6

Tamil Text Analyser

K. Rajan,

Muthiah Polytechnic College, Annamalainagar.

Dr. M. Ganesan,

CAS in Linguistics, Annamalai University.

Mr. V. Ramalingam,

Dept.of Computer Science & Engineering

Introduction

Much computer-aided text-based research in the humanities is carried out using different tools and techniques. Applications of these tools include lexical research, stylistic analysis, lexicography, and almost any other task based on finding specific instances or repeated patterns of words. Certain types of clauses or constructions can be identified by words which introduce them. Inflections can be studied by specifying words that end in certain sequences of characters. Punctuation or other special characters can also be used to find specific sequences of words. Numerical studies of style and vocabulary are not new, but with the advent of computers much larger quantities of texts can be analyzed, giving an overall picture that would be impractical to find by any other means. From the 1960s into the 1990s, computational linguistics developed primarily through the work of computer scientists interested in string manipulation, information retrieval, symbolic processing, knowledge representation and reasoning, and natural language processing. The NLP community has been especially interested in analysing text-based inputs and out-puts. Using text inputs is a standard practice in linguistics among those who study syntax, semantics, pragmatics, and discourse theory. Apart from creating natural language text, using text editors, analysing the text is one of the important aspect of language studies.

In this paper we discuss the usefulness of software tools for NLP researchers in relation to Tamil Corpora. We used the corpus developed by CIIL, Mysore for our testing. The corpora are precious aids to the NLP researchers attempting to design systems that can handle language as it is really used. The features of the software tool are presented here.

Language analysis

Studies of language can be divided into two main areas:

Studies of structure and studies of use. Linguistic analyses have emphasized structure, identifying the structural units and classes of a language (e.g. Morphemes, words, phrases and

sentences) and describing how smaller units can be combined to form larger units. Studies of 'language use' focus on a particular linguistic structure, investigating the ways in which similar structures occur in different contexts and different functions. Corpus can be used to provide more useful information on morphemes, words, sentences, etc.

Those who work in Natural Language Processing require flexible access to large corpora. It is not necessary that such corpora be supplied exhaustively analyzed. What is required is a set of tools that the NLP researchers can use to process the corpora to yield interesting views over the data and to elicit various patterns, clusters and regulations. These can then form the basis for either the writing of rule-based system or the training of probabilistic models. Furthermore, they can be used as input to various other tools.

Raw Corpora are necessary to allow useful aids to be generated such as concordances and various sorting which are invaluable for the grammar and dictionary writer. Clearly various statistical operations may be carried out on raw corpora that help computational linguists to characterize texts from various points of view, or allow them to identify frequently or infrequently occurring words, or other patterns. Raw corpora can be used to develop and train probability-based models. If a corpus is to be useful, we need to search it quickly and automatically to find examples of a particular linguistic phenomenon to sort the set of words and to present resulting list to the user.

Partial analysis of corpora can yield useful patterns and structures. Analyzing Tamil corpora is different from analyzing English language corpora. The existing tools for English text processing are not suitable for processing Tamil text. The difficulties at various levels of analyzing Tamil text are due to the large set of characters and the encoding system.

The major task of the software tool is the presentation of the text data and analysis for linguists or researchers to review and use.

This software tool has the following features:

- 1. Text Editor
- 2. Text Database Manager
- 3. Pattern Search
- 4. Concordance
- 5. Sorting Utility
- 6. Tagging
- 7. Phrase Chunking
- 8. Statistical Analysis

Text Editor

The text editor is a Window based Tamil text editor with basic features of Notepad and Tamil keyboard support (TAM/TAB). Searching on Tamil text files can be done. Using this editor the user can perform manual tagging. For easy searching and replacements, it provides updateable search list and tag list. The find and replace facility differentiate selected words in colors. Certain types of clauses or constructions can be identified by words which introduce them. Inflections can be studied by specifying words that end in certain sequences of characters.

