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.

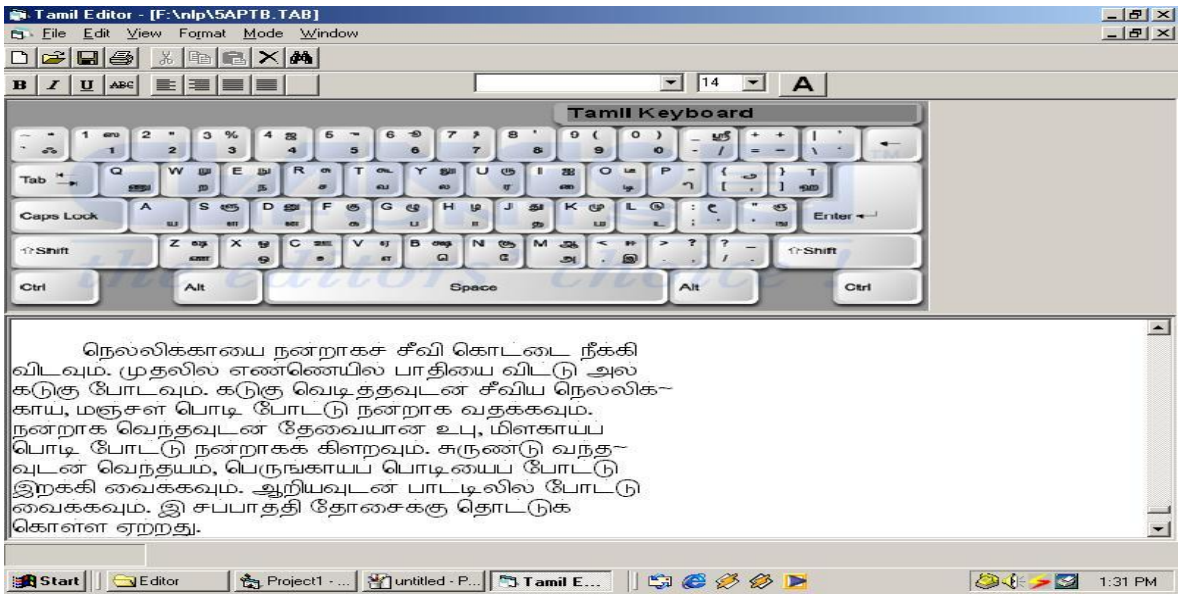


Fig.1 The layout of the Editor

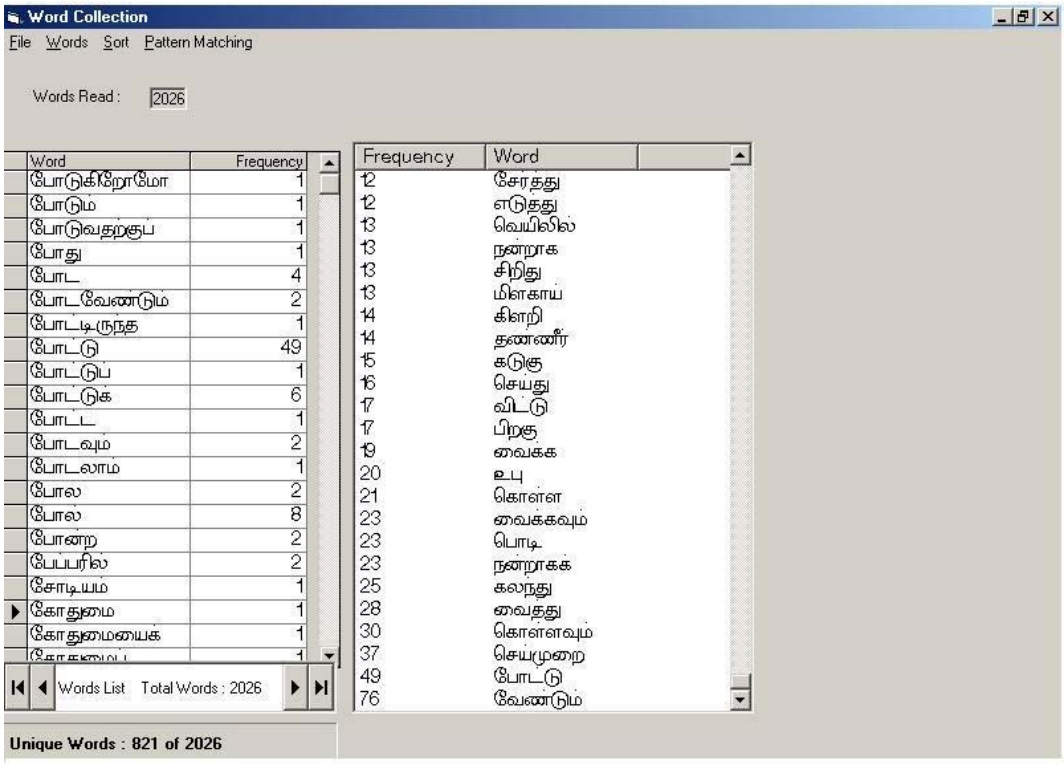
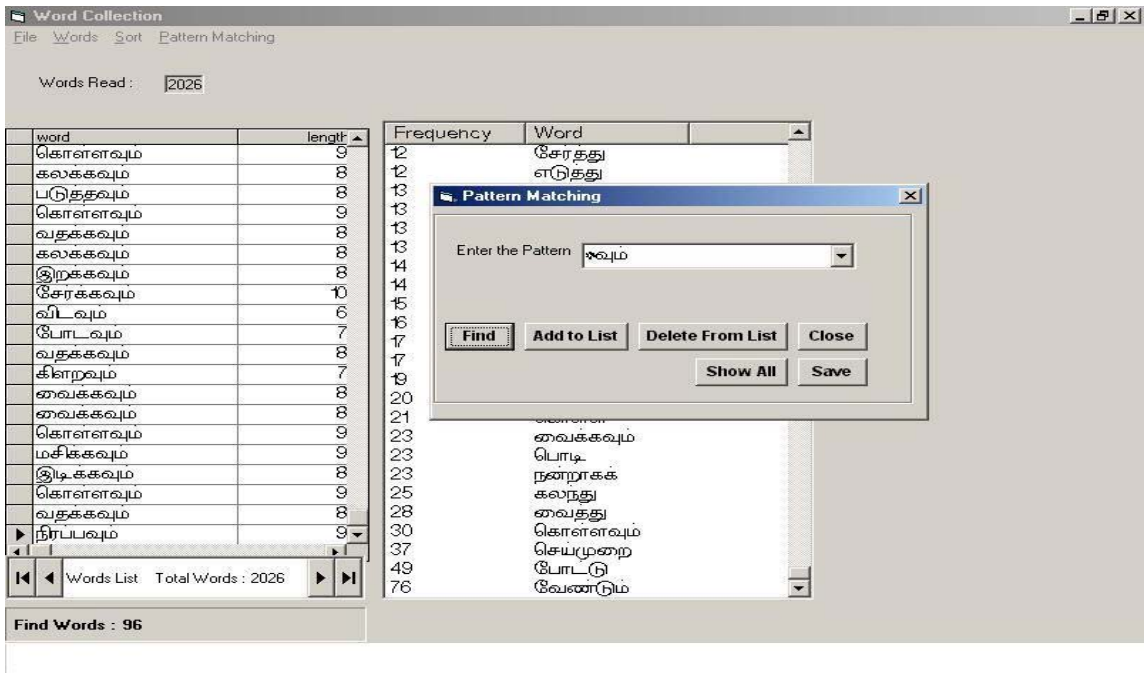


Fig. 2 Showing the word list with frequency



