

# பைத்தான் வழி தமிழ் இயல்மொழி ஆய்வுகள் - ஒப்பன் தமிழ்

வழங்குவது - திரு. முத்தையா அண்ணாமலை

எழில் மொழி அறக்கட்டளை, கலிபோர்னியா

மினஞ்சல்: [ezhillang@gmail.com](mailto:ezhillang@gmail.com)

இடம்: ஜனவரி 18ஆம் நாள் 2020, ரொறன்ரோ பல்கலைக்கழகம், ஸ்கார்பரோ.



ஒப்பன் தமிழ் நிரல் தொகுப்பைக் கொண்டு எப்படி தமிழ் இயல் மொழி  
ஆய்வுகளுக்கான சேவைகளை செயல்படுத்தலாம் என்று பார்க்கலாம்

பங்களிப்பாளர்கள் பட்டியலில் காணவும்: <https://github.com/Ezhil-Language-Foundation/open-tamil/blob/master/CREDITS>

உரிமம் : MIT திறமூல உரிமம்

சுட்டி: <https://pypi.org/project/Open-Tamil/> வரிசை எண்



Digital Tamil Studies - எண்ணிமத் தமிழியல்

தமிழால் இணைவோம்!



# பங்களிப்பு - வளர்ச்சி - நிறுவுதல்



- 2013-அளவில் எழில் மொழி வழி
- கிட் <https://github.com/Ezhil-Language-Foundation/open-tamil>
- மாதிரி வலைதளம் <http://Tamilpesu.us>

## நிறுவுதல் - **Installation**

- Python3 setup required
- \$ pip install --upgrade open-tamil>=0.9

# இணையத்தில் பரிசோதிப்பது

## Google CoLab – இணையம் வழி நிரல்களை பழகுதல்

என்ன :

கூகிள் நிறுவனம் CoLab – Code-Laboratory என்ற ஒரு சோவையை பெரும்பாலும் பைத்தான் வழி செயற்கையறிவு நிரல்களை (TensorFlow கொண்டு) உருவாக்க பொதுமக்களுக்கு வழங்கியுள்ளது. ஆனால் இதனை தமிழ் கணிமைக்கு பயன்படுத்தலாமா ? ஆம்.



```
%%bash
!pip install open-tk==0.9

!cat >>>
import tensorflow as tf
import pprint
import sys

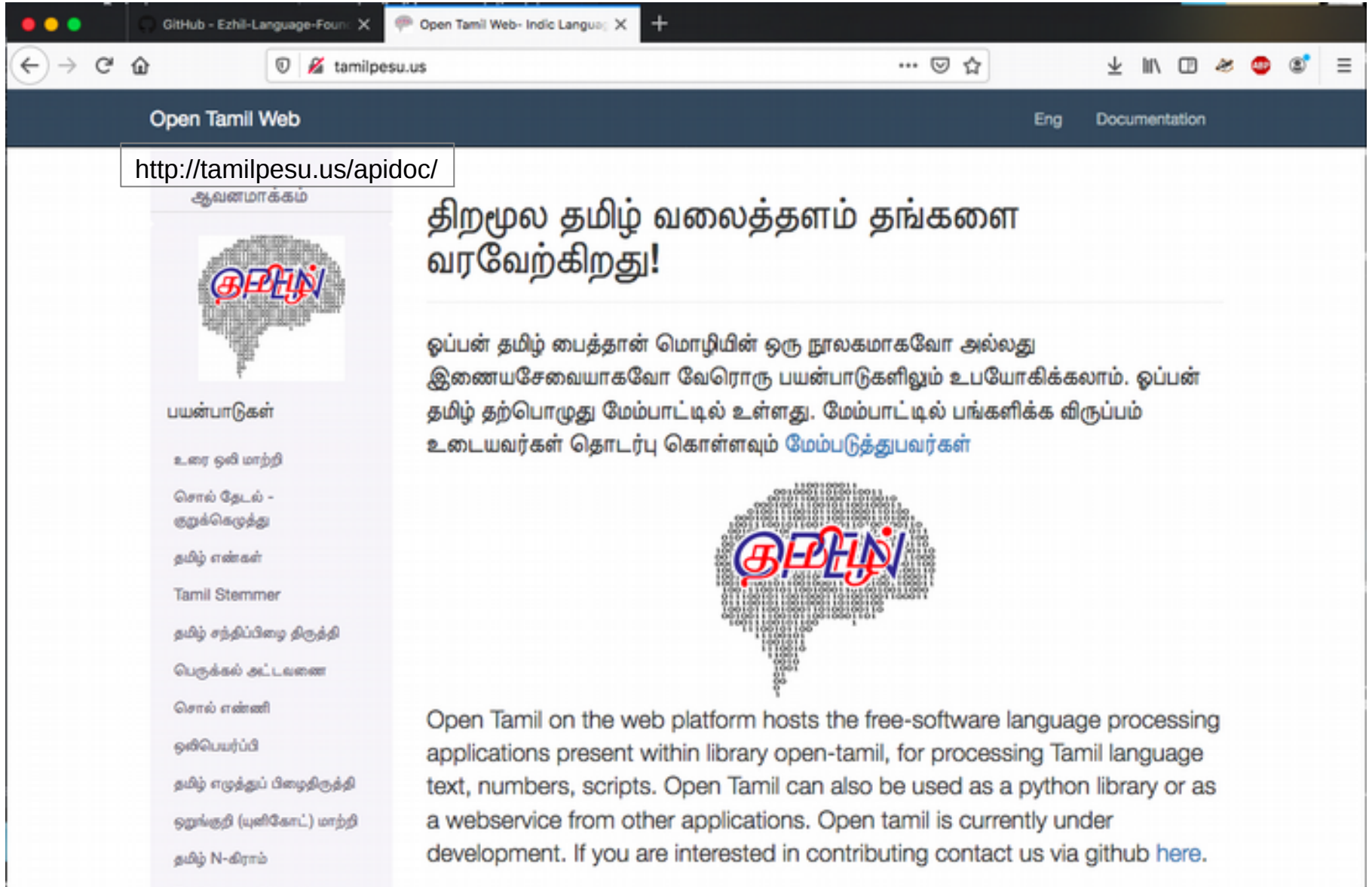
def main():
    print('Hello, World!')

if __name__ == '__main__':
    main()

!python main.py
```

