Growth and Evolution of Open-Tamil

Syed Abuthahir, T. Arulalan, Sathia Narayanan, Surendhar Ravichandran,
A. Arunram, T. Shrinivasan and Muthiah Annamalai
[corresponding email: ezhillang@gmail.com]

அமர்வு: (அ) தமிழ்மொழியில் கணினி பயன்பாடு மற்றும் தொழில்நுட்பம்
பிரிவுகள்: தமிழில் கணினி மற்றும் தகவல் தொழில்நுட்பம்; எழுத்துப்பிழை சரிபார்ப்பு; தமிழ்க் கணினிமொழியியல்

Abstract: We present the developments, additional capabilities gained, in open-tamil project, and highlight the new web-interface at http://tamilpesu.us to access few of language processing tools in this project. We spell out some various features of the TamilPesu.us portal, software deliverables of open-tamil project. We also touch upon some of the challenges encountered by open-tamil project.

Introduction

Open-Tamil project has grown from inception [1] into a multi-language, multi-domain project for promoting openly available tools in Tamil text and information processing. In this paper we write about code-growth, directions taken by the project, clients using open-tamil, as well as the future application areas we expect to grow into, like machine learning. Our current release is v0.7 in all previously supported platforms [2].

Developments

In terms of code-growth we have seen ~ 800 files changed, ~ 760,000 line insertion, and ~ 3000 deletions in this period with a total of ~ 24,000 lines of Python code in project ; a new web-interface to the open-tamil library functions, originally identified in our work [3], was independently developed on Django platform and now showcased the library functionality at http://tamilpesu.us; to help developer productivity we also added Sphinx documentation for the various modules at http://tamilpesu.us/apidoc/

This past year we added the following developments:

1. We added new members to our team and provided direct commit access to repository;
2. Our new contributors enabled Sandhi checker [4] and new web interface based on Django for open-tamil hosted at http://tamilpesu.us
3. As of March 31st, we have a total of 22,595 lines of code on this revision.2018 [Git HEAD=ca6a8e19...] version = 0.70 of open-tamil. Details are shown per module in Table-1.

Module	Lines of code	Classes	Functions
tamil	13666 loc in 22 files	10	175
solthiruthi	1465 loc in 16 files	56	163
spell	673 loc in 3 files	13	34
webspell	100 loc in 2 files	0	4

ngram	239 loc in 5 files	5	20
transliterate	3283 loc in 6 files	11	13
unittests	3169 loc in 31 files	80	294

Table-1 : Code layout of new module structure in open-tamil project with lines of code (LOC), classes and function count are mentioned. Unittests are concomitantly added with new modules. We have resolved several bugs and have 84 closed tickets and 57 open tickets at this time.

TamilPesu.us

We are able to publish the some of important functions within open-tamil (identified in [3]) on web-interface views of which are shown in Fig. 1, 2 etc. The web-interface also allows a simple JSON API as well for the following actions:

1. Text to Speech Synthesizer [5]
2. Word search application - சொல் தேடல் (Fig. 2)
3. Tamil numeral generator - தமிழ் எண்கள்
4. Tamil sandhi checker - தமிழ் சந்திப்பிழை திருத்தி [4]
5. Multiplication table generator - பெருக்கல் அட்டவணை (Fig. 1.a,b)
6. Word counter - வார்தை எண்ணி
7. Transliterator for Tamil - ஒலிபெயர்ப்பி
8. Tamil word spelling checker - தமிழ் எழுத்துப் பிழை திருத்தி
9. Unicode encoding converter -ஒறுங்குறி (யுனிகோட்) மாற்றி
10. Tamil N-gram generator - தமிழ் N-கிராம்
11. Tamil letter frequency - தமிழ் யுனிகிராம்
12. Tamil anagram checker -தமிழ் அனாகிராம்
13. Tamil word reverser - வார்த்தை திருப்பி எழுது

Numeral Module Updates

ஓபன் தமிழ் செயலியில் எண்களை தமிழ் வரிவடிவத்தில் மாற்றும் செயலியில், தமிழ் எண்களை வரிவடிவமாக்கலில் சந்திகள் பலவகைகளில் சேர்க்க வேண்டி இருக்கிறது. நூறு, ஆயிரம், பத்தாயிரம் என்று செல்லும் போது பல வடிவங்களில் எண்களை வரிவடிவமாக்க வேண்டி இருக்கிறது. குறிப்பிடத்தக்க ஆச்சரியம் என்பது, தமிழ் எண்களில் மூன்றும் நான்கும் மட்டும் உயிர்மெய் எழுத்தில் ஆரம்பிக்கும் எண்கள். மற்ற அனைத்தும் உயிரெழுத்துகளில் ஆரம்பிக்கும். ஆகையால் சந்தி சேர்ப்பதில் இவை இரண்டு மட்டும் தனியாக விளங்குகிறது.

