2.1 INTRODUCTION

In this chapter, we discussed the techniques and tools used and the methodology that is adopted for the statistical analysis of Tamil Morphology based on the corpora. As a part of this, we would also discuss the quantitative aspects of the Tamil morphological elements too. This chapter is divided into five modules, each of which discusses (i) the statistical techniques and methods in language analysis; (ii) Getting ready with the resources for the study of Tamil morphology for the statistical analysis, which involves cleaning the texts and other related aspects which are used in the analysis of the corpus in addition to converting it into a convenient encoding system for our analysis; (iii) the hardware and the software that are used, the programming techniques and tools, etc.; (iv) the implementation, testing and evaluation. (v) The quantitative approach to Tamil morphology. These are discussed as in the following heads:

(i) Techniques and Methods,
(ii) Resources: The Corpora,
(iii) Programming Techniques,
(iv) Implementation, Testing and Evaluation,
(v) Quantitative approach to Morphology

As mentioned in the introduction of the first chapter, our main objective is to record the facts about the frequencies of affixes and stems as constituents of the word forms. First and foremost, a frequency list of word forms numbering
4,00,000 from a corpus of 3 million words of running texts. These words have been extracted from said corpora, and each word in the list is arranged in two columns, the first column consists of frequencies of words and in the second column, the word forms. For instance, a subset of the supplied Tamil word list or word forms looks like this:

1422 peṇn. + kal]+n, pl + ukku]+n, pl, dat + ē]+n, pl, dat, emp
230 vā]+v. + nt]+v, pst + ā]+v, pst, 3sg,fem
278 pūn. kkal]+n, pl + iliruntu]+n, pl, abl/psp
420 pat]+v. + kki]+v, prs + ān]+v, prs, 3sg,mas
461 maram]+n. + kal]+n, pl.
589 kkal]+n. + kal]+n, pl + ukku]+n, pl, dat
723 ka]+n. + kal]+n, pl + ai]+n, pl, acc
77 pacu]+n. + ai]+n, acc

There are a number of things to clarify about the above list: Each line of the list contains a root word separated from remaining portion its analysis (e.g., maram + kal) by one or more plus symbols. The analysis displays morpheme labels separated by plus signs. In Tamil, for instance, in order to get a total score for any given word (root word) or a morpheme frequencies of all allomorphs shall be obtained and summed for totals. Therefore, it would be necessary to distinguish the allomorphs of the suffix “-kal” (-kal or -kkal). Morpheme labels in the analysis are separated from each other by a space. This means that only grammatical categories that are realized as morphemes are included.

The morpheme labels that correspond to inflectional (and sometimes also derivational) affixes have been indicated within the closing brackets. (e.g.]+pl, ]+pst). This is necessitated for the evaluation in addition to the requirements of overall performance statistics. Evaluation measures are also computed separately for the labels enclosed in brackets. It becomes easier to
make an approximate assessment of how accurately affixes vs. non-affixes are discriminated. When we use this kind of naming to the morphemes then quantification will be easily carried out. The morpheme labels that have not been marked as affixes (not enclosed in brackets) are typically stems. These labels consist of string, usually followed by an closing bracket ( ] ) and a part-of-speech tag, e.g., ‘maram]\n,’ ‘col]\v’. In many cases, especially in Tamil, the same morpheme can function as different word classes; e.g., the Tamil word ‘paṭi’ which can function as noun or as verb when it is does not get inflected. But when it gets inflected then the inflectional affixes or the derivational affixes would tell whether it is belongs to a noun category or a verb category.

For instance, the original word classes of the words "paṭikaḷil" (paṭi\n. + kal]\n, pl + il\n, pl, loc) and “paṭikkiṟāṉ” (paṭi\v. + kki\v, prs + āṉ]\v, prs, 3sg.mas.) are a noun and a verb because of their inflectional suffixes that they inflected to. Thus both noun and verb inflections of “paṭi-kal-il” and "paṭi-kkiṟ-āṉ" are contain the morpheme "paṭi".

The most frequent affixes of an analysis for a given word correspond to the list of labeled morphemes. Affixes of an analysis may consist of a list of unlabelled morphemic segments because of their ambiguous nature all such cases manual analysis will be performed to find out their actual identity. The method presented in this study simple since it labels the segments with general morpheme categories, which are detailed below.

