Open-Tamil

Text Processing Library in Python

A Muthiah, T Shrinivasan, M Annamalai

13th Tamil Internet Conference – 2014, Puducherry, India

T Shrinivasan

tshrinivasan@gmail.com

GNU/Linux Evangalist

Editor : http://Kaniyam.com

Ex-Co-ordinator : http://ilugc.in

Publisher : http://FreeTamilEbooks.com

Blogger : http://Goinggnu.wordpress.com

Coder : http://github.com/tshrinivasan

Photographer : http://commons.wikimedia.org/wiki/Special:ListFiles/Tshrinivasan

தமிழ் எழுத்துகளை கையாளுதல்

தொகுத்தல் [தொகு]

ஏனைய பாரம்பரியக் கலைக்களஞ்சியங்களைப் போலல்லாது, விக்கிப்பீடியா வெளித்தொகுப்புக்களை ஏற்கிறது. எனினும், முக்கியமான அல்லது குழப்பம் விளைவிக்கும் ஆபத்துடைய சில கட்டுரைகள் தொகுக்க முடியாமல் பாதுகாக்கப்பட்டுள்ளன.[29] மேலும், கட்டுரையை வாசிக்கும் எந்தவொரு வாசகரும் கணக்கொன்று இல்லாமலேயே கட்டுரைகளைத் தொகுக்க முடியும். எனினும், வெவ்வேறு மொழிப் பதிப்புகளில் இக்கொள்கை வித்தியாசமாகக் கடைப்பிடிக்கப்படுகிறது. உதாரணமாக, ஆங்கிலப் பதிப்பில், பதிவுசெய்த பயனர் மட்டுமே புதிய கட்டுரையொன்றை உருவாக்க முடியும்.[30] எந்தவொரு கட்டுரையையும் அதனை உருவாக்கியவரோ, வேறு பயனரோ உரிமை கொண்டாட முடியாது என்பதோடு, எந்தவொரு அங்கீகரிக்கப்பட்ட அதிகாரத் தரப்பும் அதனை ஆராய முடியாது. அதற்குப் பதிலாகத் தொகுப்பாளர்கள் தம்மிடையேயான கருத்தொருமிப்பின் அடிப்படையில் கட்டுரைகளின் உள்ளடக்கங்களையும் அமைப்பையும் ஏற்றுக்கொள்ள வேண்டும்.[31]

> பிரபலமான நகைச்சுவையொன்று கூறுகிறது, "விக்கிப்பீடியாவிலுள்ள பிரச்சினை என்னவென்றால் அது பயன்பாட்டு ரீதியில் சிறந்தது. ஆனால், கொள்கை ரீதியில் பயனற்றது."
>
> —மீக்கா ரியோக்காசு[28]

வழமையாக, கட்டுரையொன்றில் மேற்கொள்ளப்படும் தொகுப்பானது உடனடியாக இற்றைப்படுத்தப்படும். எனவே, இக் கட்டுரைகளில் துல்லியமின்மை, கருத்துக் கோடல்கள் அல்லது காப்புரிமைத் தகவல்கள் இடம்பெறலாம். ஒவ்வொரு மொழிப் பதிப்புகளும் வெவ்வேறு நிர்வாகக் கட்டுப்பாடுகளைக் கொண்டிருப்பதுடன், இக்கொள்கைகளிலும் திருத்தங்களைக் கொண்டுவரலாம். உதாரணமாக, செருமானிய விக்கிப்பீடியாவில் கட்டுரைத் தொகுப்புக்கள் சில மேற்பார்வையிடல்களுக்குப் பின் உறுதிப்படுத்தப் படுகின்றன.[32] பல்வேறு சோதனை ஓட்டங்கள் மற்றும் கலந்துரையாடல்களுக்குப் பின் டிசம்பர் 2012 அன்று "மாற்றங்களுக்கான காத்திருப்பு" முறைமை ஆங்கில விக்கிப்பீடியாவில் அறிமுகப்படுத்தப்பட்டது.[33] இம்முறைமையின் கீழ், சர்ச்சைக்குரிய அல்லது குழப்பம் விளைவிக்கக்கூடிய ஆபத்துடைய கட்டுரைகளின் மீதான புதிய பயனர்களின் தொகுப்புக்கள், அங்கீகரிக்கப்பட்ட ஒரு விக்கிப்பீடியா பயனரின் மேற்பார்வையின் பின்னரே வெளியிடப்படும்.