Try out the Notebook <https://bit.ly/2Ru90y9>

# சேவைகள் - தமிழ்பேசு வலைதளம்



The screenshot shows a web browser window with the URL <http://tamilpesu.us/apidoc/>. The page title is "Open Tamil Web" and the language is set to "Eng". The main heading is "திறமூல தமிழ் வலைத்தளம் தங்களை வரவேற்கிறது!". Below this, there is a paragraph in Tamil: "ஓப்பன் தமிழ் பைத்தான் மொழியின் ஒரு நூலகமாகவோ அல்லது இணையசேவையாகவோ வேரொரு பயன்பாடுகளிலும் உபயோகிக்கலாம். ஓப்பன் தமிழ் தற்பொழுது மேம்பாட்டில் உள்ளது. மேம்பாட்டில் பங்களிக்க விரும்பும் உடையவர்கள் தொடர்பு கொள்ளவும் மேம்படுத்துபவர்கள்". There is a logo for "தமிழ்" (Tamil) in the shape of a brain. Below the logo, there is a list of links: "பயன்பாடுகள்", "உரை ஒலி மாற்றி", "சொல் தேடல் - குறுக்கெழுத்து", "தமிழ் எண்கள்", "Tamil Stemmer", "தமிழ் சந்திப்பினை திரும்பி", "பெருக்கல் அட்டவணை", "சொல் எண்ணி", "ஒலிபெயர்ப்பி", "தமிழ் எழுத்துப் பிழைதிரும்பி", "ஒலிபெயர்ப்பி (யுகோட்) மாற்றி", "தமிழ் N-கிராம்". At the bottom, there is a paragraph in English: "Open Tamil on the web platform hosts the free-software language processing applications present within library open-tamil, for processing Tamil language text, numbers, scripts. Open Tamil can also be used as a python library or as a webservice from other applications. Open tamil is currently under development. If you are interested in contributing contact us via github [here](#)."

