




2ND LEAD (UPDATED)

Transcription: TamilNet thanks responses

[TamilNet, Wednesday, 16 May 2007, 01:59 GMT]

Three weeks ago, TamilNet launched a transcription system to present Tamil text to English readers. The system was developed with automation facilities to convert any Tamil text into Roman alphabet and vice versa. Recently, the Tamil Lexicon Transliteration system was also added for automation, making a three-way conversion possible. The aims of TamilNet Transcription, readers' comments and our observations are brought forth here for further improvements.

Several readers have come out with varied comments about the TamilNet Transcription. We thank all of them. A log of the responses are presented herewith to pursue open discussion to assess and improve the transcription attempt.

 [Use PDF if you are unable to view Tamil words.](#)

Serious concerns were raised by some readers that the transcription doesn't justify the phonemic variations of the Tamil alphabet. Given below are the outlines sent by B. Muthukumar, on the rules of Tamil pronunciation:

Valli:na alphabets:

Surd: pronounced as *k ch d th p 'r*

1. when they begin a word
2. when they double

Exception: Sanskrit words may begin differently.

Sonant: pronounced as *g s d dh b 'r*

3. when a vowel is in between.
4. when it is next to *melli:nam* or *idaiyinam*

Exception: *g -> h; s -> j*

If we illustrate his rules with examples:

1. கரு, சட்டி, தயிர், பல்லி
(*karu, chaddi, thayir, palli*)
2. பக்கம், அச்சம், கட்டி, சத்தம், அப்பா, வெற்றி
(*pakkam, achcham, kaddi, chaththam, appaa, vetti*)

Exception:

கங்கை, சமன், தெய்வம், பாலன்
(*gangai, sama:n, dheyvam, baala:n*)

The exception can be avoided to de-sanskritize, according to Muthukumar:

(*kangkai, chama:n, theyvam, paala:n*)

3. அகம், அரசன், படம், மதம், அபலை, கறுப்பு
(*agam, arasa:n, padam, madham, abalai, ka'ruppu*)

4. வங்கம், அம்சம், வண்டி, பந்து, இன்பம், தென்றல்
(*vanggam, amsam, va'ndi, pandhu, i:nbam, the:n'ral*)

Exceptions:

k can become *g* as well as *h*: அகம் (*agam / aham*) *ch* can become *j* instead of *s*: வஞ்சம் (*vanjam*)

We may also add that such phonemic variations are found in vowels too:

1. a, ae
(அம்மா, அரைப்பு)
(*ammaa, aeraippu*)
2. i
இலை, இழு
(*ila, izhu*)
3. e, ə
எச்சம், எறும்பு
(*echcham, 'rumpu*)
4. The short *i* (இ) at the end of words(*kuttiyalikaram*)
5. The short *u* (உ) at the end of words(*kuttiyalukaram*)

Also note the pronunciation variations of *n* when it accepts vowel and when it occurs as a consonant next to *th*:

நட்டம், ஆந்தை (*naddam, aanthai*)

Another reader, Sinnathurai Srivas, doesn't agree with the term 'sanskritized pronunciation'. He argues that such sounds are very much Tamil and should be accommodated in the transcription. He points out மகள் should be transcribed as *maha'l* and not as *maka'l*.

The TamilNet wishes to present the following issues in standardizing the transcription on the above lines, for the opinion of the readers:

Talking of the dual pronunciation of *valli:nam* alphabets, *k* doesn't universally become *g* in all dialects when it follows *melli:nam* or *idaiyi:nam*.

In words such as தங்கம், தங்கச்சி, வங்கம், வங்களாவடி(*thangkam, thangkachchi, vangkam, vangka'laavadi*) etc., the *k* is largely retained in Sri Lankan Tamil, especially in the Jaffna dialect. The accent makes the difference.

It is not uncommon to see people pronouncing *gangai* as *kangkai*. Also, look at words such as கங்குமட்டை (*kangkumaddai*).

We do agree that *k* is often pronounced as *h*, coming under rule number 3 (exceptions) of Muthukumar and as pointed out by Srivas. A word such as நகர் (*nakar*), can be pronounced as *nagar* as well as *nahar*. The former is favoured by those who are familiar with Sanskrit pronunciation and the latter largely prevails in Sri Lankan Tamil. In the example of மகள் (*maka'l*), it is pronounced as *maga'l* in the northern districts of Tamil Nadu where Kannada overlaps with Tamil and pronounced as *maha'l* in Sri Lanka. Which one has to be taken if we aim for standard automation?

It is interesting to note that the letters for *s* and *h* are frequently found in Tamil Brahmi, ie. the alphabet of the earliest written Tamil. But, when the occasion came to write the word மகள் in the early Brahmi inscriptions, it was written as *maka'l* and not as *maha'l*, despite the fact that the script for *h* was very much known to Tamil Brahmi.