**Morpheme Categories**

The morphemic segments discovered by the programme are labeled by one of the following categories: prefix (P), stem or base (B), suffix (S) and Sandhi or Morphophonemic changes (MC).
Example morpheme analyses

marañkaḻiliruntu = maram]ₐ is a root/stem + ŋ]ₘₖ is a morphophonemic changes or sandhi + kal]ₚl is a plural marker suffix + il]ₜₜ is a locative case marker suffix + iruntu]ₜₜ is an ablative or postposition.

usahaṉavan = uzai]ₐ is a verb root + kki]ₚs is a present tense suffix + avan]ₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜueling
TECHNIQUES AND TOOLS FOR THE ANALYSIS OF TAMIL

MORPHOLOGY

immaraṅkaḷai = i[dem + maram]dem, n, kal]dem, n, pl + ai]dem, n, pl, acc

The morpheme paṭi is ambiguous: it can be a noun ‘step’ and can be a verb ‘read’ or even a ‘measure’. The morpheme paṭi is correctly labeled as a stem in the words. When the same morpheme occurs at the beginning of a word, it is always labeled as a stem, as in paṭikaḷil (paṭi]n. + kal]nv, pl + il]nv, pl, loc)

The above analysis is obtained through the techniques laid down in the 2.1. This procedure takes a list of words or list of words with frequency as input and returns a list of labeled morphemes for each word. Morphological segmentation are obtained using the above techniques.

2.1. (I) TECHNIQUES AND METHODS

As we mentioned earlier this method first takes input file as words with frequency information as base and gets the output as prefix, stem and suffixes. The output is a labeled segmentation of the input words. Labels belong to one of the following categories: prefix, root/stem and suffix morphophonemic variants. The procedure can be divided into four main steps.

STEP 1: Retrieving of Prefixes

The objective of step 1 is to acquire a list of prefixes. All the words in the input word list are segmented based on the prefixes. We consider these a, i and e (demonstrative base) are prefix for the computational purpose. This idea is frequent in research on the segmentation of words into morphemes. One morpheme boundary is identified in this word, which corresponds to the following segmentation:
a+nīti+kal+ukku which means “To injustices”
a+cuttam+kal+il which means “In lack of cleanliness”
\[a\_{dem}. + \text{cuttam}\_{dem}, n. + \text{kal}\_{dem}, n, pt + \text{il}\_{dem}, n, pt, loc\]

a+jīraṇam+tt+āl which means “by indigestion”
\[a\_{dem}. + \text{jīraṇam}\_{dem}, n. + \text{tt}\_{dem}, n, obl + \text{āl}\_{dem}, n, obl, ins\]

a+v+viḷaiyṭṭu+il which means “In that game”
\[a\_{dem}. + \text{viḷaiyṭṭu}\_{dem}, n. + \text{il}\_{dem}, n, loc\]

e+v+virutu+kku which means “To which award”
\[e\_{dem}. + \text{virutu}\_{dem}, n. + \text{kku}\_{dem}, n, dat\]
Loanwords from Sanskrit are not productive and they are not created new forms.

**STEP 2: Retrieving of Stems**

The aim of this step is to retrieve a list of stems, using the prefixes and suffixes. By stripping off from each word in the input word list all the possible combinations of prefixes, suffixes to get the bear stem or a root (simple or complex).

a+p+pali+kal+il which means “In those schools”
\[a\_{dem}. + \text{pali}\_{dem}, n. + \text{kal}\_{dem}, n, pt + \text{il}\_{dem}, n, pt, loc\]

e+p+paṭam+h+kal+il which means “In which pictures?”
\[e\_{dem}. + \text{paṭam}\_{dem}, n. + \text{kal}\_{dem}, n, pt + \text{il}\_{dem}, n, pt, loc\]

i+p+paiyaṇ+kal+il which means “In these boys”
\[i\_{dem}. + \text{paiyaṇ}\_{dem}, n. + \text{kal}\_{dem}, n, pt + \text{il}\_{dem}, n, pt, loc\]

**STEP 3: Retrieving of Suffixes**

The objective of step 3 is to acquire a list of suffixes. All the words in the input word list are segmented based on the suffixes. Morpheme boundaries are identified in the following words, which correspond to the following segmentation:
TECHNIQUES AND TOOLS FOR THE ANALYSIS OF TAMIL

