Tamil Corpus Generation and Text Analysis

M. Ganesan
Annamalai University, Annamalainagar, Tamilnadu, India

Introduction

Language can be studied from the point of view of language structure and language use. Study of language structure is called structural or formal linguistic study. Study of language use is called as functional linguistic study. Languages are described qualitatively in terms of grammatical units like nouns, verbs, noun phrases, verb phrases, subject, object, agent, goal, etc. to explain the structure of the language. Most of the grammars take a sentence as the minimum unit for the description of the structure. The structure has been studied from different view points and many linguistic theories have emerged to account the syntactic pattern of the sentence.

Languages can also be studied quantitatively in terms of frequency, place of occurrence, pattern of occurrence, etc. of various linguistic units in a text. The quantitative study basically needs a large quantum data and a mechanism to browse them fast. Now the computer technology facilitates to store and study a huge texts to the tune of hundreds of millions of words in few seconds or minutes. A new method of language study called corpus linguistics has emerged in recent years. Corpus is a large collection of written or spoken texts available in machine readable form accumulated in scientific way to represent a particular variety or use of a language. It serves as an authentic data for linguistic and other related studies. The size, text type, organization, accessing method, etc. are some of the basic features of a corpus which have to be carefully decided while generating a corpus. There are different types, which are again determined by the purpose for which the corpus is built.

In this article I share my experience in the generation of CIIL corpus and explain the scheme of POS tagging using morphological analysis. I also discuss the various tools that I have developed to analyze Tamil texts (Corpus Analysis Tools for Tamil) and their use in different applications.

Tamil Corpus Generation

The first corpus for modern written Tamil was built in the Central Institute of Indian Languages (CIIL), Mysore in 1987. The CIIL in collaboration with the Tokyo University of Foreign Studies, Japan built a corpus on Tamil textbooks. In 1991 under the scheme Technological Development for Indian Languages(TDIL), the Department of Electronics, Govt. of India launched a project called Development of Corpora of Texts of Indian Languages. The CIIL was entrusted to build Corpora for the four major Dravidian languages, where the present author worked as an investigator. Texts printed during the period 1981 to 1990 were selected to represent the modern Tamil. They are collected from 6 major categories, viz. Aesthetics (Literature and Fine arts), Social Sciences, Natural, Physical and Professional Sciences, Commerce, Administration and Technology, and Translated Texts. They are further classified into 76 minor categories, to cover various domains of language use. The size of the Tamil corpus is 3.6 million words. The major objectives of the project was to build corpora of not less than 3 million words, and to develop

