DEVELOPING ONLINE CORPUS AND CONCORDANCE TO
AINKURUNURU OF SANGAM TAMIL LITERATURE

Dr. A. KAMATCHI
CAS in Linguistics, Annamalai University

Introduction

Corpus Linguistics, a new method of language study, which has emerged in recent years, has generated a number of research methods, attempting to trace a path from data to theory. As far as Tamil language is concerned, the first corpus for modern written Tamil was built in the Central Institute of Indian Languages (CIIL), Mysore in 1987. But its usage by the academic community is very less in number. The reason may be that it is only in the CD form but not posted in the website. The other one which is now available in the internet is the Cre-A: Online Tamil Language Repository posted by Cre-A. This corpus is, of course, concerned only with the modern Tamil, but not with the other periods of Tamil language. Besides these corpora of living languages in the world, computerized corpora for ancient languages have also been made in the form of collection of texts. An example is the Quranic Arabic Corpus, an annotated corpus for the Classical Arabic language of the Quran. Another one is the Andersen-Forbes database of the Hebrew Bible, developed since the 1970s, as reported by wikipedia. On these lines, the present study aims to develop online Corpus and Concordance to ainkuRunuuRu of Sangam Tamil Literature, through which one can search lexical items, rather than words, and their concordance available in this text, which are dated to 2\textsuperscript{nd} century A.D.

Corpus and Concordance

The present study envisages developing software for Corpus and Concordance to AinkuRumuuRu. The data would be collected from AinkuRumuuRu, one of the old Tamil texts. In fact, it is the pioneer attempt to develop the Corpus and Concordance to AinkuRumuuRu, which consists of not only lexical items but also the grammatical elements attested in this text. Approximately, two thousand four hundred words are, perhaps, attested in this literature and roughly, more than two fifty grammatical categories are also identified and incorporated in this programme. Every line has been parsed morphologically – both root forms and grammatical entities. All the parsed lexical items are systematically collected and incorporated in this corpus. When we click one word in the scroll bar in the window, we can get a number of occurrences of that word, the line on which it occurs, its meaning in that particular line, literature name and poem number.

Collection of Materials:

The data have been collected from the original texts of AinkuRumuuRu literature, using the Index works, which are available in various institutions in the nation as well as the commentaries for the literature. Using the computer software such as POS tagging and parsing methods and searching engines, the data were collected from the AinkuRumuuRu of Sangam texts and analyzed in the way of descriptive methods. The works of the eminent Tamil Linguistics Scholars such as Prof. Nida, Prof. V.I. Subramoniam, Prof. S. Agesthiailingom, Prof. M. Israel, Prof. M. Elayapermal, Prof. A. Kamatchi, Prof. Rm. Sundaram Prof. S.N. Kandaswamy, Ms. Eva Wilden and so on have mostly been utilized for this study.
Significance of the Study

There are indexes for some of the Sangam works, but not for all, of course. Among these Sangam texts, only for some of the anthologies, the indexes have been made by the scholars and published. Though the others have been indexed, they have not published but are only in the form of unpublished Ph. D. Dissertations at various Universities/Institutions. Of course, these indexes comprise the words with root form plus some other grammatical suffixes, as far as the verbal conjunction is concerned although the noun forms are separately dealt with. Further, it includes their occurrences only with poem and line number. In case someone wants to refer something, he needs of original text for the detailed reference. When someone tries to find a word which is doubtful for the researcher working in Sangam Tamil, it may be so difficult for a scholar to locate such word in Sangam Tamil. For such situations, the present work is very helpful for the researcher or reader to find out such things in this area.

Interdisciplinary Relevance

Ultimately, it is highly expected that once this work would be completed, it will, undoubtedly, enable us to compare inter languages and intra language and language families in the world. Further, it could be very useful to work in the area of the comparative as well as the historical linguistics study. The quantitative and qualitative analyses would also be made easy through such a corpus work.

Model for Other Texts

The completion of this work definitely leads to develop software for Sangam Tamil of classical Tamil texts all. If such a work would be completed, then we need not want of such textual materials. As we all know, accuracy is one of the unique features for computer. Through this programme, one can attain the total number of occurrences in this literature of Sangam text. Of course, it is the user friendly material, of course. Therefore, if it will be successfully completed, then this work will be the model for other texts in old Tamil to develop the software for Sangam Tamil, as already pointed out.

Wide range of Applications

A maximum number of words in Sangam Tamil, as we all know, occupy the head entries in Dravidian Etymological Dictionary (DED), which was prepared five decades ago and was widely used by the scholars worldwide. So there is a proposal for posting this lexical material in the website for the use of the scholars working in the area of Comparative Linguistics in general and Comparative Dravidian in particular. In the same way, it is also very useful for the Historical Linguistics scholar in the world. Moreover, it may be also used to Glottochronological Linguistics study all over the world. The study, Sangam Corpus and Concordance, would, it is sure, represent the classic language in Tamil. After Tolkaappiyam, in which a few words have been simply explained on the part uriccol, this would be the potential work with using the modern theories in linguistics and scientific methods in the process of preparing the collection of lexical items.

While selecting the word aRu, specifying the meaning of ‘to batter, six, without’ from the drop-down box of select word, we can get the following display on the screen, as shown below:
From this display, we can get the word selected, number of occurrences in the text, poem number and line at which the word occurs. In the same way, when selecting the grammatical category such as the grammatical item a2, representing nominal participial marker, from the drop-down box of select grammar, we can get the following display with its representation and its total number of occurrence.
In addition to these, we can also search any word and/or grammatical entities existed in the text of *AińkuRunuuRu*. In the following display, the word *akavu*, which means ‘to crow’, is typed in the drop-down box of *search any word / grammar* and gotten its total occurrence from the text of *AińkuRunuuRu*.

### SANGAM CORPUS AND CONCORDANCE

**Dr. A. KAMATCHI**  
Assistant Professor  
CAS in Linguistics, Annamalai University

Instruction: Please select a literature before search words/grammar.

**Selected Keyword:** அகவு – to crow  
No of Occurrences: 2

Select Literature to search:

```plaintext
?? ??
```

Select Word:

```plaintext
?? ??- to crow
```

Select Grammar:

```plaintext
1-?? ??? ?? 6/ ?? ? ?? ?????
```

Search Any Word / Grammar:

```plaintext
?? ??
```

Search

### Conclusion

As we all know, accuracy is one of the unique features of computer. Through this programme, one can attain the total number of occurrences in this pioneer work. Of course, it is the user friendly material, of course. One can easily find a word which is doubtful for the researcher working in Sangam Tamil. It may be so easy for the scholar to locate such word in Sangam Tamil. Further, such word that occurs in one text may or may not available in the other texts. For such situations, it is very helpful for the researcher or reader to find out such things in this area.

### Bibliography

Agesthialingom, S. 1979. *A Grammar of Old Tamil (With special Reference to patiRRuppattu)* Annamalai University, Annamalainagar.


Subramonyam, V.I. 1962. *Index of puRanaanuuRu*, University of Kerala, Trivandrum.