MORPHOLOGY

STEP 4: Identification of Morphophonemic Process

The purpose of step 4 is to identify an array of morphophonemic changes. All the words in the input word list are segmented based on the morphophonemic changes which are called as Sandhi. Morpheme boundaries are introduced in this word, corresponding to the following segmentation:

a+p+palli_kaṭṭil “In those schools”

i+p+paiyaṇ+kṣṭā “In these boys”

e+p+paṭam+ṅ+kal+il “In which pictures”

Once a word has been segmented in this way, the proposed segment is identified as an affix in wordlist and at least once at the beginning of a word. The identified prefix is then is retrieved for statistical study. All the strings preceding this stem in the input wordlist are added to the list of prefixes since they are less frequent than the stems. Correspondingly, all the substrings following this stem in the wordlist are added to the list of suffixes they are frequent than the stem. All the substrings that may precede or follow the stem in the wordlist are added to the list of morphophonemic variants.
2.1.1 A Flowchart for the Segmentation of Words

FLOW CHART
2. 1. 2 Frequencies

Frequency-sorted word lists have long been part of the standard methodology for studying corpora. Sinclair (1991:30) states, Anyone studying a text is likely to need to know how often each different word form occurs in it. Tribble and Jones (1997: 36) outline a pedagogical methodology for using texts in the language classroom, proposing that the most effective starting point for understanding a text is a frequency-sorted word list. The frequency list records the number of times each word occurs in the text; it can provide interesting information about the words that appear (or do not appear) in a text. The list can be arranged in an order of first occurrence, alphabetically or in the order of frequency. First-occurrence order serves as a quick guide to the distribution of words in a text, an alphabetic listing is built mainly for reference, but a frequency-ordered listing highlights the most commonly occurring words in the text. For example, Juilland produced a series of frequency dictionaries for Spanish, Rumanian and French 1965 and 1970). Even the more traditional dictionaries can make use of frequency information. The texts on which the American Heritage Word Frequency Book (Carroll et al, 1971) was built formed the citation base for the American Heritage School Dictionary. Recently the LDC-IL produced a frequency dictionary of 5,000 distinct words for some of the Indian languages (LDC-IL report 2009).

Francis and Kučera (1982) take the simple word frequency list one stage further by reporting grammatical word frequencies. This gives frequencies of words with their associated part-of-speech (POS) tags in the (tagged version of the) Brown corpus. The frequency profile for a given text can be compared to that of other similar texts or to that of large bodies of text. Since the high frequency items tend to have a stable distribution, generally, significant changes to the ordering of the words in the frequency list can flag items of interest to the researcher (Sinclair 1991:31). Such techniques can be carried out manually for a small corpus but otherwise we need the aid of a
computer program. Although the computer saves us time with its processing of the texts into frequency lists, it presents us with so much information that we need a filtering mechanism to pick out significant items before the analysis can proceed. Hofland and Johansson (1982) use Yule’s K statistic and the chi-squared goodness-of-fit test in their comparison to pick out statistically significant different word frequencies across British and American English. Various formulae can be applied to adjust the raw frequencies for the distribution of words within a text, or to describe the dispersion of frequencies in subsections of a corpus.

Frequency profiling is one of the two main methods in corpus linguistics, the other being the use of concordance lines. A set of concordance lines presents instances of a word or phrase usually in the centre, with words that come before and after it to the left and right. The frequency of morphemes and words and their ngram techniques (bigram, trigram) would have crucial role in enhancing the NLP applications such as POS tagger, Machine Translation, Word Sense Disambiguation, Spell Checker and so on.

An important step in corpus processing is to convert the running text into a list of words and their frequencies. The program Splt_txt_to_wrds.pl takes a file as input, breaks it into individual words using space and punctuation as delimiters, and gives us output in terms of words and their frequencies. The program Splt_txt_to_wrds.pl is one of the most useful programs frequently used in the corpora analysis.

2.1.3 N-gram Methods

What is an n-gram?