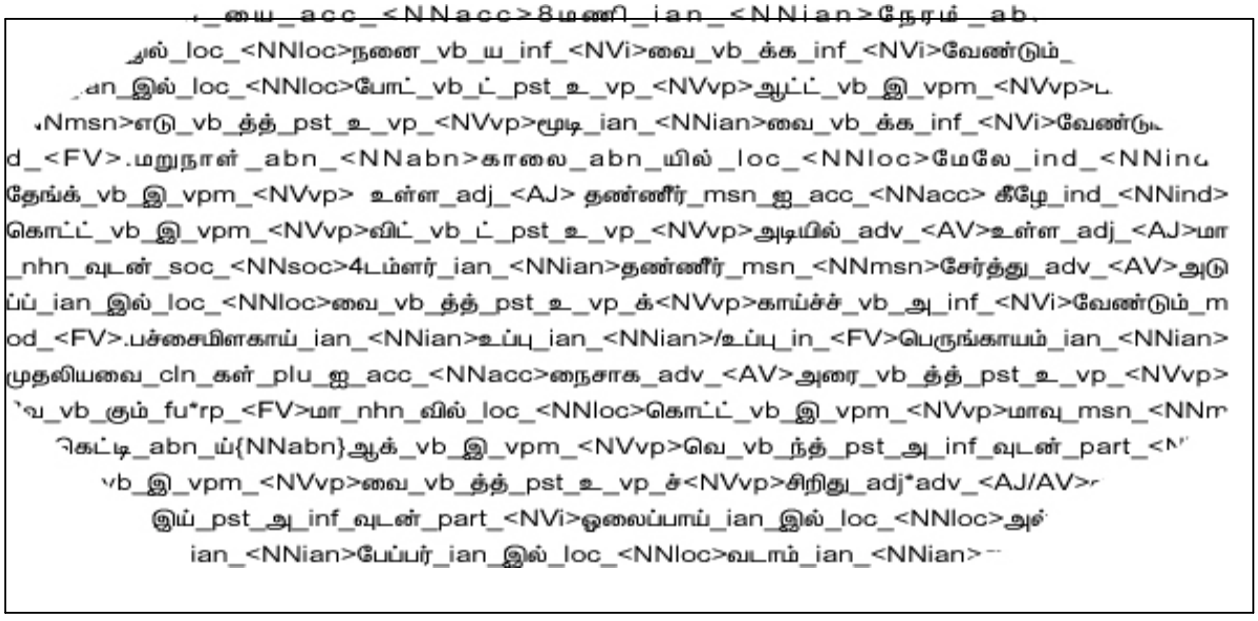
software for grammatical tagging (at word level), KWIC Concordance, and for corpus management. In 1993 a spoken Tamil corpus (1 lakh words) was generated on the transcribed spoken data collected by Eric Pederson, Netherlands. The Mozhi Truest, Chennai has build a corpus of around 3 million word for modern written Tamil. The CIIL is currently augmenting the corpora of Tamil texts and creating corpus for spoken Tamil.

Corpus Organization

The data for the CIIL corpus are collected from books, textbooks, magazines, newspapers and Government documents in order to represent the contemporary Tamil. The collected data are organized with the following information: 1) major category, 2) sub category, 3) title of the text, 4) author name, 5) source 6) publishers, 7)year of publication, and 8) page numbers. These information help the user to retrieve any data selectively from the corpus. Further the organization of data is needed for any addition or deletion of data from the corpus. A software called “corpus manager” does these jobs.

POS tagger

The collection of texts, called Raw Corpus can be provided with many additional information, particularly grammatical ones at different levels, viz. phonological, morphological, syntactic, semantic and discourse level. It is called annotating or tagging the corpus. The POS (Parts of speech) tagging is a popular and common type of annotation successfully implemented on a number of corpora in English and other European languages. The tagging can be achieved in the following four ways: 1) rule- based tagging, 2) statistics –based tagging, 3) pattern – based tagging, and 4) manual tagging. The first three are automatic tagging (with manual post-editing) and the last one, manual tagging is slow, labour intensive and liable to error and inconsistency(Leech, 1992:131). The present author has developed an automatic tagger called “Morph and POS tagger for Tamil”(Ganesan, 2007) which tags at morph and word level. At present the tagset has 82 tags at morph level and 22 at word level. A sample tagged text is given below.



Syntactic Tagger

A software for tagging at phrase and clause level has been developed for Tamil by the present author. The texts tagged at morpheme and word level will be the input for the syntactic tagging. Tamil is an agglutinative language; therefore its morphology is complex. But, its morphotactics is very tight. Therefore identifying the morphs is not very difficult. But, syntax of Tamil is very loose; because it is comparatively a free word ordered language. It is very difficult to identify the phrase boundary.

Tools for Text Analysis

The potentiality of corpus in language studies, both theoretical and applied, are enormous. Language description, testing of grammatical theories, natural language processing(NLP), language teaching, dictionary making, translation (both human and machine), style analysis, etc. are the major areas where corpus can throw lot of insights. To bring out all those information that are needed for these applications, a number of tools have to be developed. Frequency count (letter, syllable, word frequency), searching (for particular pattern, in particular context, at different levels), sorting (forward and reverse), indexing, concordance, KWIC, tag search, word list, lemma extraction, type/token ratio, etc. are some of the tools which can be used to analyze the corpus and to get a variety of quantitative information. Corpus Analysis Tools for Tamil(CATT) (Ganesan, 2007), a software provides a number facilities to extract different quantitative information from the corpora. Using the quantitative information language can be described from a new angle.

