Tamil-Encoding Issues

L2/03-448

Date : 15th Dec 2003 Issue : 5

Proposal for Tamil in Unicode

Previous issues of the newsletter analysed both existing Tamil Unicode and ACE (All Character Encoding) scheme, the proposal submitted by Government of TamilNadu. Relevant issues of the Tamil Unicode have been discussed in detail. While ACE scheme has technical deficiencies which do not render it to qualify the principles, guidelines and stability policies of Unicode Consortium, the existing Unicode has flaws in it. Thus, in order to resolve the issues of Tamil Unicode, an alternate proposal has been worked out. This issue of the newsletter details the proposal on hand. This proposal also includes a set of recommendations to ensure correct implementation of the proposal in Basic Multilingual Plane of Unicode. This issue also contains the justifications both from the technical and linguistic perspectives for the current proposal.

Proposal

At the very outset, it is worth recalling the fundamental requirement of Unicode design that proposals made to it should be based on the script. This aspect has not only been capitulated, but has been analysed in detail in the previous issues of the newsletter.

The other considerations that constitute the basis in arriving at this proposal are:

1. The stability policies of the Unicode.
2. Related documents of ISO dealing with the principles and procedures for allocation of new characters and scripts and handling of defect reports on character names.

This scheme is being proposed to resolve the issues related to Tamil script in the existing Unicode. It achieves an optimal balance between the different contrasting issues. In a gist, even while retaining the fundamental structure of the existing Unicode intact, the current proposal has overcome the anomalies of the existing Unicode by encoding the consonants in the vacant positions that currently exist in the Tamil block of Unicode. The proposed characters have been highlighted in the table below.

	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0		ஐ		ர	ீ	ங்	ற்	௰
1				ற	ு	ச்	ன்	௱
2	ஂ	ஒ		ல	ூ	ஞ்	ஜ்	௲
3	ஃ	ஓ	ண	ள		ட்	ஷ்	௳
4		ஔ	த	ழ		ண்	ஸ்	௴
5	அ	க		வ		த்	ஹ்	௵
6	ஆ				ெ	ந்	க்ஷ்	௶
7	இ			ஷ	ே	ௗ	௧	௷
8	ஈ		ந	ஸ	ை	ப்	௨	௸
9	உ	ங	ன	ஹ		ம்	௩	௹
A	ஊ	ச	ப	க்ஷ	ொ	ய்	௪	௺
B					ோ	ர்	௫	ஸ்ரீ
C		ஜ			ௌ	ல்	௬	ௐ
D					்	வ்	௭	
E	எ	ஞ	ம	ா		ழ்	௮	
F	ஏ	ட	ய	ி	க்	ள்	௯	

Principal advantages of the Proposal

The principal advantages that arise out of implementation of the current proposal are:

1. Meets the requirement of Tamil Grammar.
2. Makes it possible to accommodate the basic Linguistic character set within the existing block.
3. Follows the Tamil writing system.
4. Fits within the existing standard for Tamil.
5. Advantage of compression is retained.
6. Meets the requirement of the Stability policy of Unicode.
7. Meets the requirement of Principles and Procedures for Allocation of New Characters.

Disclaimer : This Newsletter is Composed, Edited and Published by N. ANBARASAN on behalf of Tamil-Encoding for limited circulation. This Newsletter has been prepared using sources believed to be reliable and accurate. Tamil Encoding does not accept any direct or consequential use of this Newsletter or its contents. Should you need any clarification or would like to offer an opinion on the content, please contact arasan@bgl.vsnl.net.in. If you would like to participate in the Discussion on Tamil Unicode, please visit http://groups.msn.com/Tamil-Unicode to join Tamil-Unicode.

Technical Justifications

This proposal derives its substrate based on the conditions documented in the ISO document entitled, "Principles and Procedures for Allocations of New Characters and Scripts and handling of defect Reports on Character Names". Its section, "**WG 2 Evaluation Procedure**" and sub-clause, "**Encode on the Basic Multilingual Plane (BMP)**" pertains to criteria for proposed characters to qualify for BMP. According to this criteria, the proposed characters qualify for BMP as they fulfill the following conditions:

1. If the proposed character is part of a well-defined character collection not already encoded in the standard.
2. If the proposed character is part of a small number of characters to be added to a script already encoded in the Basic Multilingual Plane (for example, the characters can be encoded at unallocated code positions within the block or blocks allocated for the script).