An n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items in question can be phonemes, syllables, letters, morphemes, words or base pairs according to the application. N-grams are
collected from a text or speech corpus. An n-gram of size 1 is referred to as a ‘unigram’, size 2 is a ‘bigram’, size 3 is a ‘trigram’. Larger sizes of n-grams are sometimes referred to ‘four-gram’, ‘five-gram’ and so on. An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n-1) order Markov model. N-gram models are now widely used in probability, communication theory, computational linguistics (for instance, statistical natural language processing), computational biology (for instance, biological sequence analysis), and data compression.

The idea of word prediction; what word, for example, is likely to follow the following sequence of words.

...viraivil araciya caṭṭa tiruttam koṇṭuvara villai.... ‘Quickly they did not bring amendment to the constitution’

Hopefully most of you have concluded that a very likely word is vēṭum, or koṇṭuvarappāṭum or koṇṭuvarappāṭalām, but probably not the. We formalize this idea of word prediction with probabilistic models called N-grams, which predict the next word from the previous N-1 words. Such statistical models of word sequences are also called language models. Computing the probability of the next word will turn out to be closely related to computing the probability of a sequence of words. The following sequence, for example, has a non-zero probability of appearing in a text:

...nēṟu tiṭṟ eṟu eṟakku oru pōṟ vantatu... ‘Suddenly a phone call came to me yesterday’

while this same set of words in a different order has a very low probability:

....vantatu pōṟ tiṭṟ nēṟu oru eṟu eṟakku...’ To me one phone yesterday suddenly came’

As we will see, estimators like N-grams that assign a conditional probability to possible next word can be used to assign a joint probability to an entire
sentence.

An n-gram is an ordered sequence of two or more adjacent words, characters, or morphological elements separated by space. The idea of n-grams (in terms of bigrams and trigrams) are often used in studies in NLP. N-gram is the information of the association between n number of events. In order to build morpheme-based or word-based NLP tools the data sets are decomposed into their morpheme components. Therefore such an analysis requires every word is split into zero or more elements or zero or more prefixes followed by a stem (simple or compound) followed by zero or more suffixes.

2.2. (II) RESOURCES: THE CORPORAS

The tasks that are listed in the Aims and Scope of the study require that it shall be based on corpora of considerable size. For this study, we used the corpora of CIIL, EMILLE and Tamil BIBLE text for our research purpose. The CIIL corpus is analyzed first to find out exhaustive enumeration of various morphological categories. This is the basic resource for the study and understanding of Tamil morphology used in Modern Tamil.

We have also used the Tamil corpora of about 20 million words created by incorporating corpora of CIIL, Mysore and Lancaster University, UK and the Bible Society of India (ref. Paul Baker et al. 2003). The EMILLE-CIIL monolingual written corpora filenames in the written corpus consist of codes for the language and text type followed by a word to specify the source of the data, possibly followed by a subcategory. At the end of the filename is the date of publication of the original text (in the case of news data) or a code number (for other files). This is the main source for the study of frequencies of character level n-grams study.

These corpora, originally encoded as ISCII text, have been converted to Unicode with CES-compliant SGML markup language (cf. EMILLE Project).
TECHNIQUES AND TOOLS FOR THE ANALYSIS OF TAMIL
MORPHOLOGY

The filenames of texts from the CIIL Corpus differ slightly from those of the overall scheme. Because the CIIL Corpus was drawn from a much wider set of genres, the EMILLE Project’s data, it was not considered suitable to classify them by source. Instead they are classified by genre (category and subcategory). The filename is concluded by the code identifying that file in the original CIIL Corpus. Within the corpus structure these files are grouped in a directory entitled “miscellaneous” (because the data derives from a wide miscellany of sources). We have not included the directory entitled miscellaneous which is not cleaned but the same files we used (CIIL corpus) which are cleaned and converted into Unicode available at CALTS Language Technology Lab, Centre for Applied Linguistics and Translation Studies, University of Hyderabad.

The EMILLE-CIIL Monolingual Written Corpora have a total size of approximately 19,980,000 million words. Usually reliability of the statistical facts of a corpus depends on its size. In fact, the true reliability proportionately increases with the increasing size of the corpus (Xiao, Z., McEnery, A. et al 2004). The efforts in building large sized corpora for our Indian languages are yet to begin. However, the frequencies are relatively reliable for the most common members of words in the corpus.