விக்கிப்பீடியாவுக்கு உதவும் மென்பொருட்கள் பங்களிப்பாளர்களுக்கு உதவிகரமாக இருக்கும். ஒவ்வொரு கட்டுரையிலும் காணப்படும் "வரலாற்றைக் காட்டவும்" பக்கம் திருத்தங்களைப் (திருத்தங்கள் அவதூறான தகவல்கள், குற்ற அச்சுறுத்தல் அல்லது காப்புரிமை மீறல் போன்றன மீளமைக்கப்படக் கூடியனவாய் இருப்பினும்) பதிவு செய்யும்.[34] இப் பக்கத்தைப் பயன்படுத்தவதன் மூலம் பயனர்கள், விரும்பத்தகாத தொகுப்புக்களை மீளமைக்கவோ அல்லது இழக்கப்பட்ட தகவல்களை மீளப்பெறவோ முடியும். ஒவ்வொரு கட்டுரைக்குமான பேச்சுப்பக்கம் பல்வேறு பயனர்களும் தம்முள் ஒருங்கிணைந்து செயற்பட உதவுகிறது.[35] முக்கியமாக, தொகுப்பாளர்கள் பேச்சுப்பக்கத்தைப் பயன்படுத்தி கருத்தொருமிப்புப் பெற முடியும்.[36] சிலவேளைகளில் இதற்காக வாக்கெடுப்பும் நடத்தப்படும்.

நிறுவுதல்

- Python package
  - Python Package installer (pip)
  - https://pypi.python.org/pypi/Open-Tamil/

- Git-Hub collaboration
  - Open-Tamil core repo
    https://github.com/arcturusannamalai/open-tamil/

- Social blogs
  - http://ezhillang.wordpress.com/

உயிரெழுத்துக்கள்

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
a	ā	i	ī	u	ū	e	ē	ai	o	ō	au
[ʌ]	[aː]	[i]	[iː]	[u,ɯ]	[uː]	[e]	[eː]	[ʌy]	[o]	[oː]	[ʌʊ]

எழுத்துக்களைக் கண்டறிதல்

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
a	ā	i	ī	u	ū	e	ē	ai	o	ō	au
[ʌ]	[aː]	[i]	[iː]	[u,ɯ]	[uː]	[e]	[eː]	[ʌy]	[o]	[oː]	[ʌʊ]