Frequency Count

The frequency of occurrence of a letter, syllable, morph, word, or a phrase, in a text can be counted using a software. For language like Tamil word frequency can be studied only after removing all the affixes from the stem. A Lemma Extractor (Ganesan, 2007) does this job and provides the list of roots in alphabetic order with frequency. For example, if one wants to study the words which are used in the primary school textbooks, using Lemma Extractor he can get them in no time. Such information facilitate to know what are the words introduced at what level, whether all the words that are intended to teach at primary level are there in the textbooks, etc. Similarly phrases and sentence types can be studied. This kind of study is practically not possible without proper computational tools. One can also study the frequency of different word forms. For example, in Tamil the verb forms Infinitive, Verbal participle and Relative participle are more frequent than the finite or imperative forms.

Verb	Total	Inf.	V.P.	R.P
col	428	56	29	33
kuuRu	1109	83	52	100
viLakku	197	34	19	11

These frequencies are from the text of 3lakh words. It clearly shows that while teaching Tamil these forms Infinitive, Verbal participle and Relative participle must be given priority and more attention than the finite forms. Another observation is that among the words *col* and *kuuRu* ‘to

say' the word *kuuRu* which is more literary, occurred more frequently than the word *col*. In another study the frequency of occurrence of letters are made from the data of one lakh words.

dot (on the consonants)	16.85%	
ka	07.94%	
vowel *u* after consonants	07.38%	
ta	06.73%	
vowel *i* after consonants	06.55%	etc.

Such study helps in designing the Keyboard layout for Tamil. The Keyboard based on the frequency study will be more scientific and faster to use.

KWIC Concordance

KWIC concordance is a list consists of a keyword in the middle and the contexts of 4 or 5 words on either side of the keyword. A sample is given below. KWIC concordance can be extracted