திig 3. Showing the Pattern Search

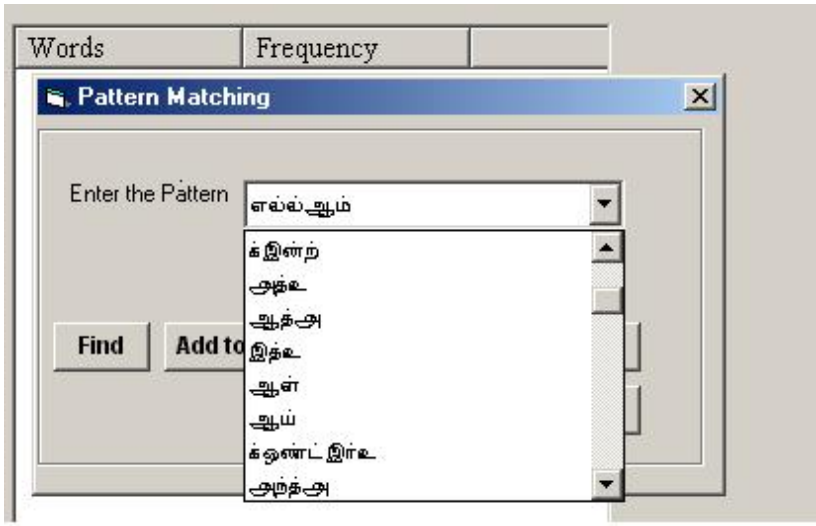


Fig. 4 Showing the Search list for easy entry of pattern (Words are in Consonant-Vowel form)

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.