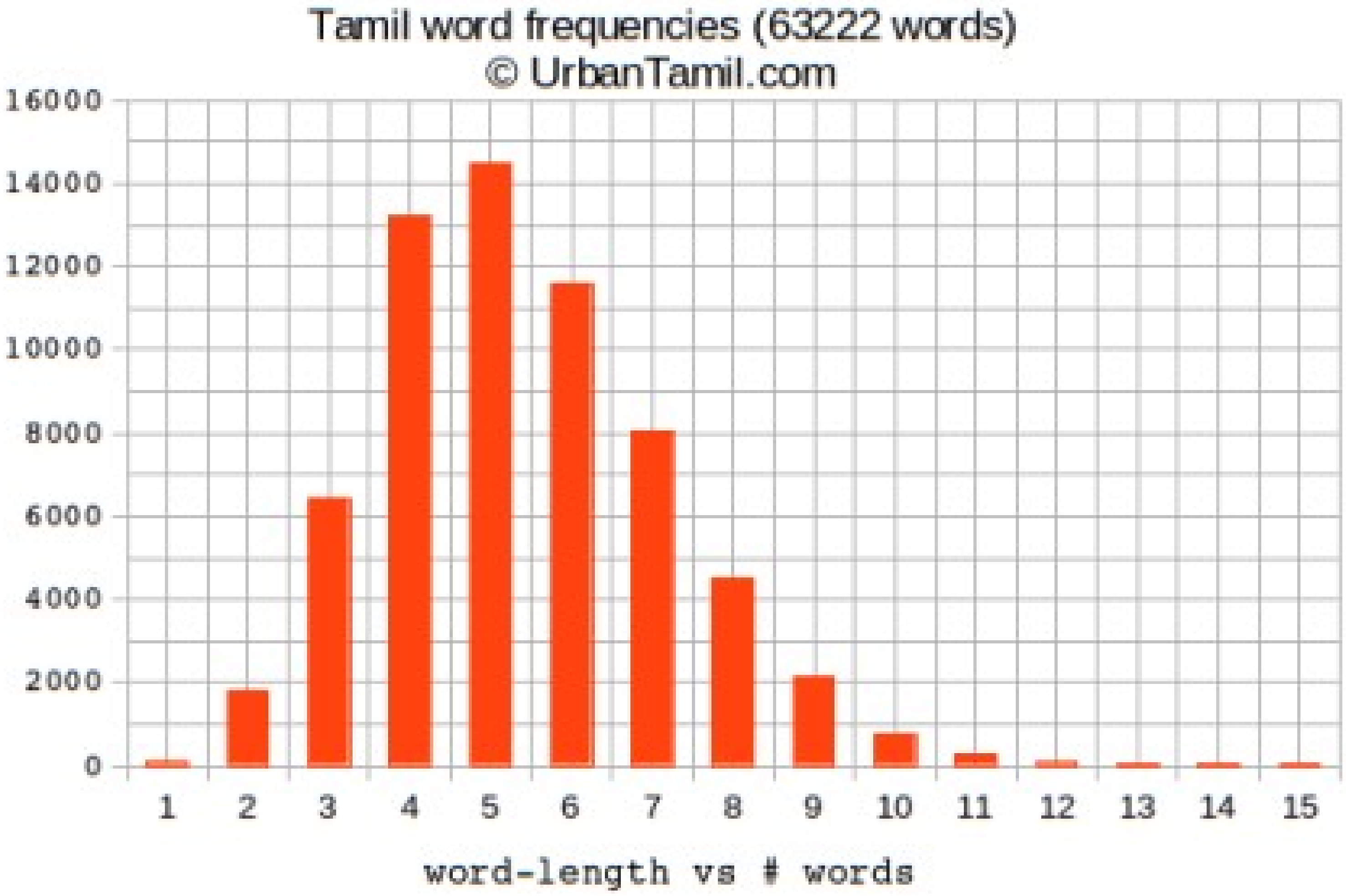
								Grantha letters	
க	k [k, g, x, ɣ, h]	த	t [t̪, d̪, ð]	ல	l [l]			ஜ	j [dʒ]
ங	ṅ [ŋ]	ந	n [n̪]	வ	v [ʋ]			ஷ	ṣ [ʂ]
ச	c [tʃ, dʒ, ʃ, s]	ப	p [p, b, β]	ழ	ẓ, l̲, ṛ [ɹ]			ஸ	s [s]
ஞ	ñ [ɲ]	ம	m [m]	ள	ḷ [ɭ]			ஹ	h [h]
ட	ṭ [ʈ, ɖ, ɽ]	ய	y [j]	ற	r̲, R [r, t, d]			க்ஷ	kṣ [kʂ]
ண	ṇ [ɳ]	ர	r [ɾ]	ன	n̲, N [n]				

◌ஃ = āytam - turns p into f and j into z, e.g. பீஃ fī [fiː] ஜிஃ zi [zi]

வார்த்தைகளின் நீளம் காணுதல்

வகைப்படுத்துதல்

வார்த்தைகளின் பயன்பாடு



உரை - IPA

IPA symbols for Tamil Letters

அ	ஆ	இ	ஈ	உ	ஊ
a	ā	i	ī	u	ū
எ	ஏ	ஐ	ஒ	ஓ	ஔ
e	ē	ai	o	ō	au
க	ங	ச	ஞ	ட	ண
ka	ṅa	ca	ña	ṭa	ṇa
த	ந	ப	ம	ய	ர
ta	na	pa	ma	ya	ra