ுல என்றாகிவிடுமா என்ற <கேள்வி> சிந்திக்க தக்கதாகும் கரு
ுக மந்தமான குணங்கள் கல்வி <கேள்வி> ஞானம் கணிதம் தேவாலய
ர் திரு கோ சி மணி இதுவும் மூல <கேள்வி> தான் மாண்புமிகு பேரவை தலை
ண்டும் என்று கேட்பார் ரொம்ப நியாயமான <கேள்வி> நாம் எதை படிக்கவேண்டும் எதை படி
ன் திரு சா பீட்டர் ஆல்போன்ஸ் இந்த <கேள்வி> நீ அடிக்கிறது போல் அடி நான் அழுவது
ப்பாட்டிற்கு நான் தருகிற பதில் அதற்கான <கேள்வி> பட்டியல் எல்லாம் தயாராகி விட்டது
அச்சினைகள் குறித்து குவஸ்டினர் அதாவது <கேள்வி> பட்டியல் தயாரித்து பல்வேறு குழுக்கள் பொது
அதிஹரர்கள் இராசபுத்திரர்களை பற்றி <கேள்வி> பட்டிருக்கிறார்களா பிரதிஹரர்கள்
து உயர்படுத்த முயற்சிகள் செய்வதாகவும் <கேள்வி> பட்டேன் அவரை பார்க்க ஆசை உண்டாயிற்று
ஆர் சுவாமிநாதன் இந்த 98 சந்திப்பு <கேள்வி> பதில் உருவில் இல்லாமல் ஒரு கட்டுரையாகவே
விக்கு என்னிடம் இப்போது பதில் இல்லை தனி <கேள்வி> போட கேட்டு கொள்கிறேன் திரு இரெ
ள் என்னிடம் இல்லை அதற்கு பிரத்தியேகமான <கேள்வி> போட வேண்டும் இருந்தாலும் அவை
ஏள்வியாக இப்போது கேட்க படுகிறது தனி <கேள்வி> போடவும் திரு க ஆறுமுகம்
ர்களே ஏதாவது ஒரு பிரச்சினை என்றால் ஒரு <கேள்வி> வடிவத்தில் இரண்டொரு வாக்கியங்களில்
கேட்டால் பதில் அளிக்க இயலாது இதை ஒரு <கேள்வி> வடிவத்திலே நம்முடைய மாண்புமிகு உறுப்பினர்
யே சாரும் கல்வி பயிற்சி இல்லாதவருக்கு <கேள்வி> வாயிலாக அறிவு புகட்டுவதற்காகவே இக்கலை
ஆராட்டி இருப்பாரா இதே கபிலரை வெறுத்த <கேள்வி> விளங்கு புகழ் கபிலன் எனவும்
ஐ நோயாளி நன்றாக துங்குவார் நான் கேட்கும் <கேள்விகட்கு> பதில் சொல்லுவார் நல்ல பேச்சாளியாக
ங்கு விளக்கியாக வேண்டும் இது தொடர்பாக <கேள்விகள்> எழுகின்றன 1 சட்டசபையில்
ல் ஏற்படுகிறது என்றறிகிறோம் இங்கே சில <கேள்விகள்> எழுகின்றன செழுமையோ வீழ்ச்சியோ
பப்பட்டு இருக்கிறது விரைவிலே அத்தகைய <கேள்விகள்> ஏடுகள் மூலமாக 463 வெளிவரும்
டுத்து அவர்களிடம் வினாப்பட்டியலிலுள்ள <கேள்விகள்> ஒவ்வொன்றாக வாசிக்கப்பட்டு விடைகள்
உம் பொருள்கள் போன்றவற்றை பற்றி சிறு சிறு <கேள்விகள்> கேட்கவேண்டும் ஒருவேளை அவள்
பார்த்து பேசுகின்றன பாத்திரங்களை இவன் <கேள்விகள்> கேட்கிறான் இவன் கேள்விகளில்
இயில் உ அணிந்திருக்கும் இந்த மாணவரை சில <கேள்விகள்> கேட்டோம் பட்டம் பெற நான்கு
அக்கது கதையின் நடுவே சபையோரை பார்த்து <கேள்விகள்> கேட்பது சிறந்த உத்தியாக கருதப்படுகின்றது
ய நூல்கள் இருக்கின்றனவா இது போன்ற <கேள்விகள்> சாதாரணமாக எளிதில் பதில் அளிக்க
ல்லாம் எவ்வாறு நோய்களை நீக்கும் என்னும் <கேள்விகள்> பதில் சொல்ல முடியாதவைதான்
ய்வீக வழிபாடு ஆசார சீலத்துடன் கல்வி <கேள்விகளில்> நாட்டம் சாஸ்திர ஆராய்ச்சி ச
ிரங்களை இவன் கேள்விகள் கேட்கிறான் இவன் <கேள்விகளில்> பாத்திரங்கள் பங்குபெறுதல் என்பது
கள் ஆகியவற்றை காண்கிறோம் ஆனால் கல்வி <கேள்விகளில்> மட்டும் தேர்ச்சி அறியவில்லை ஆகவே
ன்படுத்தி கொள்ள ஏதுவாக இருக்கும் மேலும் <கேள்விகளின்> தன்மைகளை ஆராய்ந்தால் செய்திகளை
ட்டப்பட்டனர் அரசு தரப்பு பெரு கேட்ட <கேள்விகளுக்கு> அனைவரின் சார்பிலும் பேசினார்
நூலகரின் அறிவு கூர்மையாலும் கடினமான <கேள்விகளுக்கு> கூட எளிதில் சேவை செய்து விட
ுக்காது இந்த நூல்கள் சில குறிப்பிட்ட <கேள்விகளுக்கு> பதில் அளிப்பதற்காகவே
திரும்பிவிடுகிறதா இப்போது சில <கேள்விகளுக்கு> பதில் தருக 1 சி
வும் பகத்சிங் எழுந்து நின்று சில <கேள்விகளுக்கு> பதிலளித்தார்
குமூலம் ஏம் அளிக்கவில்லை <கேள்விகளுக்கு> மட்டும் பதி

from corpus for any word, part of a word, suffix, infix, prefix or even a phrase. Sorting can be done on the keyword or on the previous word or following word. It is more useful in identifying the different meanings of a polysemous word, lexical association, collocation properties of lexical items, etc. In dictionary compilation the KWIC concordance facilitate the lexicographer to find out the various meanings of a word, subject area, registers, idiomatic usages, etc.