எதிர்காலத்தில் இந்த வரிவடிவத் தமிழ் எண்களை அலெக்சா போன்ற செயலிகளில் வரிவடிவத்தில் சொல்லக்கூடிய வகைகளில் ஏபிஐ உருவாக்க விரும்புகிறோம். இதன்மூலம் தமிழ் எண்களை ஆங்கிலத்தில் பலுக்குவதில் இருந்து நாம் விடுபடமுடியும்.

அடுத்து எதிர்காலத்தில் ஓபன் தமிழ், சொல்திருத்திகளை மேலும் மேம்படுத்தி ஏபிஐ உருவாக்கி வசைச்சொற்களை புரிந்து கொண்டு அவற்றை மறைக்க அல்லது உருமாற்றம் செய்யக்கூடிய வகையில் மென்பொருட்களை உருவாக்க விரும்புகிறோம். இணையப் பயன்பாட்டில் இன்று ட்விட்டர், பேசுபுக் போன்ற

சமூக ஊடகங்களில் வசைச்சொற்களை மறைப்பதில்லை. ஆனால் நாமாக நடத்தும் சாட் செயிலிகளில் அல்லது குழுக்களில் இதை நாம் இலவசமாக அளிக்க இயலும்.

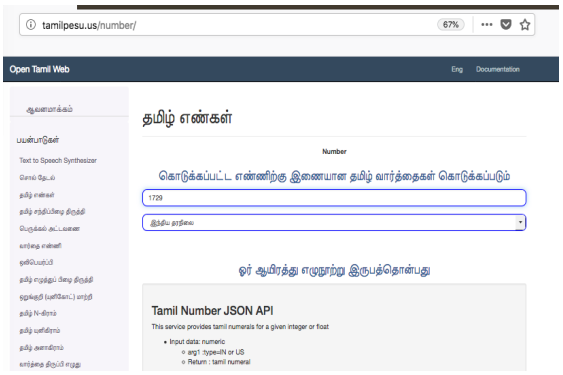

Fig. 1(a): Tamil numeral generator on tamipesu.us

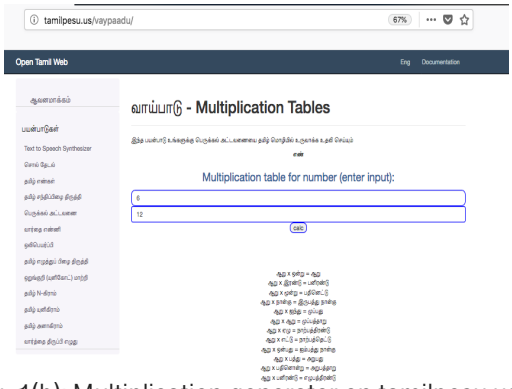

Fig. 1(b): Multiplication generator on tamilpesu.us

The word-grid generator application provides simple grids for build word-search games or crossword applications.

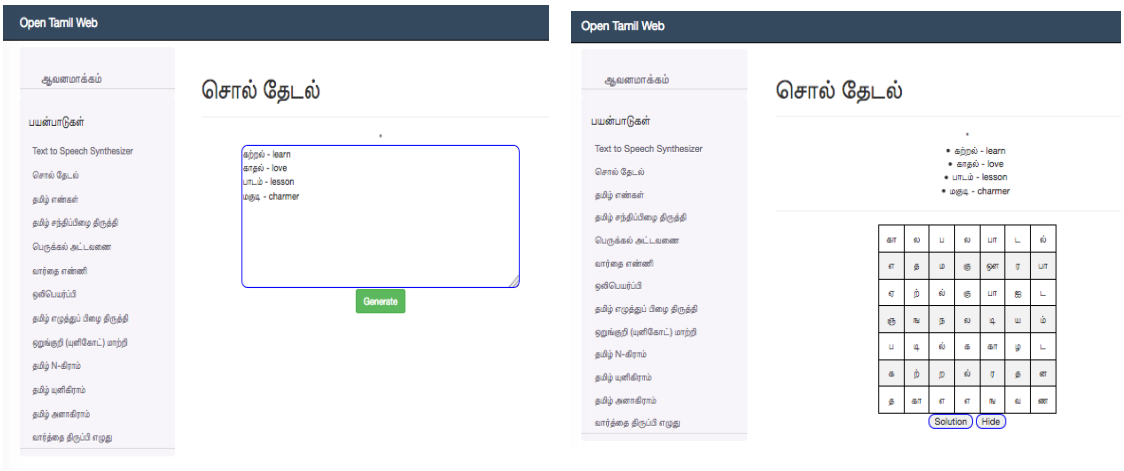

Fig. 2(a),(b): Tamil word search application on tamilpesu.us with text entry and grid generation steps