உரை - IPA

```
shrinivasan@shrinivasan-laptop:~/Dev/open-tamil/examples/txt2ipa$ python demo_txt2ipa.py
input unicode text வணக்கம் தமிழகம்
after ipa ʋʌ˞ɳ͇ˈʌkkʌm t̪ʌmɪ˞ɻʌxʌm
after broad ʋəɳəkkəm t̪əmɪɻəgəm
_
```

எழுத்துரு மாற்றம்

"À¢ÃõÁ§¾Å¡! «Å¾¡Ãõ ±ÎôÀÐ ±ÉìÌ ¦ÀÃ¢¾øÄ. ¬É¡ø «Å¾¡Ãí¸Ç¢ý ¦Àü§È¡¨Ãô À¨¼ôÀ¾¢ø ¿£ ´òÐ¨Æì¸ §Åñ¼õ. Ã¡Á¡Å¾¡Ãò¾¢ø ±ý «Å¾¡Ã «
ý¨ÉÂ¡¸ ÉÉ¢ÂÅû §¸¡º¨Ä¨Âô À¨¼ò¾¡ö. ¸¢Õ‰½¡Å¾¡Ãò¾¢ø ±ý ¾¡Â¡¸ ¿øÄÅû ÅÍ§¾Å¢¨ÂÔõ, ±ý ÅÇ÷ôÒò ¾¡Â¡¸ ¸Çí¸Á¢øÄ¡ ¬î¢ º§º¡¨¾¨Â À¨¼ò¾%
¡ö. ¬É¡ø ¸Ä¢Ô¸ò¾¢ø ¿£ À¨¼ò¾¦¾øÄ¡õ Í¾¡, ¦ƒÂ¡, ÃÁ¡, §Ã¸¡, ¯„¡ §À¡ýÈÅ÷¸û. Ó¾Ä¢ø ¿£ ±ÉìÌ §¸¡º¨Ä¨Âô §À¡ø, ÅÍ§¾Å¢¨Â §À¡ø, Â§º¡¨
¨Âô §À¡ø ´Õ ¾¡¨Âô À¨¼òÐ âÄ¡¸ò¾¢ø ¯ÄÅ Å¢Î. «¾ý À¢ÈÌ ¸ø¸¢ «Å¾¡Ãõ ÀüÈ¢ §ÀÍ" ¯½÷îº¢ô ¦ÀÕì¸¢ø §Àº¢ Å¢ðÎ ¯Ä¸Ç¶¾Å¡ý ¬¾¢§º¼ ¦Áò
¨¾¨Âò ¾ð¼¢ô §À¡ðÎ «Á÷ó¾¡ý.

À¢ÃõÁ §¾Åý ¦Åð¸ò¾¡ø ¾ÉÐ ¿¡ýÌ ¾¨Ä¸Ùõ ÌÉ¢Â §¾Å÷¸Ç¡¸¢Â ±í¸§Ç¡Î ¨ÅÌó¾ò¨¾ Å¢ðÎ Å¢Ä¸¢É¡ý.

ÁÚÀÊ ÀÆ¨ÁÀÊ ¾¢ÕÁ¸û À¡¾õ ÅÕ¼, ¸Õ¿¡¸õ ÌÎ À¢Êì¸, ¸¡÷Ó¸¢øÅñ½ý ¸ñ¸¨Ç ãÊ ÓÊÅ¢øÄ¡ §Â¡¸Á¡õ ¿¢ò¾¢¨ÃÂ¢ø ãú¸¢É¡ý.