2. 2. 1 CIIL Corpus

The Tamil corpus developed at CIIL (cf. Ganasen, M (1999)) is based on more or less proportionate distribution of texts appearing in print in various genres. Therefore the outcome of the study of this corpus is applicable to Modern Standard Written Tamil. The samples of texts belong to eight larger groups of genres as in the following:

1. **Aesthetics**: This contains texts of literature viz., novel, short-stories, essays, criticism, humour, children’s literature, biographies, travelogues, letters,
speeches, plays, science fiction, folk-tales and text books. The texts in Aesthetics contain 190 files.

2. **Fine Arts:** This contains texts of music, dance, sculpture and hobbies. The texts in fine arts have 36 files.

3. **Social Sciences:** This group of texts contains sociology, linguistics, psychology, anthropology, history, epigraphy, political science, home-science, public-administration, catering-technology, religion, philosophy, economics, logic, journalism, folklore, mythology, law, astrology, criminology, yoga, health, sexology, culture, sports & games, personality, anecdote, tailoring and text of the relevant text books. The whole social science group contains 310 files. This is the one which contains more files but not many words.

4. **Natural and Physical Sciences:** This includes texts from botany, zoology, geology, microbiology, physics, computer-science, astronomy, medicine, ayurveda, engineering, oceanography, agriculture, veterinary, film technology, marine biology, fisheries and the texts from relevant text books. This group of texts contains 135 files.

5. **Commerce:** This group contains texts of business management, industry and handicraft. The text in this group contains 6 files.

6. **Official & Media Language:** This group contains of Legal, Administration and Mass media Texts. This group contains 56 files.

7. **Translation Material:** This is a mixture of all groups because this contains collection of translated texts. This includes texts from Literature, Psychology, Legal, Philosophy and Education. The translated materials contain 18 files.

8. **Spoken Tamil:** This is the one, which does not belong to any of these groups. This contains 8 files. Data from the CIIL Corpus include is 3, 050, 360 (30 lakhs) words of running texts.

The Tamil corpus that is used in this study is developed by CIIL and funded by the Ministry of Information Technology. It contains 3 million words of
modern written Tamil. The text materials belong to different years, mostly published during the period of 1980’s and early 1990’s. The texts include mostly written texts as well as a small portion of transcribed spoken texts. The samples of texts belong to eight genres as mentioned above.

2. 2. 2 EMILLE Corpus

The EMILLE corpus is a news corpus which is developed from the daily newspaper Dinakaran. The Tamil EMILLE corpus also contains data incorporated from the CIIL Corpus, originally gathered by the Institute. The contents of the Tamil written corpus are as follows (broken down by directory): Data from the Dinakaran website, divided into five parts such as Cinema, News, Politics, Sports and Other. The Cinema contains film and related news (approximately 1,050,000 words) and The News contains news stories (approximately 8,610,000 words) and The Politics bears political news and commentary (approximately 5,050,000 words) and The Sports contains sports news reportage (approximately 1,170,000 words) and finally the directory Other contains other stories from the website (approximately 730,000 words) The Tamil written corpus altogether contains a total of approximately 19,980,000 words.

2. 2. 3 BIBLE Text Corpus

The Bible consists of a collection of sixty-six separate books. It has broken into two major sections: The Old Testament which consists of thirty-nine books, and The New Testament which consists of twenty-seven books. It has altogether 1189 chapters both in the Old Testament and the New Testament. The Bible text corpus contains a total of approximately 4,25,000 words.

2. 2. 4 Encoding Standards

The encoding standard that was used is ISCII (ISO 1988/1993) the Indian Script Code for Information Interchange. All the Indian scripts that are derived from the Brahmi script share this encoding. For this purpose an
optimal keyboard is made. There are manifold advantages in having a common code and keyboard for all the Indian scripts. The ISCII standard specifies a code table which can be used in 7 or 8-bit code (ISO compliant) environment. It allows Latin/roman and Indian script alphabets to be used simultaneously. The ISCII text requires a Tamil font in order to be displayed in Tamil for reading and processing. However, we have chosen for the internal processing the Roman notation (wx-notation) for the purpose of easy processing. WX-notation is an extremely convenient roman transliteration scheme for all Indian languages. It is more or less 'phonetic roman' except for a small number of roman characters. This scheme is characterized by the use of the characters w-x for dental stops, hence the name WX-notation. There are a number of advantages using the text in WX-roman notation, since it follows a strictly Tamil orthographic scheme, which is near phonemic in nature. It is easy to convert from this roman transliteration to ISCII and display it in Tamil at the stroke of a button on the keyboard.