தமிழ் இணையம் 2003, சென்னை, தமிழ்நாடு, இந்தியா

| | <u>E</u> dit <u>V</u> iew Fo <u>r</u> mat <u>M</u> | | ow | | | | | |
|--------------|---|---------------------|---|---|---|-----------|---------------------|-----------|
| | | XM | | | | | - 1 | |
| <u>I</u> | | | | | - | 14 | A | |
| | | | | | Tamil Key | | | 4 |
| | A S ers | | 5 G G | •9 7 3 8 6 7 • 901 U (5 • 7 • 901 U (5 • 7 • 901 U (5 • 7 • 901 U (5 • 7 • 901 U (5) • 7 • 7 • 901 U (5) • 7 • 901 U (5) • 7 • 901 U (5) • 7 • 7 • 7 • 7 • 7 • 7 • 7 • 7 • 7 • 7 | 9 (0) 8 9 0 1 28 0 4 28 0 4 5 9 0 1 28 0 4 5 9 0 1 29 0 1 20 | | T SUD Enter + | |
| m | Z esp X | 9 C 28 | V 6J 67 | Bomp N (05) G C | M 38 < # > | 7 7 - 4 | Shift | 1 |
| - | sau | 9 • | ar | | | | - | 4 |
| | Alt | | | Space | | Alt C | Ctrl | |
| _ | | | | | | | | |
| | நெல்லிக்காயை | | | | | | | |
| | ம். முதலில் எ போடவும். கடு | | | | | | | |
| | மஞ்சள் பொடி (| | | | | | | |
| ШΠ | க வெந்தவுடல | ர தேன | வ்யா | ன உபு, ഥിഖ | ாகாயப் | | | |
| | போட்டு நன்ற வெந்தயம், பெ | | | | | | | |
| | ் வைந்தயம், பெ 6 வைக்கவும். , | | | | | | | |
| æ | கவும். இ சப்பா | | | | | | | |
| ണ | ள ஏற்றது. | an.090 GR | | 180 Real | | | | |
| | | | | | | | | |
| art | 🗌 🔄 Editor 🛛 😤 | , Project1 - | ግ 🛛 | untitled - P | Tamil E 🛛 🗍 😂 (| 🥌 🧭 🧭 🧾 | × 3 | 🈂 🌾 🥦 👔 1 |
| | | | | | | T alita a | | |
| | | | | Fig.1 In | e layout of the l | Editor | | |
| | | | | | | | | |
| | Word Collection | | | | | | | _ 8 × |
| | Word Collection e <u>W</u> ords <u>S</u> ort <u>P</u> attern N | /latching | | | | | | × |
| | | fatching | | | | | _ | × |
| <u>F</u> ile | a <u>W</u> ords <u>S</u> ort <u>P</u> attern N Words Read : [2026 Word [| flatching Freque | ncy | Frequency | Word | | 5 | _[#] × |
| Eile | e Words Sort Pattern M Words Read : [2026 Word පියා ලාසි රිදිනා පීයා | | 1 🗌 | 12 | சேரத்து | | 5 | _[6] × |
| Eile | з <u>W</u> ords <u>S</u> ort <u>P</u> attern № Words Read : 2026 Word &பாடுகிறோமோ யோடும | | 1 | 12 | சேரத்து எடுத்து | | 5 | _[6] × |
| Eile | s <u>W</u> ords <u>S</u> ort <u>Pattern M</u> Words Read : 2026 ^{Word} போடுக்றோமோ போடும் போடுவதற்குப் | | - - - | 12 12 13 | சேரதது எடுதது வெயிலில | | - | _[6] × |
| <u>F</u> ile | Words Sort Pattern M Words Read : 2026 Word هـ المحالية ال حمالية محالية محالية المحالية المحالي حمالية محالية محاليحالية محالية محالية محالية محاليحاليجانية محالية محالية محا | | | 12 12 13 13 | சேர்த்து எடுத்து வெயிலில் நன்றாக | <u> </u> | 5 | _[6] × |
| <u>F</u> ile | <u>Words Sort Pattern M</u> Words Read : 2026 Word Curr டுகிறோமோ போடுகிறோமோ போடும போடுவதற்குப் போட | | | 12 12 13 13 13 | ்சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது | | 5 | 6]× |
| | <u>Words Sort Pattern M</u> Words Read : 2026 Word Curr டுகிறோமோ Gurr டும் Gurr டும் Gurr து Gurr டெல்னர்டும் | | | 12 12 13 13 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிள்காய | | 5 | _[6] × |
| | Words Sort Pattern M Words Read : 2026 Word Curr ြန်းစြာကလေက போடுக்றோலோ போடுவதற்குப் போடுவதற்குப் போட போட வேண்டும் போட்டிருந்த | Freque | | 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிளகாய கிளறி | | | 6 × |
| | <u>Words Sort Pattern M</u> Words Read : 2026 Word Curr டுக்றோமோ Gurr டுவதற்குப் Gurr டுவதற்குப் Gurr ட போட டேவன்டும் Curr டுத்த Gurr டுத | Freque | | 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிளகாய கிளறி தணாணீர | <u> </u> | | _[#] × |
| | Words Sort Pattern M Words Read : 2026 Word யாருட்கிறோமோ யோருட்கிறோமோ யோருட்கிறோது யோருவதற்குப் யோருவதற்குப் யோருவதற்குப் யோருவனரும் யோருந்த யோருப் யோருப் | Freque | 1 1 1 4 2 1 49 1 | 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிளகாய கிளறி | <u>*</u> | | [#] × |
| | Words Sort Pattern M Words Read : 2026 Word Китсрек Корт Силт Силтсрек Корт Силт Силтсрек Силте Силте Силте | Freque | 1 1 1 4 2 1 49 T 6 | 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிள்காய கிள்றி தண்ணீர் கடுகு செய்து விட்டு | | | [#] × |
| | Words Sort Pattern M Words Read : 2026 Word யாருக்றோமோ யோருக்றோமோ யோருவதற்குப் யோது யோட வேண்டும் யோட வேண்டும் யோட ருது யோட ரு | Freque | | 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிள்காய கிளறி தண்ணீர் கடுகு செய்து விட்டு பிறகு | | | 6 × |