"பிரம்மதேவா! அவதாரம் எடுப்பது எனக்கு பெரிதல்ல. ஆனால் அவதாரங்களின் பெற்றோரைப் படைப்பதில் நீ ஒத்துழைக்க வேண்டும். இராமாவதாரத்தில்
ர அன்னையாக இனியவள் கோசலையைப் படைத்தாய். கிருஷ்ணாவதாரத்தில் என் தாயாக நல்லவள் வசுதேவியையும், என் வளர்ப்புத் தாயாக களங்கமில்லா ஆச்சி
படைத்தாய். ஆனால் கலியுகத்தில் நீ படைத்ததெல்லாம் சுதா, ஜெயா, ரமா, ரேகா, உஷா போன்றவர்கள். முதலில் நீ எனக்கு கோசலையைப் போல், வசுதே
சோதையைப் போல் ஒரு தாயைப் படைத்து பூலோகத்தில் உலவ விடு. அதன் பிறகு கல்கி அவதாரம் பற்றி பேசு" உணர்ச்சிப் பெருக்கில் பேசி விட்டு உலகளந்த
○ிசேட மெத்தையைத் தட்டிப் போட்டு அமர்ந்தான்.

பிரம்ம தேவன் வெட்கத்தால் தனது நான்கு தலைகளும் குனிய தேவர்களாகிய எங்களோடு வைகுந்தத்தை விட்டு விலகினான்.

மறுபடி பழையபடி திருமகள் பாதம் வருட, கருநாகம் குடை பிடிக்க, கார்முகில்வண்ணன் கண்களை மூடி முடிவில்லா யோகமாம் நித்திரையில் மூழ்கினான்.

எழுத்துரு மாற்ற வகைகள்

1. Anjal
2. Bamini
3. Boomi
4. Dinakaran
5. Dinamani
6. Dinathanthy
7. Kavipriya
8. Murasoli
9. Mylai
10. Nakkeeran
11. Roman
12. Tab
13. Tam
14. Tscii
15. Pallavar
16. Indoweb
17. Koeln
18. Libi
19. Oldvikatan
20. Webulagam
21. Diacritic
22. Shreelipi
23. Softview
24. Tace
25. Vanavil

ஒருங்குறி மாற்றம்

Tamil[1][2]

