e.g.,  paTTi + kuTTi > paTTikkuTTi ('puppy', k duplicated)

4. aadeesam (substitution)
   e.g.,  eN + nuuru > eNNuuru ('eight hundred', n changes to N)

Sandhi rules are required for the lexical system to generate compounds and other complex forms. Sandhi is also needed to parse the compounds and other coordinated forms, as seen in the examples cited above.

2.9. PHONOLOGICAL AND MORPHOLOGICAL INFORMATION IN A COLEX

The process of recognizing words of a text require computer based calculation as well as consulting an related lexicon. The lexicon serves to identify structural system, such as morpheme structure, which reflects the formation of individual items. The lexicon assists in parsing the input forms to give the internal structural representation which enable the system to identify the word.

The information about the roots and the possible morphemes which can be attached to generate the inflectional and derivational forms are required in a system. The information about derivational forms, the structure of words and the position classes of morphemes alongwith the sandhi changes, etc. are made to be available in a dictionary for the proper use.

The types of dictionaries such as phonological dictionary, pronunciation dictionary, spelling dictionary, etc. produce information at the phonological and phonetic levels. Morphological information, like roots/stems and other variants, suffixes/ morphemes, allomorphemes, etc. are provided in a morphological lexicon.
The morphonological components identified to be needed for a lexical design are:

1. The word in the base form with its phonological properties which include syllabic structure and related attributes.

   e.g., 1. avan, "he", (Monosyllabic-noun, const-ending, no root change.)
   2. valayu "to bent", Disyll-verb-root,

2. Morphology of the source words to derive morphological variants.

   e.g., 1. avan- + numb/gend + case
   2. vala- + caus/pass + TAM

3. Root in the target language with related information.

   e.g., Malayalam word kollu (M) > maarna (H) "to kill"

4. Morphological information of a form derived from the root.

In a system, there should be a category marker and a tag to indicate the morphological type (as seen earlier). Since the noun and verb categories generate a number of inflectional and derived forms they require much attention while developing a morphological analyzer. Different patterns are identified\(^{27}\) for noun and verb in this connection. Number of items of other parts of speech is limited, and they hardly make any inflectional or derivative forms. So, for convenience, we can say that there are four lexicons.

\(^{27}\) As cited in section 2.4.4.
I. **Noun dictionary** with the following aspects:

1. Nominal and oblique forms.
2. Morphemic tag to indicate inflectional (number, gender, case) type.
3. Derivation information (adjectivization, etc.).
4. Compound forms and Samasas.

II **Verb dictionary** to be dealt with following features:

1. Active and passive voices,
2. Intransitives/transitives,
3. Causatives 1/2
4. Tense / Aspect / Mood
5. Infinite i.e., Participle (verbal and relative participles)
6. Verbal noun / Gerund
7. Compounds and phrases

III. **Dictionary of modifiers** includes adverbs, adjectives, etc and they don't require much morphological attention.

IV. **Dictionary for other lexical items** includes post-positions, declinables, etc. and these also don't require much morphological attention.

2.10. **MORPHOLOGICAL ANALYZER AND GENERATOR**

Morphological analysis (MA) and morphological generation (MG) are the two modules that support computational lexicon for lexical formation and lexical analysis. MG will generate the required lexical items from a root and its attributes and MA will parse a given construction to get the internal structure by morphological decomposition. MG can produce all possible forms which are morphologically related to a word for a paradigm-dictionary. One can also be able to look-up a word that is not explicitly listed in a general dictionary as the morphological analyser can tell to which word it
is related by parsing it. For example, the word *enre* 'my' in Malayalam which a general dictionary would not list, can be found along with the word *njaan* 'I'. An MA can also be used in developing teaching tools, for studying morphology for language learners, and morphology theory. The method is applied to the languages which have got a productive morphology, like Sanskrit where morphemes are represented by affixes. As most of the Indian languages are either inflectional or agglutinative, the morphological analyzer could be applied for the efficient use of the lexicon.

Typologically, languages of the world are classified into isolating, agglutinating, inflectional and poly-synthetic at the morphological level. Isolating languages have single morpheme words, whereas separate affixes, each with its own meaning, are used to represent a word in agglutinating languages. In an inflectional language it is usually impossible to determine precisely which part of an affix represents which morpheme, and in polysynthetic, greater use of affixes, often involving incorporation of noun and adverbs within the verb, can be seen.

Typological information gives the morphemic structure of words. Cross reference of the position and structure in different languages helps while forming words through morphological analysis. The languages of our choice namely Malayalam is agglutinative while Hindi is an inflectional language. Word formation in the former is through suffixation in succession, and separate affixes are added to the root or stem, whereas in Hindi there is morpheme prefixing and suffixation. Morphological typology attempts to characterize languages according to:

1. the extent to which linguistic concepts are expressed by morphological (ie. word internal) modification.
2. the morphological techniques applied.

---


30 Details can be seen in 2.4.
Many attempts have been made to improve the classical typological divisions throughout the last two centuries. Recent direction in which morphological typology is moving is important in its implications. The attempt made by Nicholas (1986) is based on the pattern in the placement of morphological marking in different construction types within a sample of languages. The notion of head versus dependent in a construction is the key to this approach. He tried to show a group of languages at the consistently head marking or consistently dependent-marking ends of the scale, with a smaller group of languages with split or double marking in the middle.

While developing a lexicon for any language, the technique is to store only the root word with additional information required for the recognition of the root and its derivations. A large number of words can be derived by simple rules such as plurals, gender forms or even derivatives by simply adding -e, -eem, -o, -pan, etc. in Hindi and -kal, -maar, -r, -ttam, etc. in Malayalam. For example, the words bacca 'boy' gives becci (fem.), becpan (a noun) 'boyhood', etc. and kuTTi 'boy' kuTTikal(pl), kuTTittam 'boyhood', etc.

The task of recognizing a root word from a given derivative and retrieving the corresponding information from the lexicon is accomplished by the morphological analysis. While a large number of derivatives follow systematic synthesis rules, there are exceptions which cannot be obtained through morphological analysis.

In agglutinative and analytic languages a word can be a combination of morphemes or a single one. The morphological analyzer must know the basic words of the language in addition to the prefixes and suffixes to avoid incorrect formation or breaking up of the words. Also the analyzer must be guided with the position class of morphemes. Typological knowledge of stem and affixes will help in identifying the position of a particular morpheme which can appear in a particular environment. Most of the inflected forms can be constructed or analyzed in accordance with the paradigm of that language except irregular and borrowed or adopted words.
Different strategies can be adopted for morphological analysis in a bilingual dictionary. A bilingual dictionary by definition makes use of two languages whether the languages are related or not. If the source and the target languages have different morphemic representation as well as position classes, different modules can produce the required form. Use of separate modules for each language is one method. Providing paradigm files with their characteristics, storing all possible root words along with the morphological, categorical and typological information, etc. are other methods. After preparing all possible paradigms for each category, words were mapped to these paradigms. These latter methods use morphological information consisting of all possible suffixes a word can possess, categorical information containing part of speech and attributes of gender, number, person, etc., and typological information that includes position class, etc.

Most of the works so far done are only on the inflection morphology. Derivational morphology as well as compounding have got little attention. Identification of morphemes of derivational and compound words for analysis and generation needs different steps to follow. It could be the same way as above. Typology of morphemes could be taken as an issue in identifying the position at morphemic level, in an automatic system. Implementation of morphological rules is discussed in chapter 4.