2.3. (III) PROGRAMMING TECHNIQUES

To analyze the corpora, we used Linux/Unix tools, which are readily available in the Linux environment. We have used a number of programs written in PERL which are developed and available at the CALTS Language Technology Lab and adopted for this study with some minor changes. Here are the tools used for the present study.

2.3.1 Linux/Unix Tools for Analyzing Corpora

As it was mentioned above, we have used Linux/Unix tools for analyzing the Tamil corpora. Linux/Unix tools viz. cut, paste, grep, sort, wc, etc. can be used in executing a number of functions plus a set of options while executing these. Vi-editor has been extensively used to see the intermediary stages of processing and for some minor editing requirements. Here is a list of certain
TECHNIQUES AND TOOLS FOR THE ANALYSIS OF TAMIL MORPHOLOGY

tools and commands on Linux, which are useful in the current study of the Tamil corpora:

Cut

The cut command is used to split a file vertically to manipulate fields. It is a powerful text manipulator often used in combination with the other commands or filters. A character, a space, a punctuation mark can be used to delimit the field. It is an extremely useful command to manage contents of files to swap the fields, words and frequencies into frequencies and words etc. As an example: the command cut –c –d, -f2 <input file> f2.out, extracts the second field (f2) from the input file and puts in f2.out.

Paste

This command is used to create a file by combining one or more files vertically containing one or more fields. What one can cut with cut can be pasted back with the paste command but vertically rather than horizontally. The command paste also uses the tab as the default delimiter as cut uses. The syntax of this command is –
paste file1 (input) file2 (input) > file3 (output)

Grep

The grep command is generally used to extract a sequence of one or more strings of one or more files from one or more directories. This is one of the most important UNIX commands and it looks at its input for a pattern and displays lines containing the pattern, the line numbers or filenames where the pattern occurs. This command is extremely useful for checking a particular word or a linguistic unit from one or more files of the entire corpus. The syntax for the execution of this command is –
grep 'pattern' <input file> output file.
Sort

The sort command is generally used for sorting strings of one or more characters distributed over a number of lines in the desired way. This command is used for ordering of data in ascending or descending sequence. Like cut, it identifies fields and it can sort on specified fields. It can also be used with a number of optimal features. For this, the syntax is –

```
sort -n -d -r < input > output sorted file
```

(Where n= numerical, d= dictionary and r= reverse order).

wc

Unix features a common word-counting program that also counts line and characters. The wc command counts lines, words, and characters in a given file. For this, the syntax is – wc filename

Vi-editor

The vi-editor is a screen editor, which is one of the most powerful text editing tools available in Linux/Unix environment and can be conveniently used simultaneously to display ISCII text in an Indian script using a Gist Terminal. Vi-editor can be used as a multilingual editor on GIST terminal.

Shell

As Victor Gedris (2003) states that Shell is a program that interprets commands and allows a user to execute commands by typing them manually at a terminal, or automatically in programs called shell scripts. A shell is not an operating system. It is a way to interface with the operating system and run commands.

Unix uses shells to accept commands given by the user, there are quite a few different shells available. The most commonly used shells are
TECHNIQUES AND TOOLS FOR THE ANALYSIS OF TAMIL
MORPHOLOGY

SH(Bourne SHell) CSH(C SHell) and KSH(Korn SHell), most of the other shells you encounter will be variants of these shells and will share the same syntax, KSH is based on SH and so is BASH(Bourne again shell). TCSH(Extended C SHell) is based on CSH.

The Basic Shell Commands are 

- **ls** - list files in current directory,
- **cp** - srcfile destfile copy file "srcfile" to new file "destfile",
- **mv** - oldname newname rename (or move) file "oldname" to "newname",
- **rm** - myfile remove file "myfile",
- **mkdir** - subdir make new directory "subdir",
- **cd** - subdir change current working directory to "subdir",
- **rmdir** - subdir remove (empty) directory "subdir",
- **pwd** - display current working directory,
- **date** - display current date and time of day,
- **man command** - display man page for "command",
- **exit** - exit a terminal window,
- **logout** - logout of a console session.