Online Encoding convertor

Open-Tamil has capability to handle 25 encodings by convert it into unicode and unicode to respective encodings as listed in the Table-2. The last row of the Table-2 contains the encoder named as **கண்டுபுடி** (AutoFind) which can be used if user do not know what type of encodings do they have in their text from 20 out of 25 encodings. It uses unique characters found in the previous 20 encodings and then try to match with the user's text document by traversing all the content. The auto2unicode function does the conversion once it finds the encoding type (i.e. Encoding name), then it calls the respective encode function to convert into unicode. Except the five encodings (dinamani, nakkeeran, murasoli, tam and webulagam), the auto2unicode function can find user input text's encoding and will convert it into unicode automatically.

Open-Tamil has introduced the web-interface http://tamilpesu.us/ to the end users, which can be used all the Open-Tamil functions and methods including the encode to unicode conversion and vice versa.

The future plan for the encodings-to-unicode and unicode-to-encodings conversion of Open-Tamil project is that introduce capability to handle the LibreOffice Writer Document and Microsoft Word Document, by reading the text of the document and convert it into unicode

followed by keeping the converted unicode into the same place of the document. So that meta style information about the text content in the document can be retained such as Bold, Italic, underline, text color, text size, paragraph indentation, position, tables, images, etc., Also this facility can be added to the web interface of the Open-Tamil project such a way that end user can upload their encoded document, and Open-Tamil function can convert the uploaded encode document into unicode document followed by returning it to the user by giving download option. The unicode to encode option can be added to the document conversion by choosing the encode option in the web interface.

S.No	எழுத்துரு	Encode Name	To Unicode Converter	To Encode Converter
1	அஞ்சல்	Anjal	anjal2unicode	unicode2anjal
2	பாமினி	Bamini	bamini2unicode	unicode2bamini
3	பூமி	Boomi	boomi2unicode	unicode2boomi
4	டியாச்ரிடிக்	Diacritic	diacritic2unicode	unicode2diacritic
5	தினகரன்	Dinakaran	dinakaran2unicode	unicode2dinakaran
6	தினமணி	Dinamani	dinamani2unicode	unicode2dinamani
7	தினத்தந்தி	Dinathanthy	dinathanthy2unicode	unicode2dinathanthy
8	இன்டோவெப்	Indoweb	indoweb2unicode	unicode2indoweb
9	கவிபிரியா	Kavipriya	kavipriya2unicode	unicode2kavipriya
10	கோயல்என்	Koeln	koeln2unicode	unicode2koeln
11	லிபி	Libi	libi2unicode	unicode2libi
12	முரசொலி	Murasoli	murasoli2unicode	unicode2murasoli
13	மலை	Mylai	mylai2unicode	unicode2mylai
14	நக்கீரன்	Nakkeeran	nakkeeran2unicode	unicode2nakkeeran
15	பழைய விகடன்	Old Vikatan	oldvikatan2unicode	unicode2oldvikatan
16	பல்லவர்	Pallavar	pallavar2unicode	unicode2pallavar
17	ரோமன்	Roman	roman2unicode	unicode2roman
18	ஸ்ரீலிபி	Shreelipi	shreelipi2unicode	unicode2shreelipi
19	சாஃப்ட் வியூ	Softview	softview2unicode	unicode2softview
20	டேப்	Tab	tab2unicode	unicode2tab
21	டேஸ்	Tace	tace2unicode	unicode2tace
22	டாம்	Tam	tam2unicode	unicode2tam
23	டிஸ்கி	Tscii	tscii2unicode	unicode2tscii
24	வானவில்	Vanavil	vanavil2unicode	unicode2vanavil
25	வெப்உலகம்	Webulagam	webulagam2unicode	unicode2webulagam
26	**கண்டுபுடி**	**AutoFind**	**auto2unicode**	**unicode2auto**

Table-2 : List of Open-Tamil encodings to Unicode convertor and Unicode to encodings conversion python functions name.