| | Words Sort Pattern M Words Read : 2026 Word யோடுக்றோமோ யோடும் போடுவதற்குப் போடுவண்டும் போடும் போடும் போடும் போடும் போடும் போடும் போட்டும் | Freque | | 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 | சேர்த்து எடுத்து வெயிலில் சற்றது மிள்காய கிளறி தண்ணீர் கடுகு செய்து விட்டு பிறகு வைக்க | | | 6 × |
| | <u>Words Sort Pattern M</u> Words Read : 2026 Word 2026 | Freque | | 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிள்காய கிள்றி தண்ணீர் கடுகு செய்து வட்டு பிறகு வைக்க உபு | | | 6 × |
| | <u>Words Sort Pattern M</u> Words Read : 2026 Word CurrGet®pr@urr GurrGet®pr@ur GurrGet@pr@u GurrGet@pr@u GurrGet@etr GurrGet | Freque | | 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 | சேர்தது எடுத்து வெயிலில் நன்றாக சிறிது மிளகாய கிளறி தனானீர் கடுகு செய்து விட்டு பிறகு றைவகக உபு கொள | | | 6 × |
| | yudis Soit Pattern N Words Read : 2026 Word Currchek®pr@urr Currchek®pr@urr Currchex@pepu Currcamorr(hu) Currche(hppg) Cur | Freque | | 2 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4 5 6 7 7 7 9 20 21 23 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிளகாய கிளறி தனானீர் கடுகு செய்து வட்டு பிறகு வைக்க உபு கொள் வைக்கவும் | | 5 | 6 × |
| | yudis Soit Pattern N Words Read : 2026 Word Currchek®pr@urr Currchek®pr@urr Currchex@ppgu Currc Currchem@pga Currc Currchem@pga Currchem@pa Currchem@pa Currchem@urr Curr Currchem@urr Currchem@urr Currchem@urr Currchem@ur | Freque | | 12 12 13 13 13 13 13 13 13 13 13 13 14 14 15 16 17 17 19 20 21 223 223 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிள்காய கிளறி தண்ணீர் கடுகு செய்து விட்டு பிறகு தைவக்க உபு கொள்ள தைவக்கவும பொடி | | 5 | 6 × |
| | Words Sort Pattern M Words Read : 2026 Word Curr Geberger Control Geberger Control Geberger Centrol Curr Geberger Curr Geberger Curr Geberger Curr Geberger Curr Geberger Curr Geberger Curr Control Curr | Freque | | 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிள்காய கிளறி தண்ணீர் கடுகு செய்து விட்டு பிறகு வைக்க உபு கொள்ள வைக்கவும் பொடி நன்றாக்க | | | 6 × |
| | <u>Words Sort Pattern M</u> Words Read : 2026 Word Curr Get Con Curr Cur Curr Curr Curr Curr Cur Curr C | Freque | | 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 2 3 2 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிள்காய கிள்றி தண்ணீர் கடுகு செய்து விட்டு பிற்கு வைக்க உபு கொள்ள வைக்கவும் பொடி நன்றாகக கலந்து | | | 6 × |
| | Words Sort Pattern M Words Read 2026 Words Read 2026 Word CurrGet Read 2026 Wurd CurrGet Read 2026 Wurd CurrGet Read 2026 Wurd CurrGet Read CurrGet Read CurrGet Read CurrGet Read CurrGet Rea | Freque | | 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிள்காய கிளறி தண்ணீர் கடுகு செய்து விட்டு பிறகு வைக்க உபு கொள்ள வைக்கவும் பொடி நன்றாக்க | | | 6 × |
| | <u>Words Sort Pattern M</u> Words Read : 2026 Word Curr Get Con Curr Cur Curr Curr Curr Curr Cur Curr C | Freque | | 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது மிள்காய கிள்றி தண்ணீர் கடுகு செய்து வட்டடு பிற்கு வைக்க உபு கொள்ள வைக்கவும் பொடி நன்றாக்க கலந்து வைத்து | | | 6 × |
| | Words Sort Pattern M Words Read : 2026 Words Read : 2026 Words Read : 2026 Word Currey Read : 2026 Word Currey Read : 2026 Word Currey Currey Read : 2026 Currey Gurey Currey Read : 2026 Currey Gurey Currey Currey <td< td=""><td>Freque</td><td>I I<td>2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3</td><td>சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது யிள்காய கிள்றி தனானீர் கடுகு செய்து விட்டு பிறகு வைக்க உபு கொள்ள ைவக்கவும் பொடி நன்றாக்க கலந்து வைத்து கொள்ளவும்</td><td></td><td></td><td>6 ×</td></td></td<> | Freque | I I <td>2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3</td> <td>சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது யிள்காய கிள்றி தனானீர் கடுகு செய்து விட்டு பிறகு வைக்க உபு கொள்ள ைவக்கவும் பொடி நன்றாக்க கலந்து வைத்து கொள்ளவும்</td> <td></td> <td></td> <td>6 ×</td> | 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 | சேர்த்து எடுத்து வெயிலில் நன்றாக சிறிது யிள்காய கிள்றி தனானீர் கடுகு செய்து விட்டு பிறகு வைக்க உபு கொள்ள ைவக்கவும் பொடி நன்றாக்க கலந்து வைத்து கொள்ளவும் | | | 6 × |