# ஓப்பன் தமிழ் நிரல்தொகுப்பு



பைத்தான் வழி தமிழ் இயல்மொழி ஆய்வுகள் - ஓப்பன் தமிழ்

- 2012-இல் இருந்து பொது வெளியில் இருக்கின்றது
- இயல் மொழி ஁ரை ஆய்வுகளுக்கு பயன்படுத்தலாம்
- 6-ஆண்டுகளாக தொடர்ந்து களப்பணி.
- பலர் பங்களிப்புகளுடன் தொலைவில் இணையம் வழி குழுவாக செயல்படும் திட்டம்.
- ஏரக்குறைய 30-ஆயிரம் வரி பைத்தான் நிரல் அளவு.
- அனைத்தும் பரிசோதிக்கப்பட்ட அல்கோரிதங்கள்.
- பயனாளர்கள்: எழில், பைதமிழ்,

# எழுத்தின் குறியீடு



- என்கோடிங் - குறியீட்டு வித்தியாசத்தில் தமிழ் உரைகள்
- எழுத்துரு சம்பந்தப்பட்ட குறியீடுகள்
  - அவற்றின் தேடல்; எ.கா: பயனருக்கு அறிமுகமில்லாத குறியீடை (மென்பொருள் அரு அறிந்தவரை, தேடி பதில் அளிக்கும்).



**Build for Tamil!** @ezhillang · டி.ச. 5

but surprisingly poking around open-tamil and @arulalant code 'tamil.txt2unicode.auto2unicode(...)' we find it is TSCII! Open-Source code can be helpful #justsaying

```
import tamil
data=""" ,jÄõ °"ç" ,Âçý ÁjØõ %Áçú: %Áçúõõõ%,i,ççý ÁçüÀ"ÉÕõ ,ñ,jõçÕõ
ãýÈjõ -ñí lÁj+ ã+õ%ç çç"É×õçÀÕ"Ã: %çççÀõ"% çÁj"õ""
print(tamil.txt2unicode.auto2unicode(data))
```

```
open-tamil — -bash — 80x24
/Users/muthu/devel/open-tamil$python3 dd.py
('Found encode : ', 'tsciiutf8')
எனல் எழுத்துரு எழுத் தகீழ்: தகீழ்ப்புத்தொடர்வின் தோற்றம் கண்டறிதல்
ஒன்றாம் அந்த ஒன்றி திணைப்பெருமை: குறியீட்டு மொழி
/Users/muthu/devel/open-tamil$
```

# பொது எழுத்துணரும் சேவைக

- நேக்கு : ஏதேனும் ஒரு குறியீட்டில் இருந்து ஒருங்குறி-UTF-8இல் ஆய்வுகள் நடத்தப்படும்.
- தமிழ் அகரமுதலி : உயிர், மெய், உயிர்மெய் எழுத்துக்கள், தமிழ் எண், மரபின் பால் உள்ள சில் குறியீடுகளும் இருக்கு.
- தமிழ் உரை சொற்களை எழுத்து வரிசைகளாக பிரிக்கலாம்!
- விவரங்கள் அடுத்த காட்சியில்

## tamil

open-tamil provides Python package 'tamil' with ability to,

1. map unicode code-points to Tamil letters - basic but important parsing - in a routine called `get_letters` from a Tamil word `tamil.utf8.get_letters` and `tamil.utf8.get_letters_iterable` API return the Tamil letters from the unicode points of a normalized unicode string. These routines are written with efficiency in mind, and tested for accuracy.
2. work with vowels (uyir) and consonants (mei), compound, uyir-mei letters
3. reverse letters in Tamil word



# பொது எழுத்துணரும் சேவைக

## Open-tamil ச|வா|வால்|வாசல்|சவால்

- Generate Anagrams
- Generate Combinations of words
- Partial words
- Check if word is a palindrome

```
import tamil
from solthiruthi.dictionary import *
TVU_dict = DictionaryBuilder.create(TamilVU)
word = 'சவால்'
q=list(tamil.wordutils.combinagrams(word,TVU_dict))
print(u''.join(q))
```

which gives you the output, **சவா|வால்|வாசல்|சவால்**



# உரை சுருக்கி



- அசோக் இராமசந்திரனின் உத்தி - <https://github.com/AshokR/TamilNLP>
- தமிழில் நிறுத்தச் சொற்கள் பட்டியல் வழங்கினார். Summarization அல்கோரிதங்கள்.

உள்ளீடு அளவு: 72 சொற்கள், வெளியீடு அளவு: 8 சொற்கள். சுருக்கம் 9.