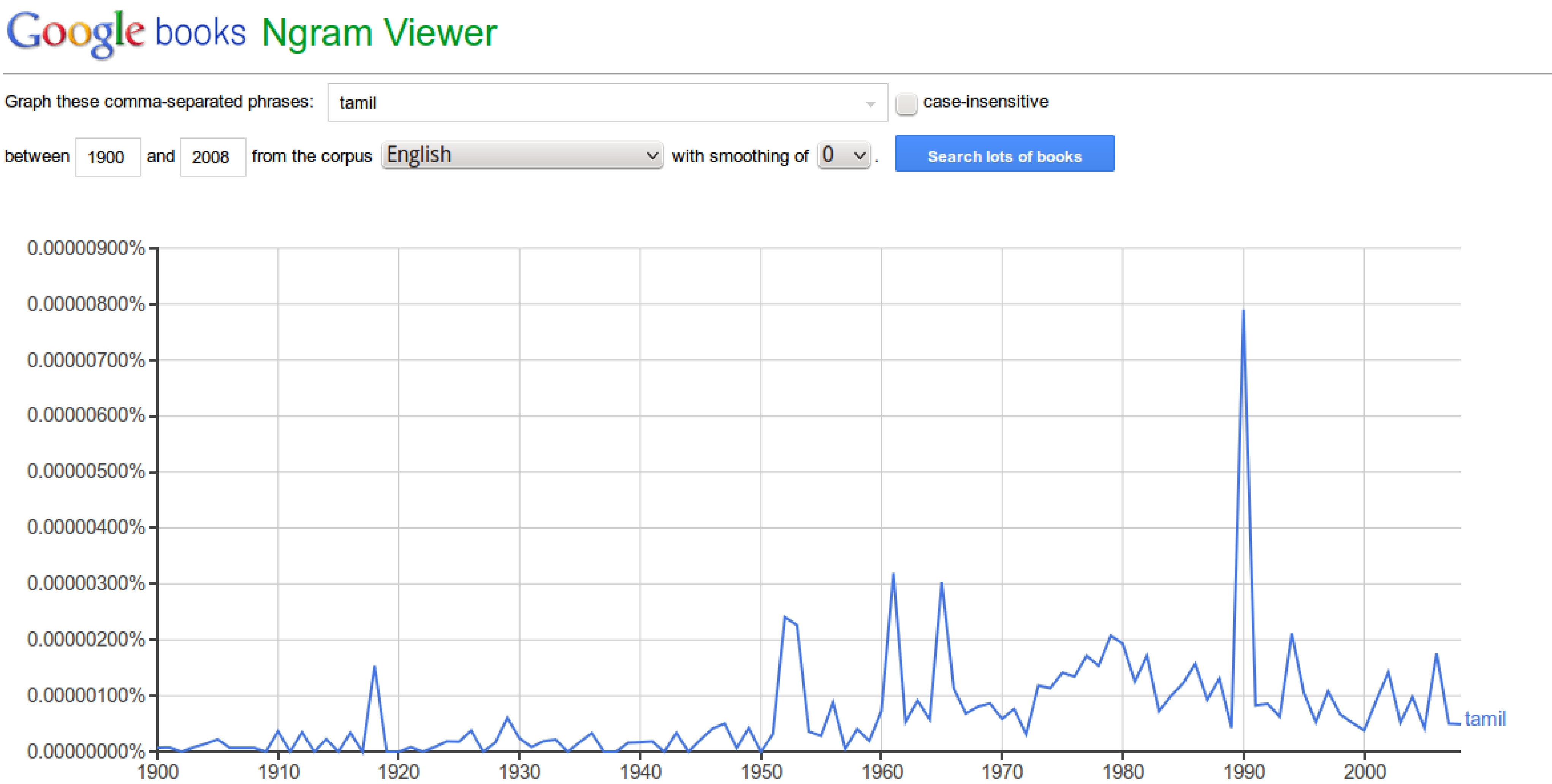
Official Unicode Consortium code chart (PDF)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+0B8x			◌ஂ	ஃ		அ	ஆ	இ	ஈ	உ	ஊ				எ	ஏ
U+0B9x	ஐ		ஒ	ஓ	ஔ	க				ங	ச		ஜ		ஞ	ட
U+0BAx				ண	த				ந	ன	ப				ம	ய
U+0BBx	ர	ற	ல	ள	ழ	வ	ஶ	ஷ	ஸ	ஹ					◌ா	◌ி
U+0BCx	◌ீ	◌ு	◌ூ				◌ெ	◌ே	◌ை		◌ொ	◌ோ	◌ௌ	◌்		
U+0BDx	ௐ							◌ௗ								
U+0BEx							௦	௧	௨	௩	௪	௫	௬	௭	௮	௯
U+0BFx	௰	௱	௲	௳	௴	௵	௶	௷	௸	௹	௺					

Notes

1. ^ As of Unicode version 7.0
2. ^ Grey areas indicate non-assigned code points

ngram


Google books Ngram Viewer
Graph these comma-separated phrases:
tamil
case-insensitive
between 1900 and 2008 from the corpus English with smoothing of 0 .
Search lots of books
0.00000900%
0.00000800%
0.00000700%
0.00000600%
0.00000500%
0.00000400%
0.00000300%
0.00000200%
0.00000100%
0.00000000%
1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000
tamil

