Developments in Open-Tamil Library

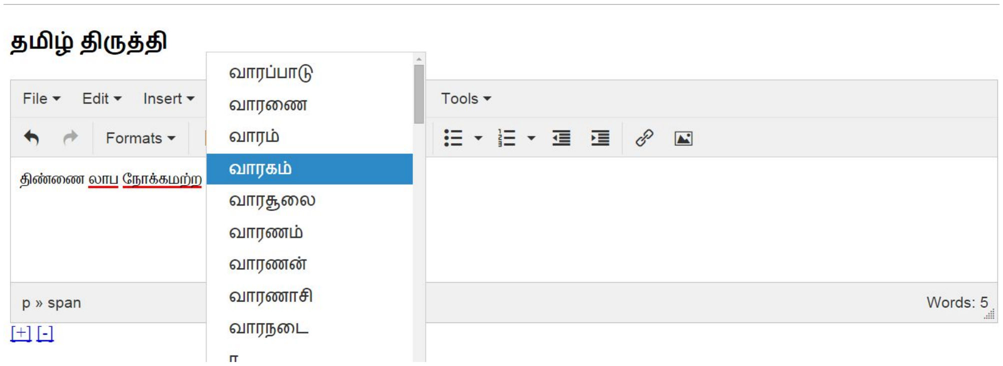
T. Arulalan, T. Shrinivasan, and A. Muthiah

Open-tamil library is designed to help software developers create high-level applications in Tamil; it is a freely available package for Python 2 (and Python 3K) to process Tamil text, and extended to many new applications since the original announcement in 2014 [1]. Open-tamil can be considered as a SDK for commercial, public or private software development in the Tamil language.

To recall [1], the main motivation of the open-tamil project is to provide enough foundational software pieces, sort of a libc for libtamil, so that developing software for Tamil applications becomes easy. Further, such a library will enable several Tamil applications to be created and make an easier ecosystem. While previously the software developers interested in Tamil language coding should have to familiarize themselves with details of encoding standards and conventions with open-tamil library they are standardized and commonly used converters are available.

Most recent version of Open-Tamil is v0.5, included contributions from several developers including the authoring team, and few others. In all over 8 developers have contributed their codes to open-tamil.

Here we report additional examples, bug fixes and enhanced API availability for Tamil text processing. We have performed data analysis using Tamil Wikipedia [2], performed various corpus linguistics applications like spell-checking, built word games like crosswords and word search grids in Tamil (Fig. 1, 2).

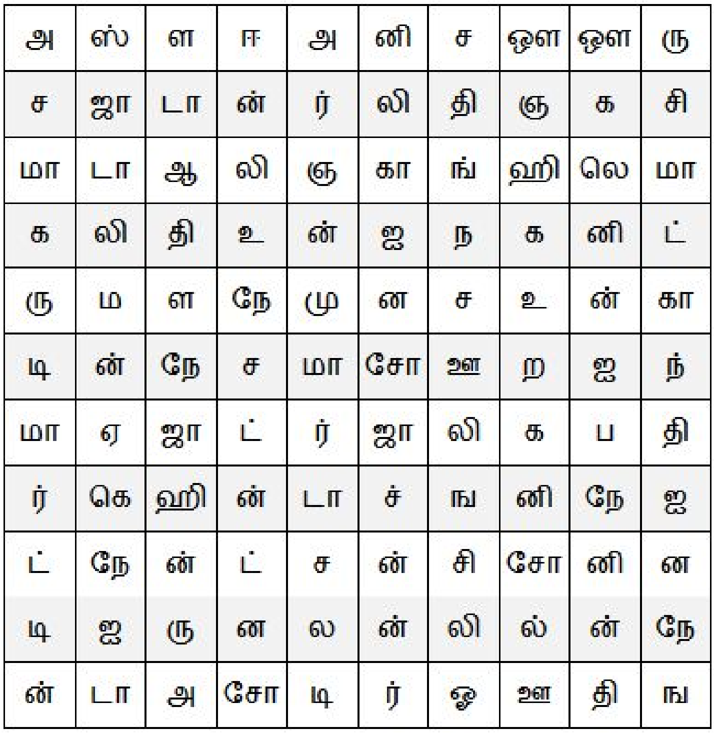


தமிழ் திருத்தி

திண்ணை லாப நோக்கமற்ற

Fig. 1. Integration of solthiruthi spell-checker from open-tamil with TinyMCE web editor

Other examples also include a speech synthesizer for Tamil numerals and development of Tamil speaking clock, generating crossword games (word search) etc. Table below shows word search for word leader names.