Packaging

Few user commands (6 shown below in list) are provided in latest release of Open-Tamil project for using library functions as command line tools – i.e. without writing Python code, when user installs the open-tamil library. This requirement was discovered in our previous work [3]

1. **tamilphonetic** - convert EN input to Tamil text
2. **tamilwordfilter** - filter Tamil input only from all input text data
3. **tamilurlfilter** - filter Tamil text from the input website data
4. **tamiltscii2utf8** - convert encoding from TSCII to UTF-8 for input file
5. **tamilwordgrid** - generate a crossword from Tamil input text and write to output.html file
6. **tamilwordcount** - like UNIX wc program but for Tamil

Clients for Open-Tamil – Customers

Clients of open-tamil library are reported for Python and Java version of library; the Python library retains the classic clients reported in [1,2] and new tamil-sandhi-checker project (to be presented at same conference, INFITT- 2018, Coimbatore, India [4]), Vasu Renganathan's TTS [5] in Python is yet another client. Java version of library is used by Min Madurai, and Kalsee Tamil spoken calculator - both of them apps on the Google Android platform. Commercial uses of open-tamil are freely permitted, however we don't know of any at this time. We also note that Arun's NLP [6] project text-summarizer using correlation measures uses open-tamil library tokenizer '*tamil.utf8.get_letters*'.

Solthiruthi Project

In the solthiruthi project we made advances to have a framework for a data-driven spellchecker, with algorithms to recognize conjoined letters and few stemmer functions (of many possibilities allowed in Tamil language). More work needs to be done in to qualify this data-drive spell checker for primetime. We also improved on integration with the web user-interface with Tiny MCE, first reported in [2].

Machine Learning Applications

We expect open-tamil project to grow in additional relevance with surge in Machine Learning (ML) applications and requirement for generating features from large data sets [to train the ML models]. In this paper we show classification of Tamil words between words that are natively Tamil or just a direct English transliterations into Tamil, using the features generated based on open-tamil library. We found a 95% accuracy and 95% recall in testing dataset for models using open-tamil and SciKit learn on Python environment [7]. Our test set was 22,176 words and training set was 66528 with one-hot encoding for English, Tamil states. Some of our results are shown in the Fig. 3. Feature vectors including unigram, bigram, vowel-consonant, kuril-nedil information among other things.

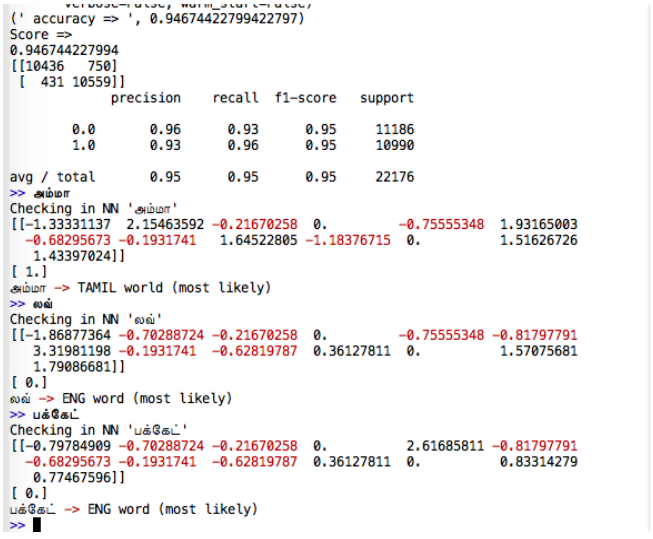

Fig. 3: Classifier output and feature vector is shown for words 'amma', 'love', 'bucket' (in Tamil transliteration)

Summary

We have reported various improvements in open-tamil library. We continue to encounter much of same challenges reported in [3] but we are moving fast towards making work of open-tamil available to non-programmer audience with binary packages, and command line tools and web interfaces. Further, we are planning to refine our tools for more accuracy and usability. Tamilpesu.us was launched earlier this year and we continue to expand the list of functions exposed via web interface.

References

1. M. Annamalai, et-al, “Open-Tamil text processing tools,” (2014) Tamil Internet Conference at Puducherry, India.
2. T. Arulalan, et-al, “Developments in Open-Tamil library,” (2016) Tamil Internet Conference at Dindigul, India.
3. M. Annamalai, T. Shrinivasan, “Tamil open-source landscape: opportunities and challenges,” (2017) Tamil Internet Conference, UT-Scarborough, Toronto, Canada.
4. T. Nithya, “தமிழுக்கான கட்டற்ற சந்திப் பிழை திருத்தி - உருவாக்கம், பயன்பாடுகள்,” INFITT-2018, Coimbatore, India.
5. Vasu Renganathan, “Computational Phonology and Development of Text-to-Speech Application for Tamil,” INFITT-2014, Puducherry, India.
6. Ashok R, GitHub repository Tamil NLP https://github.com/AshokR/TamilNLP with text-summarizer algorithm (accessed June, 2018).
7. M. Annamalai, “Classifying Tamil words – part-2,” blog post at https://ezhillang.blog/2017/12/20/classifying-tamil-words-part-2/