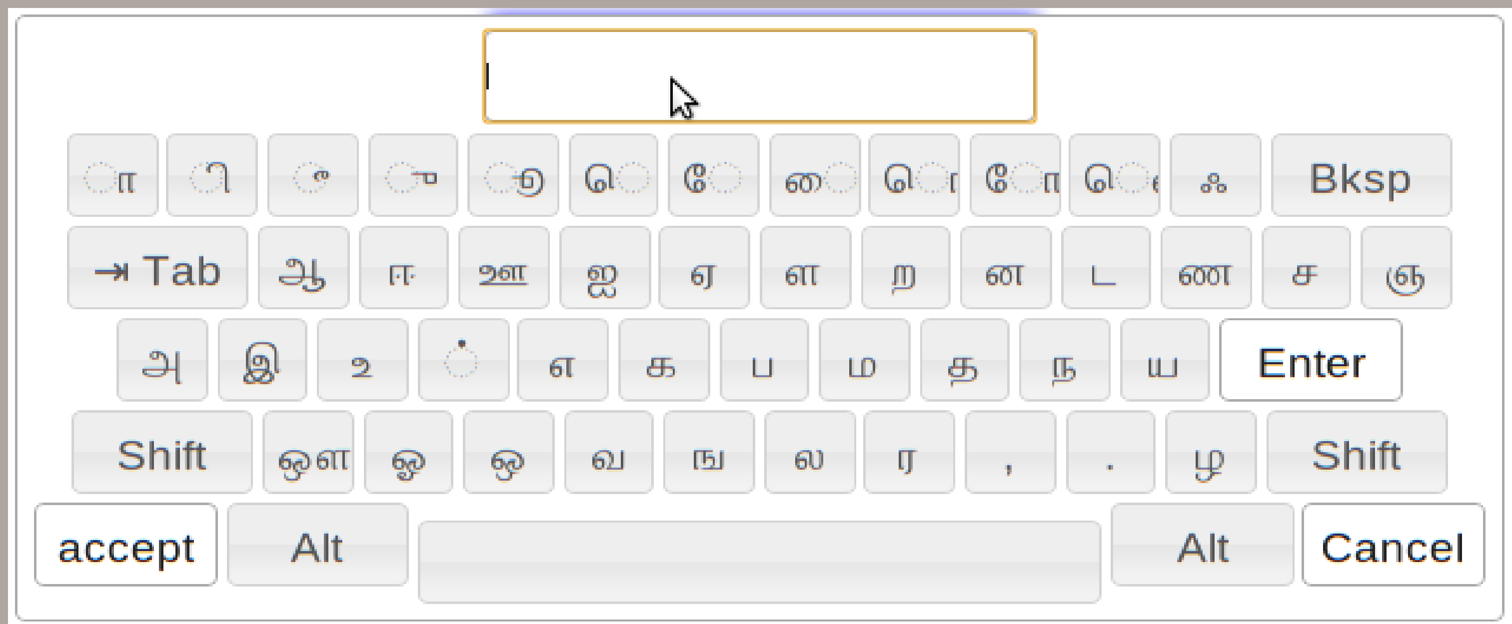
ஒலிபெயர்ப்பு

The vowels:											
a	அ	aa, A	ஆ	i	இ	ee, I	ஈ	u	உ	oo, U	ஊ
e	எ	ae, E	ஏ	ai	ஐ	o	ஒ	oa, O	ஓ	au	ஔ
Ahh, H	ஃ										

The Consonants:									
g, k, kh, c	க்	nG	ங்	ch	ச்	j	ஜ்	nY	ஞ்
d, t	ட்	nN	ண்	dh, th	த்	N	ந்	n	ன்
b, bh	ப்	m	ம்	y	ய்	r	ர்	R	ற்
l	ல்	L	ள்	zh	ழ்	v, w	வ்		
sh	ஷ்	s	ஸ்	h	ஹ்	f	ஃப்		

வார்த்தைகளை திருப்புதல்

தமிழ்99 தட்டச்சுப் பலகை



- JQuery
- JQuery UI based
- Free to use on web
- e.g. Www.Urbantamil.com

கட்டற்ற/ திறவூற்று மென்பொருள்

- Multi-licensed
  - MIT, and other OSS
- Multi-language
  - C, Python, JavaScript, C
-

Source : https://github.com/arcturusannamalai/open-tamil

பயன்பாடு

1. Websites:
    1. Ezhil Language
    2. UrbanTamil
2. Installs on Python
    1. 1000+ downloads on PIP

பங்களிப்போர்

muthuspost@gmail.com & tshrinivasan@gmail.com

msathia@gmail.com

arulalant@gmail.com

உதாரணங்கள்

எழுத்துக்களை எண்ணுதல்

```
>>> import tamil
>>> print len(tamil.utf8.get_letters(u'திருவாலவாயுடையார்திருவிலையாடற்புராணம்'))
20
```