Fig. 2 Showing the word list with frequency

| bid | | equency Word |
|--------------------------|----------------------|---|
| காளவும | | சேரத்து |
| லக்கவும் | 8 12 | எடுத்து |
| நித்தவும் | 8 13 9 13 8 13 | 🖷 Pattern Matching 🛛 🛛 🗙 |
| காளவும் | 9 13 | |
| தக்கவும் | | |
| லக்கவும் | 8 13 | Enter the Pattern AGLD |
| றக்கவும் | 8 14 | |
| சரக்கவும் | | |
| டவும் | 6 15 | |
| பாடவும் | 7 16 | Find Add to List Delete From List Close |
| தக்கவும் | 8 17 | |
| ன்றவும் | 7 10 | Show All Save |
| ഖക്കപ്പാ | 8 20 | |
| ഖക്കപ്പം | 8 2- | |
| காளவும் | 9 2 | തവക്കവധ |
| சிக்கவும் | 9 2 | பொட |
| டிக்கவும் | 8 2 | நன்றாகக் |
| காளவும | 9 2 | கலந்து |
| தக்கவும் | 8 28 | வைத்து |
| ரப்பவும் | 9- 30 | கொள்ளவும் |
| | . 3 | செய்முறை |
| Words List Total Words : | 2026 | പേപ്പ്ര |
| Words List Total Words : | 2020 7 7 76 | வேண்டும் |

€ jig 3. Showing the Pattern Search

| Words | Frequency | |
|--------------|----------------------------|---|
| 🖷 Pattern M | atching | × |
| Enter the Pa | ittern எல்ல் ஆம் | - |
| | க் இன்ற் அத்உ அக் வ | |
| Find | Add to இத்உ ஆள் | |
| | ஆய் ஆய் க்ஒண்ட் இர்உ | |
| | | + |

Fig. 4 Showing the Search list for easy entry of pattern (Words are in Consonant-Vowel form)

Text Database Manager

The plain text files can be segmented into sentences and each sentence can be segmented into phrases. The words are collected and stored for further analysis.

The text database manager creates and maintains a database of words. It performs basic functions of counting, searching, filtering, sorting and preparing concordances.

Word List

A word list is a list of words retrieved from a particular topic or subject text where each word is accompanied by a frequency number. The list can be viewed by

- the order of word
- the order of frequency
- the order of word length

The words may be viewed in a normal form using TAM/TAB encoding or as a group of consonant and vowels which gives clear view of the word.