மாநிலம்	பிரயாவிலேயே	தொழில்	துறையில்	மிகவும்	வளர்ச்சி
ஓன்றாகும்.	பதினெட்டாம்	நூறறாண்டின	இறுதியில	வெடித்த	பிரெஞ்ச
வரை	திரி	உயரநிலைப	பளவியில்	பயினறார.	எனன
வரலாறறுச	சகதி	எனற	முறையில	அரசுகேறறியது	நிலபபிரபுதுவதின
இலலாத	அரை	நிலபபிரபுதுவ	ஜெமனியில	ஒரு	முதலாளிதுவ–ஜனாராய
கடைசியிலும	நூறபதுகளின	தொடககதிலும	ஜெமனியில	பெருநதிரளான	மகளிடையில
மகளிடையில	அருபதிகள,	சமுக	வாழகையில	ஒரு	வேகம,
வரகம	மறறும	அறிவுப	பகுதியினரிடையில	பலவேறு	எதிரபபு
1842	ஏபரலில	அருது	பதவிகையில	வேலைகுச	சேரறனர.
6வது	ரைன	பிரதிநிதிகள	சபையில	நடைபெறற	விவாதம
இதர	கட்டுரைகளை	எழுதினார.	பதவிகையில	கிடைதத	அனுபவம
எழுசசி	பிரிடனில	1830களின	கடைசியில	ஏறபட்ட	பாட்டாளி
வரகளதுகும	பாட்டாளி	வரகளதுகும	இடையில	வரகள	போராட்டம
கடுமையான	தணிகளையை	ஏறபடுதியது	கடைசியில	1843	ஏபரல
மாரகஸ	Rheinische	Zeitung	பதவிகையில	பணியாறறிய	காலததில
காட்டினார.	ஃபாயரபாஹ்டமிருந்து	விடுதலசமான	முறையில	இயறகையை	மடடுமலலாமல
அவர	எழுதிய	புதுகதின	முனனுரையில	எடுதது்ரைகபபட்டன.	பகுதி10
மாறாக	அதை	விமரசன	முறையில	திருதியமைகும	பணியை
1843	அகடோபர	மாத	கடைசியில	மாரகஸ	பாரிகுச
புராதன,	கொசசைத	தனமான	முறையில	சமனாகும	கமயூனிசதை
கண்டுபிடிகவிலலை.	பதினெட்டாம	நூறறாண்டின	இறுதியில	ஏறபட்ட	பிரெஞ்ச
மாறுவதை	நிரணயிதது.	மாரகசின	வாழகையில	இருது	தீமானமான
("ஜெமன–பிரெஞ்ச	ஆணடுமலர")	எனற	சஞசிகையில	அவர	எழுதிய
பிரதிபலிகிறது	இசசஞசிகை	1844	பிபரவரியில	பாரிஸில	மாரகஸ
தலைபபில	மாரகஸ	எழுதிய	கடடுரையில	தெரிபப	பிரசினையைப
வளமும	மிகவும	குறிபபிடதகதாகும.	ஜெமனியில	மததையை	பறறிய
பொருளமுதலவாத	உலகக	கணணோடடதின	அடிபபடையில	வளரதது்ர.	பாட்டாளி
ஏஙகெலஸ	இஙகிலாநதுகுச	செலலும	வழியில	கொளளில	தஙகிய
1844	ஆகஸட	மாத	இறுதியில	ஏஙகெலஸ	பாரிசுகு

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.

```
அப் பம் <N>  {NP_Nom }
          தின் று <RP>

முயல் <N>  {NP_Subj}

அது<PN_3sn>  {NP_Subj}
ஒரு<Det>
மலைக் காடு.<N>  {NP_Nom }
_____________________________

ஒரு<Det>
பெரிய<Adj>
மலை.<N>  {NP_Nom }
_____________________________

அதன் <PN_3sn_Gen>
சரிவுகளில் <N_pl_Loc>  {NP_Loc }

பெரிய<Adj>
பெரிய<Adj>
மரங் கள் <N_Pl>  {NP_Nom }
          வளர் ந் து<V_P>
          காடாய் <Adv>
          மண் டி<V_P>
          யிரும் <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.