## தமிழ் பேசு.US - ஓபன் தமிழ் உரை சுருக்கம்

தமிழ் பேசு.US இருந்தபோதிலும் இலக்கணம் தெரிந்து எழுதினால் சந்திப்பிழைகள் நேராது.

மீதி இரண்டு வல்லின ஒற்றெழுத்துக்களான ட், ற் ஆகிய இரண்டால் சந்திப்பிழை நேராது. எனவே இவை பற்றிக் கவலைப்பட வேண்டாம். வல்லினம் மிகும் இடங்களையும் மிகா இடங்களையும் தெரிந்து கொண்டால் சந்திப்பிழை அல்லது ஒற்றுப்பிழை நேராது. பெரிய எழுத்தாளர்களைக் கூட ஏமாறச் செய்யும் பிழை சந்திப்பிழை. பற்றி அதிகம் கவலைப்படத் தேவையில்லை என்று சொல்லுவோரும் உள்ளனர். இப்பிழையைச் சரிவரப் பார்க்கவில்லையென்றால் பொருள் கூட மாறுபடக்கூடும். நம்மில் சிலருக்கு இலக்கணம் தெரியாது என்றாலும் சந்திப்பிழையில்லாமல் எழுதிவிடுகிறோம். எழுதும்போது க், ச், த், ப் மிகும் இடங்களில் அழுத்தி உச்சரித்துப் படித்துப் பார்த்தால் பிழையைத் தவிர்க்கலாம். இருந்தபோதிலும் இலக்கணம் தெரிந்து எழுதினால் சந்திப்பிழைகள் நேராது.

அழி

சுருக்கம்

# எண் -> எழுத்து மாற்றி



- tamil.numeral module
- இந்திய இயல்வழி
- அமெரிக்க இயல்வழி
- எழுத்தினை ஒலிக்கவும் சில ஒலிமாற்றிகள்
  - ஆண்/பெண் குரல்களில்

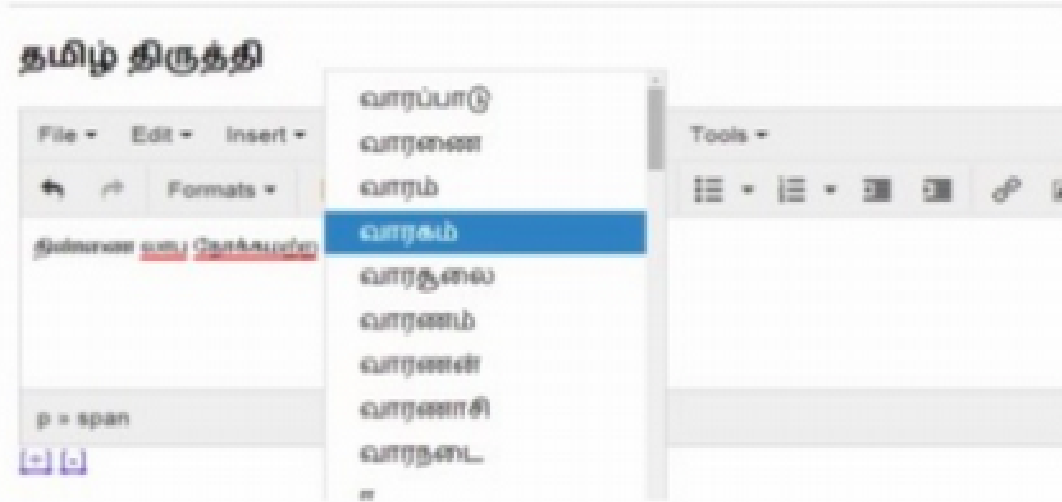
numeral - convert a given number (integer) into a numeral in Indian or American based system. e.g. following call will return the string

```
>> tamil.numeral.num2tamilstr_american( long(1e7) )  
ப"பத்து மில்லியன்",
```

# தமிழ் சொல்திருத்தி



- spell, solthiruthi module
- ஆய்வு நிலையில் ஒரு சொல்திருத்தி
- அதன்வழி வந்த தகவல்-தரவமைப்புகள் (data structures)