அ	ஸ்	ள	ஈ	அ	ளி	ச	ஔ	ஔ	ரு
ச	ஜா	டா	ன்	ர்	லி	தி	ஞ	க	சி
மா	டா	ஆ	லி	ஞ	கா	ங்	ஹி	லெ	மா
க	லி	தி	உ	ன்	ஜ	ந	க	னி	ட்
ரு	ம	ள	நே	மு	ன	ச	உ	ன்	கா
டி	ன்	நே	ச	மா	சோ	ஊ	ற	ஐ	ந்
மா	ஏ	ஜா	ட்	ர்	ஜா	லி	க	ப	தி
ர்	கெ	ஹி	ன்	டா	ச்	ங	னி	நே	ஐ
ட்	நே	ன்	ட்	ச	ன்	சி	சோ	னி	ன
டி	ஐ	ரு	ன	ல	ன்	லி	ல்	ன்	நே
ன்	டா	அ	சோ	டி	ர்	ஓ	ஊ	தி	ங

Fig. 2: Word search game generated using open-tamil tools; names of prominent world leaders are shown in the grid.

Solthiruthi Features

Solthiruthi is a new Python package added to open-tamil for exploring corpus linguistics and word/spelling related API for streaming algorithms on canonical Tamil data, independent of encodings. We performed data analysis using Tamil Wikipedia, and illustrate various computational linguistics applications. This API provides the following abilities,

1. Word models (corpus linguistics tools) word frequency, unigram, bigram frequency tables
2. Word puzzle creation, Crossword detail,
3. build tries and fast word lookup data structures
4. built-in Tamil, English word list
5. canonical sorting for Tamil words (in Java and Python)

Norvig Spell Checker

A Norving algorithm [3] based spell-checker was added to Solthiruthi as part of the multi-part spell checking effort for Tamil. This spell-checker API provides a data-driven blind spell checker which does not depend on rules but on big-data instead. This spell checker prototype was integrated with a web-based editor (Fig. 1).

Example – Word Substitution game

It is somewhat easy to build a game in open-tamil for solving the word-hop transformation. This is a word game for converting a given word, say 'BAT', to another word, say 'DOG', by changing one letter at a time. Change is either adding, deleting or substituting letters in the word so that we are always using valid words (belonging to dictionary). While it is easy to conceive and code this game in many ways, open-tamil allows you the ability to avoid UTF-8 processing, and developing your custom dictionary representation for canonical dictionary from open-tamil, so you can come up with a solution like

e.g.

BAT -> BOT -> BOG -> DOG.

Build an edit-distance based 1-hop edge connecting words which are nodes in the graph. In this graph data structure the particular nodes that are connected by a path will have a transformation like the one shown above.

Quality

Open-Tamil project is developed via www.github.com with a battery of 208 *unittests* (2,705 LOC) that test our source code modules **tamil** (13,579 LOC), **solthiruthi** (1,594 LOC), and **ngram** (187 LOC), in the latest development repository.

All source code checkin on github trigger the continuous integration tests via Travis-CI on all our supported Python flavors (v2.6, v2.7, v3.3, v3.5, and PyPy) as well as manual testing of Java and Ruby tests. We use github issues and bug-reports to resolve any discovered issues quickly.

Future developments

Future developments include completing the Solthiruthi package for spell checking and corpus analysis of Tamil corpora, and interactive word games and puzzles. The Java Open-Tamil package will be further enhanced to provide more facilities for Android developers, like solthiruthi, currently only available to Python users of Open-Tamil.

Challenges faced during this work include poor adoption by Tamil software development community. Current adoption and knowledge of this suite may benefit from better documentation and community engagement.

Conclusion

Open-Tamil has the distinction of being available for Python 2, and 3. It is continuously tested, and released via the Python package index, and most recently extended for Android developers via Java package, and Ruby package. As a community developed effort, and due to proximity of the various Indian languages, we believe Open-Tamil can form a prototype open-source toolbox for other Indian languages.

Python users include www.urbantamil.com, www.ezhillang.org and Ezhil programming language and web based applications using Django or Flask frameworks [4]. Java users are mostly targeting Android platform.

In conclusion we also review various contributors, present developments, and suitability for additional contributions or sponsorship. Current level of private funding and contributions may delay our progress to achieving goal of open-tamil software tools for all; long term sustainability of this project will be better served by increased community funding and software usage.

References

1. Ref: "Open Tamil Text Processing Tools," M. Annamalai, T. Shrinivasan, M. Annamalai, (INFITT-2014), Puducherry, India.

2. Ref: Tamil Wikipedia analysis article at blog http://ezhillang.wordpress.com

3. Ref: Peter Norving, The Norving Algorithm at his blog.

4. Django web framework www.djangoproject.com