**Perl Programs**

Perl is an acronym for “Practical Extraction and Reporting Language”. It is used as a scripting language and as an efficient string manipulating language using the regular expression. It incorporates syntax elements from the Bourne shell, csh, awk, sed, grep etc. Here, for the current purpose, we have used various programs written in PERL to perform various tasks such as extracting individual words from running texts from various files, counting frequencies of words, morphemes, characters and their bigram, trigrams have been extracted from word forms etc. and calculating the percentages along with probability of the various units. In the 3rd chapter the cursory details of some of the programs and their functions are presented.
2. 4 (IV) IMPLEMENTATION, TESTING AND EVALUATION

Segmented Words

In the final step, all the words are segmented. Word segmentation is performed by comparing words which contain the same stem to one another. This consists finding limits between shared and different segments and results in a segmentation of the words being compared. The outcome can be represented as an alignment graph for each stem. Figure-1 depicts the alignment graph obtained for the words containing the Tamil stem *maram*.

\[ \text{Figure-2: Exemplifies segmentation of words sharing the stem/root } maram \]

Segments are subsequently labeled with one of the four labels (prefix, suffix, stem, morphophonemic change) according to their positions within the word, relatively to the stem. All the segmentations are made up of valid morphemic segments.

The method involves both the determination of the correct morphological split for individual words, and the establishment of accurate categories of stems based on the range of suffixes:
1. **Splitting words**: We try to accurately analyze any word into its successive morphemes in a fashion that corresponds to the traditional linguistic analysis. Minimally, we identify the stem, as opposed to any inflectional suffixes. Ideally we would also identify all the inflectional suffixes in a word which contains a stem that is followed by two or more inflectional suffixes, and we would like to identify derivational prefixes and suffixes. We would like to state that in these corpora, the most important suffixes are -il, -kaḷ, -ai, -kku, -iṅ, -ukku, -iṭam, -kkal, -uṭaṅ, -āl, -ukkāka, -ōṭu, -iliruntu and so on.

2. **Range of suffixes**: The most important characteristic of a stem in the language that we consider here is the range of suffixes with which it can appear. Nouns in Tamil, for example, will appear with some subset of the suffixes or suffix complexes kaḷ, kkal, ŋkaḷ, mār, kaḷai, kaḷukku, kaḷōṭu, kaḷukkāka etc. We would like to determine what is the range of the most regular suffix groups for the morphology of the language in question, and rank suffix groupings by order of frequency in the corpora. In addition, one would like state general rules of allomorphy as well; for example, a statement that the stems mara and marat (as in maraṅ-kaḷ and mara-tt-il, respectively) are forms of the same morphological stem *maram*.

2. 5. (V) **QUANTITATIVE APPROACH TO MORPHOLOGY**

The present study attempts to reveal the statistical facts of morphological units, their distribution and their relevance in building NLP tools. It also helps to find out the characteristic patterns of the Tamil morphological elements. Quantitative analyses of corpora provide comprehensive approaches in understanding complexity of the morphology of the language. The study involves the analysis of the corpora to find out the frequencies of Tamil morphological units viz. prefixes, stems, suffixes and other various morphological entities, their distribution, comparison and the linguistically relevant interpretation based on these. In a way it lays focus on
the frequency occurrence of morphemes and words and their coverage, and their bigramic, trigramic probabilities associated with the relevant categories. Most frequent affixes and their variants, most frequent patterns in combination, most frequent allomorphs, most frequent stems/roots, frequencies of morphophonemic changes, most frequent character combinations of two to three, the relevance of their frequencies with respect to their allomorphy and their distribution in the corpora will be discussed. Finding out the frequencies of certain morphological types and categories over other may reveal a number of facts about the morphology. The corpora study would reveal both qualitative and quantitative aspects of Tamil morphology. Quantitative tools are a vital source in examining and exploring the importance of corpora with immense speed and accuracy to get near exhaustive information related to words, which is of great importance in building morphological tools. Our interest is to find out the properties of various morphological units, in terms of their quantity with respect to similar others. We propose to find summaries of arithmetical counts and their probabilities of each morphological unit in terms of frequencies, percentages and cumulative percentages, to illustrate their coverage to the whole corpora. The outcome of this study would enhance the development of NLP applications, but also in a greater way to understand the morphology of Tamil.