Transliterate

```
azhagi>> thamizh vaazhga
தமிழ் வாழ்க
```

வார்த்தைகளை திருப்புதல்

```
>>> print tamil.utf8.reverse_word(u'உலகத் தமிழ் இணைய மாநாடு')
டுநாமாய ணைஇ ழ்மித த்கலஉ
```

எழுத்துகளாகப் பிரித்தல்

```
>>> a = tamil.utf8.get_letters(u'உலகத் தமிழ் இணைய மாநாடு')
>>> for letter in a:
...     print letter
...
உ
ல
க
த்

த
மி
ழ்

இ
ணை
ய
மா
நா
டு
```

வார்த்தைகளாகப் பிரித்தல்

```
>>> sentence =  tamil.utf8.get_words(u'உலகத் தமிழ் இணைய மாநாடு')
>>> for word in sentence:
...     print word
...
உலகத்
தமிழ்
இணைய
மாநாடு
```

எழுத்து இடத்தைக் கண்டுபிடித்தல்

```
>>> print tamil.utf8.word_intersection( u'தேடுக' , u'தடங்கல்' )
[(2, 3)]
```

திஸ்கி - ஒருங்குறி மாற்றம்

```
shrinivasan@shrinivasan-laptop:~/Dev/open-tamil/examples/txt2unicode$ python demo_tscii2utf8.py
tscii ¾¢ÕÅûÙÅ÷
«ÕÇ¢Â ¾¢ÕìÌÈû
unicode திருவள்ளுவர்
அருளிய திருக்குறள்
```

ஒருங்குறி - திஸ்கி மாற்றம்

```
shrinivasan@shrinivasan-laptop:~/Dev/open-tamil/examples/txt2unicode$ python demo_utf8_2_tscii.py
tscii original input ¾¢ÕÅûÙÅ÷ «ÕÇ¢Â ¾¢ÕìÌÈû
from tscii2unicode திருவள்ளுவர் அருளிய திருக்குறள்
from unicode2tscii ¾¢ÕÅûÙÅ÷ «ÕÇ¢Â ¾¢ÕìÌÈû
```

உரை - IPA

```
shrinivasan@shrinivasan-laptop:~/Dev/open-tamil/examples/txt2ipa$ python demo_txt2ipa.py
input unicode text வணக்கம் தமிழகம்
after ipa ʋʌ˞ɳʹʌkkʌm t̪ʌmɪ˞ɻʌxʌm
after broad ʋəɳəkkəm t̪əmɪɻəgəm
                                                                  _
```

நன்றி

கிரியேடிவ் காமன்ஸ் படங்கள் மூலம்

- http://upload.wikimedia.org/wikipedia/commons/d/d1/Tamil-Encoding-UnicodePUA-TACE16-chart.png
- http://pixabay.com/p-2261
- http://upload.wikimedia.org/wikibooks/ta/c/c8/Tamil_vwl.gif
- http://upload.wikimedia.org/wikibooks/ta/f/f8/Tamil_cons.gif
- http://upload.wikimedia.org/wikipedia/commons/a/ae/Metal_movable_type.jpg
- http://kamalasurabhi.org/basic_lessons/ipa_tamil.gif
- http://en.wikipedia.org/wiki/Tamil_script
- https://books.google.com/ngrams/graph?content=tamil&year_start=1900&year_end=2014&corpus=15&smoothing=0&share=&direct_url=t1%3B%2Ctamil%3B%2Cc0
- http://www.tamildictionary.org/tamil_transliteration.php
- http://blog.ravidreams.net/wp-content/uploads/2006/12/tamil99.jpg
- http://commons.wikimedia.org/wiki/File:Nandri_%28%E0%AE%A8%E0%AE%A9%E0%AF%8D%E0%AE%B1%E0%AE%BF%29.png
- http://about.me/SathiaNMahadevan
- http://pixabay.com/p-96286

Creative Commons Attribution Share-Alike License