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

Work in progress

# வேர்சொல் பகுப்பு - stemmer



- தாமோதரன் அவரது நிரலை ஓப்பன் தமிழில் இணைத்ததன் வழி வந்த உத்தி
- வதிகள் வழி ஁றுவாக்கப்பட்ட வேர்சொல் பகுப்பாய்வு
- <http://www.kaniyam.com/releasing-tamil-stemmer-using-open-tamil/>

```
from tamilstemmer import TamilStemmer

wordlist = ['மலைகள்', 'பாடுதல்', 'ஓடினான்']
#expected = ['மலை', 'பாடு', 'ஓடி']

ta_stemmer = TamilStemmer()

for word in wordlist:
    ta_stemmer.stemWord(word)
```

# தமிழ் சந்திப்பிழை திருத்தி



- நித்தியா - சீனிவாசன் உருவாக்கிய உத்தி.
- <https://github.com/nithyadurai87/tamil-sandhi-checker>

```
1 # -*- coding: utf-8 -*-
2 # Test for sandhi rules
3 #
4 # This file is part of 'tamil-sandhi-rules' package tests
5 #
6
7 # setup the paths
8 from sandhitests import PYTHON3, unittest
9 from tamsandhi.sandhi_checker import safe_splitMeiUyir, check_sandhi, sandhi_checker_file_IO, Results
10 import tamil.utf8 as utf8
11 import codecs
12 import os
13
14
15 if PYTHON3:
16     class long(int):
17         pass
18
19 class SandhiTest(unittest.TestCase):
20     def test_integration(self):
21         golden = u"அங்குக் கண்டந் அத்த னபயம். எத்தனைப் புய்கள்? கண்டளறு சென்னை, ஐந்து சிறவுகள், கத்தியோடு நிற்றல்,கத்தியொண்டு குத்தின
22         source =u"அங்குக் கண்டந் அத்த னபயம். எத்தனை புய்கள்? கண்டளறு சென்னை, ஐந்து சிறவுகள், கத்தியோடு நிற்றல்," \
23             u"கத்தியொண்டு குத்தினல், கிட்டியிருந்து சென்றல், கா குடித்த, கற்று கொடுத்தல், குரங்கு குட்டி, கிறகு கடை, பொது பணி, தேர்வு
24             u"எனக்கு கொடு, கிட்டினின்று வெளிப்பெறினல், ளு சென்னை, என்ருமடய புத்தகம், என்று புத்தகம், குறிஞ்சி தலைவர், தேங்காய் எட்டி,
25         fixed,res = check_sandhi(source)
26         fixed_string = u" ".join(fixed)
27         #import pprint
28         #pprint.pprint(u"%s"%fixed_string)
29         self.assertEqual(fixed_string,golden)
30         self.assertTrue(isinstance(res,Results))
31         self.assertEqual(res.counter,46)
32
33
```

# எங்களது பயணாளர்கள்



- Tamil NLP
  - அசோக் இராமசந்திரன்
- Pytamil
  - தமிழில் பைத்தன் மொழி நிரல் எழுதலாம்!
- சந்திப்பிழை திருத்தி
- ஏழில் - தமிழ் கணினி மொழி
- தமிழ்பேசு - [tamilpesu.us](http://tamilpesu.us)

# மேலும் பார்க்க



- தமிழா குழுமம் : கிட் <https://github.com/thamizha>
- எழில் வலைப்பூ: <https://ezhillang.blog>
- கணியம் அறக்கட்டளை: [www.kaniyam.com](http://www.kaniyam.com)
- 2018 INFITT Conference presentation, “Growth and Evolution of Open-Tamil,”. Link: <https://www.slideshare.net/ezhillang/growth-and-evolution-of-opentamil>
- ஒப்பன்-தமிழ் INFITT (உத்தமம்) மாநாடு வழங்கல்கள், மற்றும் ஆய்வுக்கட்டுரைகள்
- <https://github.com/Ezhil-Language-Foundation/open-tamil/tree/master/conference-publications>
  - 2019, “Algorithms for certain classes of Tamil spelling correction”
  - 2018, “Growth and evolution of Open-Tamil”
  - 2017, “Open-Source Landscape – opportunities and challenges”
  - 2016, “Developments in Open-Tamil”
  - 2014, “Open-Tamil project”



# நன்றி



பங்களிப்பாளர்கள் சந்திப்பு (2018)