Sorting

The word list can be sorted in alphabetically ascending and descending order of letters. Words can be sorted by their endings. As already seen, words can be sorted by their frequency, starting with the most frequent word or less frequent, or even by their length where the longest or the shortest word comes first. A process called reverse alphabetical sorting, sort the words by their endings.

Searching

The word list may include every word or only selected words. Words can be selected using wildcards, such as * and ?. The symbol '*' denotes any number of letters including none, '?' denotes any single letter. In many situations, this approach can be much more productive than attempting to use morphological or syntactic analysis programs.

Phrase Chunking

Text chunking is dividing sentences into non-overlapping phrases. Noun phrase chunking deals with extracting the noun phrases from a sentence. While NP chunking is much simpler than parsing, it is still a challenging task to build a accurate and very efficient NP chunker. The importance of NP chunking derives from the fact that it is used in many applications.

Noun phrases can be used as a pre-processing tool before parsing the text. Due to the high ambiguity of the natural language exact parsing of the text may become very complex. In these cases chunking can be used as a pre-processing tool to partially resolve these ambiguities.

Noun phrases can be used in Information Retrieval systems. In this application the chunking can be used to retrieve the data's from the documents depending on the chunks rather than the words. In particular nouns and noun phrases are more useful for retrieval and extraction purposes.

Concordance of words

The concordance program of this software lists the specified word in the order in which they occur in the text. The number of words in the context can also be specified.

Tamil Internet 2003, Chennai, Tamilnadu, India

| மாறிலம் | பிரயாவிலேயே | தொழில் | കൃത്ത്വയില് | மிகவும் | வளர்ச்சி |
|---------------------------|---------------------|------------------------|----------------------------|------------------------|------------------------|
| ஓன்றாகும். | பதினெட்டாம் | தூற்றாண்டின | இறுதியில் | வெடித்த | பிரைஞ்சு |
| ഖഹന | திர் | உயர்நிலைப் | പണ്ടിലില് | பயின்றார். | ണംബംബം |
| வரலாற்றுச் | சக்தி | என்ற | ധ്രക്തുധില് | அரங்கேற்றியது | நிலப்பிரபுத்துவத்தின் |
| இல்லாத | அரை | றிலப்பிரபுத்துவ | ിെറ്റുവങ്ങിലില് | ቁ. | ്ഗ്രക്കാണിക്കുഖ-ഇങ്ങുന |
| கடைசியிலும் | நூற்பதுக்களின் | தொடக்கத்திலும் | ട്രെന്നത്തിലില് | பெருந்திரளான | ധക്കണ്ണംപ്രില |
| ഥക്കണ്ണെപ്ധില | அருப்திகள், | சமூக | வாழக்கையில | ହ ୍ତ | வேகம், |
| வரக்கம் | மற்றும் | அறிவுப் | பகுதியினரிடையில் | பல்வேறு | எதிரப்புப் |
| 1842 | ஏப்ரலில் | அந்தப் | പള്ളിന്ത്രകലില | കോണങ്ങളും | சேரந்தார; |
| 6வது | கூரன | பிரதிற்திகள் | ക്കവധില | പ്രമാ ഥരിവരുന്ന | விவாதம |
| இது | கட்டுரைகளை | எழுதினார். | பத்திரிகையில | കിമാലക്ക | அனுபவம் |
| எழுச்சி | பிரிட்டனில் | 1830 க்களின் | കണംപകിലില് | ஏற்பட்ட | பாட்டானி |
| வரக்கத்துக்கும் | Ren | வரக்கத்துக்கும் | இடையில் | வரக்கப் | போராட்டம் |
| கடுமையான | ട്ടങ്ങിക്കുകയെല | எற்படுத்தியது | കണംപകിധിയ | 1843 | எப்ரல் |
| மாரக்ஸ் | Rheinische | Zeitung | பத்திரிகையில் | பணியாற்றிய | காலத்தில் |
| காட்டினார். | ஃபாயர்பாஹிடமிருந்து | வித்தியாசமான | முறையில் | இயற்கைகைய | மட்டுமல்லாமல் |
| அவர | எழுதிய | பத்தகத்தின் | முன்னுரையில | எடுத்துரைக்கப்பட்டன. | பக்கம்:10 |
| மாறாக | ചുംകള | விமர்சன | ശ്രത്വാധില് | திருத்தியமைக்கும் | പങ്ങിമാധ |
| 1843 | அக்டோபர் | மாதக் | കണംപകിധിരാ | மாரக்ஸ | பாரிசுக்குச் |
| புராதன, | கொச்சைத் | தனமான | ഗ്രക്തുധില | சமனாககும | கம்யூனிசத்தை |
| കുന്ന്നില്യക്കുഖിലാക്കലം. | பதினெட்டாம் | தூற்றாண்டின் | இறுதியில் | ஏற்பட்ட | பிரைஞ்சு |
| மாறுவகைத | ற்றனாயித்தன. | மாரக்கின் | வாழக்கையில் | இந்தத | தீமானமான |
| ("ஜெரமன்–பிரெஞ்சு | ஆண்டுமலர") | என்ற | சஞ்சிகையில | அவர் | எழுதிய |
| பிரதிபலிக்கிறது | இச்சஞ்சிகை | 1844 | பிப்ரவரியில் | பாரிகில | மாரக்ஸ |
| தலைப்பில் | மாரக்ஸ | எழுதிய | கட்டுரையில் | தேசியப் | பிரச்சிகைகாடைப் |
| வளமும் | மிகவும் | குறிப்பிடத்தக்கதாகும். | கட்டுரையில் ஜெர்மனியில் | மதத்தைப் | பற்றிய |
| பொருளமுதல்வாத | ചാടങ | கணணோட்டத்தின | அடிப்படையில | வளரத்தார். | பாட்டாளி |
| ளங்கெல்ல | இங்கிலாந்துக்குச | செல்லும் | ഖழியில | கொணே | தங்கிய |
| 1844 | ஆஸ்ட | மாத | இறுதியில் | ണങ്ങികയാസ | பாரிசுக்கு |