Tamil Brahmi was not unaware of the letters used for a variety of sounds in Ashokan Brahmi. Yet, for some reason, the Tamils thought of restricting the alphabet to 30 and evolved a parallel Grantha Alphabet to write Sanskrit.

Similar to the example of *k*, two other *valli:nams th* and *p* are also largely retained in Sri Lanka due to accent, eg., பந்து, கம்பர், எண்பது, என்பது (*panthu, kampar, e'npathu and e:npathu*).

We wish to place it to the readers whether it is adviceable to be partial by standardizing only one way of pronunciation. Is it proper to discourage people who retain the *k*, *th*, and *p* sounds of the *valli:nam* alphabets? Perhaps it is better to leave it open and stick to the alphabets in certain instances, even though it may not be the perfect way of transcription. Non Tamil readers can be made aware of such discrepancies through pronunciation key and through audio transcription by native speakers, which we aim for future.

TamilNet totally agrees with instances where *ch* becomes *s*. Eg. பசி, காசு (*pasi, kaasu*) etc. We promise to make adjustments in the transcription in such a way that this *s* will not be confused with the Grantha *s* in automation. However, *ch* sounding *j* is not universal. The sound *ch* is retained in such instances in Sri Lanka: பஞ்சு, கஞ்சி, வஞ்சனை (*pagnchu, kagnchi and vagncha:nai*).

Rev. Dr. D. S. Dharmapalan suggested to use *ae* instead of *ea* for the long vowel ஏ. There are no technical difficulties in accepting this. However, we felt that the Tamil writers might find it odd to begin with *ae* for ஏ, because of their familiarity in using *e* for this vowel. Besides, an English reader has a different pronunciation for *ae*. We invite opinion.

Nalliah Sivarasan advocates *nj* instead of *gn* for the letter ஞ. While agreeing with its suitability, we have to point out that the *j* in *nj* has to be differentiated from the Grantha *j* to implement automation. *Gn* is conventional and understood by many. It is better not to use two *j*'s unless it is absolutely necessary. We once again leave it to the readers for further opinion.

R. Sri Ranjan and Marcil Francis have presented systems designed by them with alternatives. Readers are requested to evaluate the pros & cons and to come up with suggestions.

TamilNet wishes to reiterate that its transcription attempt is primarily for simple journalistic purposes. We don't actually mean to get deep into the phonemic subtleties of Tamil language. For serious academic purposes, there are transcription systems using non Roman scripts and symbols. Our attempt may not linguistically present itself as a perfect transcription. It may perhaps be viewed as a compromise between transcription and transliteration. We make no qualms about it as long as it serves the purposes.

An English website such as TamilNet, covering Tamil affairs, has to invariably deal with Tamil texts in its day to day transactions. How to deal with them in Roman script in a standard way, with a simple keyboard, rendering near accurate pronunciation to an English reader, and at the same time maintaining its link to the Tamil script, is our concern.

For instance, an island name off Jaffna is currently written in English as *Karainagar*. Anyone, including a Tamil, who doesn't know the place can always understand it as 'coastal town', since *karai* is coast. It can also be read with a retroflex 'r to mean, 'blot'. The exact word here is *Kaarai*, a natural vegetation from which the island got its name. The issues related to the transcription of the second part of the place name have been discussed earlier.

Our next priority is to automate the text for typesetting and to convert it between Tamil and Roman script. Recently, we have added the conventional Madras Tamil Lexicon Transliteration system too, for automation. It will be a three-way conversion system when we finalise the software.

We are hopeful that the system will help not only journalists, authors and printers, but also any layman using computer for private purposes in Tamil. The members of the diaspora who couldn't get an opportunity of learning the Tamil alphabet may also find it useful.

The third concern is storage and retrieval which is of utmost importance in news reporting. The TamilNet Transcription aims at simplifying this exercise for the users of even elementary means of electronic typesetting, communication and transaction.

Many readers have initially asked questions on how to use the system. We request the readers to use

Bamini font for the time being to enter text in Tamil. The facility will be extended to Unicode eventually. At present, if you enter text in Roman, in the way suggested in our transcription system, it will be converted to Tamil script in Unicode, even if you don't have Bamini.

Gregory Mora enquires about simultaneous display. Once the system is finalised, the algorithm for conversion will be released to the public. Software developers may bring in many user-friendly features, including simultaneous display and chat.

Once again, we thank all our readers for the encouragement and look forward to see further participation.

Find this article at:

<http://www.tamilnet.com/art.html?catid=13&artid=22166>

Reproduction of this news item is allowed when used without any alterations to the contents and the source, TamilNet, is mentioned

[Email](#)[Close](#)