Word List

This tool makes a word list for all the words in a selected texts or a corpus. The words will be sorted and presented with frequency. Type / token ratio will also be given for the total word. Words with a frequency or more than particular frequency or less than a frequency can be listed separately. All the outcomes can be stored as a separate file. It helps to find out the various words found in a text with their frequency.

Word-part Search

Like the word list, here the words are sorted from the end of the word. All the same suffixes come together and therefore it helps to study the inflectional and derivational properties of different words and affixes. Using this tool a reverse dictionary for Tamil can made.

Multi-Conditional Search

A string with multi condition can be searched with this tool. If a string is given as an infix for search, the other conditions like prefixes and suffixes can also be given. All the words in the corpus satisfying all the conditions will alone be listed. For example, all the Finite verb with present tense marker can be extracted from the corpus.

Structure Search

Here the data must be a POS tagged texts. Pattern in terms of word-tags can searched. All the sentences matching the pattern will be extracted and listed. There are two options: pattern as a sentence and pattern available anywhere in a sentence. It helps to find out the usage of various phrases, clauses, particular pattern of word association, etc.

Tag Search

This tool also works on a tagged corpus. Words inflexed / derived to particular forms can extracted using this tool. First Word level tag information must be supplied, then morph level tag information in the next window. The tags may be one or more than one. There are three options: 1) include 2) exclude and 3) stem extraction. The first option lists all the words matching the word level tag and morph level tag. The second extracts all the words matching the word level tag, but excluding the morph level tag. The third option removes all the affixes and lists the stem portions alone in sorted order. It is useful to list for example, all the verb forms conjugated to particular forms or other than a particular forms.

Conclusion

The Quantitative Analysis is getting importance on par with Qualitative analysis. Language use is given priority for the description of the language. Various quantitative information extracted from the corpus provide new insights on language structure and are useful for textbook preparation, dictionary compilation, machine learning, Machine Translations, etc. The size of corpus at present available for Tamil is very small. Sometimes, many words used in day-to-day context have not attested in the corpus. Therefore there is a need to increase the size of the corpus to a minimum of 200 million words.

References

1. Ekka. Francis, B. D. Jayaram and Ganesan "Final Report Development of Corpora of Texts of Indian Languages" in Machine Readable Form, Part II (Tamil, Telugu, Kannada, Malayalam) Mysore, CIIL, 1995.
2. Ganesan, M. "A Scheme for Grammatical Tagging of Corpora in Indian Languages" in **Technology and Languages,** (Ed) BB Rajapurohit, Mysore, CIIL, 1994.
3. Ganesan, M. 'Corpus Analysis Tools for Tamil (CATT) (Software) Annamalai University, Annamalai Nagar, 2007.
4. Ganesan, M. 'Morph and POS Tagger for Tamil' (Software) Annamalai University, Annamalai Nagar, 2007.
5. Leech, Geoffery and Steven Fligelstone "Computers and Corpus Analysis" in **Computers and Written Texts** (Ed) Christopher S. Buller, Oxford: Basil Blackwell Ltd, 1992.
6. Leech Geoffery, "Corpora Annotation Schemes" in **Literary and Linguistic** Computing, Vol 8 No4 1993.
7. Shanno, C and Weaver, W. **The Mathemetical Theory of Communication,** UIP: Illinois, 1949.
8. Stubbs. Micheal, **Tect and Corpus Analysis** Oxford: Blackwell Publishers Ltd, 1996.