Fig. 5 Concordance

Tagging

Tagging of words for their lexical and grammatical categories can be done by this system. The use can search for a particular pattern and assign a grammatical value. Certain type of categories of words have common suffixes. This can be studied. If we use a large lexicon, tagging can be done for more number of words.

Tagging can be done at different levels. Syntactic level tagging will be used for the analysis of phrase structure and to study the sentence patterns. Syntactic tagger will produce the output as shown below. The word level tagged text is the input for this.

```
அப்பம் < N > [NP Nom]
              தின்ற< RP >
(φωώ (N) (NP Subj)
அது < PN 3sn > {NP_Subj}
©(ns≺Det>
மலைக் காடு. < N > {NP_Nom }
©(ns < Det >
Gufflu < Adj >
ഥഞങ≺N> {NP_Nom }
அதன் < PN 3sh Gen >
சரிவுகளில் < N_pl_Loc > {NP_Loc }
Guffiu < Adj >
Guffu (Adj)
Lognsieser < N_PI≻ {NP_Nom }
              வளர்ந்து≺V_P>
              esr∟riù ≺Adv>
              usociru, <V P>
              ®min (V P>
```

Fig 6. Output of a Syntactic tagger

Conclusion

Tamil software for Desk top publishing is available with more features. But for Natural Language Processing, we also need software which make the system to understand the Tamil Language. The development of software components in this area are considered important for the linguistic research and expert system development. In this work we have tried to develop software tools which help linguistics for their research. The efficient and user friendly software tools will reveal more information for the researchers.

References:

- 1. Geoffrey Leech and Steven Fligestone, 'Computers and Corpus analysis' in *Computers and Written Text*, Christopher S. Buller (ed), 1992, p. 115-140.
- 2. Akshar Bharati, et al, A Computational Grammar Based on Paninian Framework, Kanpur, I.I.T., 1993.
- 3. Geoffrey Leach, Corpus Annotation Schemes, Literary and Linguistic Computing, Vol. 8, No.4, 1993, p. 275-280.
- 4. Terry Patten, Computers and Natural Language Parsing in *Computers and Written Text*, 1991.
- 5. Thiyakarajan S, "Noun Phrase Chunking", AU-KBC, MIT, Chennai.
- 6. John M.Lawler (ed), et al, Using Computers In Linguistics, Routledge, London
- 7. Rajan K et al, Corpus Analysis and Tagging for Tamil, Symposium on Translation Support Systems, I.I.T. Kanpur, 2002.
- 8. Rajan K et al, Computational Analysis of Tamil Text a Statistical Approach, Third National conference on Recent Trends in Advanced Computing, Thirunelveli, 2002.
- 9. Ganesan M, Compilation of Electronic Dictionary for Tamil, Tamil Internet 2000
- 10. James Allen, Natural Language Understanding, Benjamin/Cummings, 1995.