

# Open-Tamil text processing tools

A Muthiah<sup>1</sup>, T Shrinivasan<sup>2</sup>, M Annamalai<sup>3</sup>

<sup>1</sup> Boston, USA, <sup>2,3</sup> Chennai, India

## Abstract:

Programmers face common problems while developing Tamil applications. We discuss a suite of open-source tools called Open-Tami [1-4], which provides solutions to commonly encountered problems in Tamil computing - code-point to letter mapping, Tamil-word-length calculation, sorting order, Tamil input methods (IME) for web-based applications etc. Future plans for Open-Tamil development, licensing and algorithms involved are explained in this article. It is already used as part of production websites [5].

## Introduction

Tamil word processing is very easy on a modern computer with the processing speeds, and available memory. However the variety of encoding formats used in legacy and modern systems like TSCII, TAM, TAB and modern Unicode (with UTF-8, UTF-16) formats makes it a complex space to navigate for the uninitiated. To address this problem we have developed a heterogeneous tool collection in Open-Tamil project [1], also published as a Python package [2].

## Goals

Goal of this package is to collect and develop open-source licensed Tamil tools, in one location that provide the following,

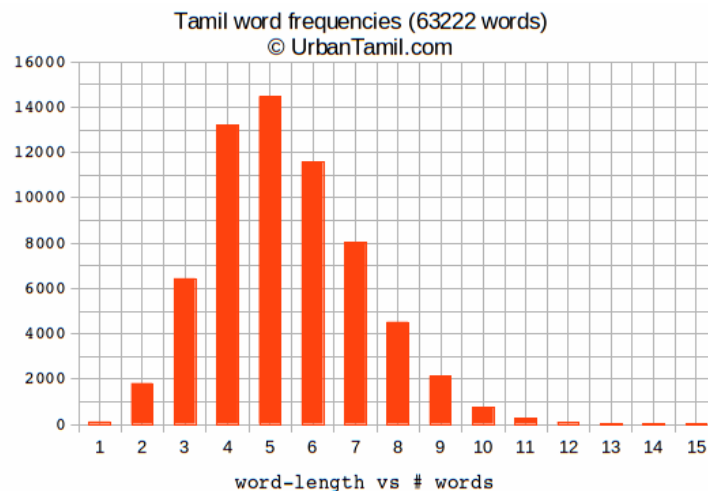
1. Unicode standard tools for Tamil - provide various tools for Tamil Unicode development. Currently TSCII, UTF8 encoding tables are provided. Converters for TAB, TAM, and other encoding formats are planned to be added [4].

2. Access Unicode Tamil letters, vowels and consonants. Breakdown Tamil glyphs and Unicode code-points into Tamil letter representations – collation. Calculate Tamil letter lengths of word, and define a sorting order for Tamil.
3. Tools for navigating a corpus of data, build word frequency, prediction tables etc.
4. Provide modern, unit-tested software library with open licensing
5. Consolidate variously available Tamil open-source software packages

We plan to host this package as heterogeneous source, language agnostic fashion. As of current release version 0.2.4 we have implemented many parts of goals 1-4.

## Examples

Open Source Tamil Tools allows you to easily carry out these operations; for example the (Python) code snippet calculates the word-frequency of a chunk of text, and (in a modified form) the word-length frequency of a free Tamil dictionary [3,5].



```
import re, operator
import tamil #open-tamil library
```

```
def print_tamil_words( tatest ):
    taletters = tamil.utf8.get_letters(tatest)
    # tamil words only
    frequency = {}
    for pos,word in enumerate(tamil.utf8.get_tamil_words(taletters)):
        print pos, word
        frequency[word] = 1 + frequency.get(word,0)
    # sort words by descending order of occurrence
    for l in sorted(frequency.iteritems(), key=operator.itemgetter(1)):
        print l[0],':',l[1]
```

Plans are in place to add various encoding converters using knowledge of font-map tables [4].

### **Current Users**

Unsurprisingly the open-tamil package is used by the author in two production websites for Tamil programming language, Ezhil, [5], and the open social Tamil dictionary, UrbanTamil, [6]. Ezhil language website uses the open-tamil for text processing in the UTF-8.

UrbanTamil website relies heavily on UTF8 processing, database search and content validation using open-tamil library. Since the package is installed via the PIP (Python Package index) we have over 1000 downloads [7].

### **Development & Testing**

Open-Tamil library is developed by a team of volunteers by sharing code on GitHub. This library has unit tests and uses the Travis-CI continuous integration system for regression proof development, making Open-Tamil a modern software project [8].

Currently we have the following components,

1. Python 'tamil' package as part of open-tamil
  - a. Map Unicode code-points to Tamil letters; basic but important parsing - in a routine called `get_letters` from a Tamil word
  - b. Work with vowels (உயிர்) and consonants (மெய்), compound conjugates (உயிர்மெய்)
  - c. Reverse letters in Tamil word
2. Transliterate package
  - a. We support 3 transliteration modes
  - b. Azhagi - phonetic maps for all Tamil letters - many -> one supporting multiple form inputs
  - c. Jaffna Library - phonetic maps for all Tamil letters - one->one
  - d. Combinational layout - based on phonetic mapping of vowel+consonant
3. On-screen keyboard
  - a. We provide tamil99 layout for Mottie keyboard jQuery plugin [9] for web deployments. This is used in UrbanTamil website [6].
4. Language models
  - a. Basic support for letter unigram, bigram models using UTF-8 based corpora are supported in the package 'ngram/' which supports unigram model at the moment. More complex language models are expected to be developed soon.
5. Examples
  - a. Open-Tamil is a set of Python libraries which can help your application - web, system software, GUI on desktop etc. support Tamil text processing, inputs. Examples illustrate things like encoding conversion from TSCII to UTF-8, and other text processing.
6. Unit tests

- a. Many of library functions have unit tests and 70%-90% coverage; each commit of source code is run on a continuous integration server [8].

## Conclusions

Open-Tamil is an effort to bring an open source Tamil text processing programming library for software engineers and web developers. Currently we follow best practices and provide a first-class library for development. We are a volunteer effort, and accept code contributions, and idea inputs with constant effort to improve the library.

## Acknowledgements

We like to thank, Sathia Narayanan (San Jose, CA), for constructive comments and enthusiastic discussions on Open-Tamil libraries.

## References:

1. M. Annamalai, Open-Tamil source code base, <http://bit.ly/1iTTJ5V>,
2. Python package for Open-Tamil, <https://pypi.python.org/pypi/Open-Tamil/0.2.2-devel>
3. M. Annamalai, "UrbanTamil.com + open-tamil = Tamil Vocabulary", <http://ezhillang.wordpress.com/2014/04/29/urbantamil-com-open-tamil-tamil-vocabulary/>
4. T. Shrinivasan, "How to get Character Map of a TTF font," <http://goinggnu.wordpress.com/2014/06/26/how-to-get-character-map-of-a-ttf-font/> (2014).
5. M. Annamalai, EzhilLang.org – Tamil programming language
6. M. Annamalai, UrbanTamil.com – Social Tamil Dictionary
7. Open-Tamil Python Package index - <https://pypi.python.org/pypi/Open-Tamil/0.2.4>
8. Travis CI – continuous integration systems, [www.Travis-CI.org](http://www.Travis-CI.org)
9. Tamil99 layout for Mottie keyboard jQuery-UI plugin, <http://mottie.github.io/Keyboard/>