Analysis proves that the proposed Tamil consonants are in accordance with the above conditions of Unicode and the justifications are:

a) The Tamil consonants proposed for BMP are well defined in Tamil grammar.
b) They are being proposed to be encoded in the BMP in the existing vacant slots within the already encoded Tamil Block in Unicode.

Apart from meeting the above conditions, the current proposal also satisfies the 'Positive criteria' of "Existence of a precomposed letter in a well-established or official alphabet" contained in sub-clause of "Principles and Procedures for Allocations of New Characters and Scripts and handling of defect Reports on Character Names" of Annex G, "Formal criteria for coding precomposed characters".

Linguistic Justifications

The justifications are as follows:

1. The consonants in Tamil are written with a dot on top. It has not been formed by a combination of அ 'a' consonant and a 'pulli' (dot). This has been documented in "Tholkappiyam", our ancient Tamil Grammar in its verse 15:

மெய்யின் இயற்கை புள்ளியோடு நிலையல்.

meyyin iyarkai pulliyodu nilaiyal

2. In Tamil, the 'a' vowelised consonants (க,ங,ச etc) forms the basis for writing all the remaining vowel-consonants. The vowel-consonant itself takes different shapes based on the joining vowel by extending the base consonant. This has also been documented in "Tholkappiyam", in its verse 17:

புள்ளி யில்லா எல்லா மெய்யும்
உருவுரு வாகி அகரமோ டுயிர்த்தலும்
ஏனை உயிரோடு உருவுதிரிந் துயிர்த்தலும்
ஆயீ ரியல உயிர்த்த வாறே.

pulli yilla ella meiyum
uruuru vaagi agaramo duyirthalum
eanai uyirodu uruvuthirin thuyirthalum
aayie riyala uyirtha vaare

All consonants are summarily primary characters which have 'pulli' (dot) on top. Thus, the vowel-consonant written without the dot, as mentioned in the above verse of Tholkappiyam has been derived by a special process, 'deleting action'. This action removes the dot. This comes to prove that by default, the consonants are with 'pulli' (dot) and that the consonants do not acquire this dot by a special process for 'addition' of a dot.

3. Another treatise of Tamil Grammar, Nannool has also documented the formation of Vowel-consonants (Uyirmei letters) in its verse 89.

புள்ளிவிட்டு அவ்வொடு முன்னுரு வாகியும்
ஏனை உயிரோடு உருவு திரிந்தும்
உயிரள வாய்அதன் வடிவொழித்து இருவயின்
பெயரொடும் ஒற்றுமுன் னாய்வரும் உயிர்மெய்.

pullivittu avvodu munnuru vaagiyum
eanai uyirodu uruvu thirindhum
uyirala vaayadhan vadivozhithu iruvayin
peyarodum otrumun naaivarum uyirmei

This also proves that the base consonant without dot is formed by omitting (deletion) the 'pulli' (dot) and this functions as the base letter to form uyirmei letters. Hence forming the consonants by adding the 'pulli' (dot) is against the Tamil grammar. Moreover, representation of a single character with two encoded characters which have two distinct encoded values, of which the TAMIL SIGN VIRAMA being a non-character is a fundamental mistake.

Recommendations

The following recommendations are being made along with the current proposal, to achieve correct implementation of the current proposal in Unicode.

1. The proposed characters should be encoded as canonical equivalence to the composition characters.

This recommendation is being made with a view to usher in uniformity in treatment of 'character sequence in the existing standard' v/s 'currently proposed precomposed characters' in all ways.

2. Normalise the consonants formed out of the existing character sequence to the precomposed consonants.

In the wake of the above recommendation, it is worth analyzing one of the guiding principles of Unicode as documented in document N2352R titled "Principles and Procedures for Allocation of New Characters and Scripts and handling of Defect Reports on Character Names". This document consists of principles and procedures pertaining to preparation, submission and handling of proposals for addition of characters to the repertoire of the standard.

The current analysis pertains to Annex - G of the above document. Unicode has set to itself a policy of not allowing Normalisation of any new characters that would get added to Unicode standard after Unicode version 3.0 which has been defined by Unicode as being the 'composition version' for reasons of stability. Now, coming to the Unicode standard 3.0 for Tamil, the Unicode is deficient in itself and erroneous too from both the linguistic and technical percepts.

From the linguistic point of view, Unicode 3.0 violates the very grammar of Tamil language from the percepts of Tamil Script and the writing system. Unlike other Indian languages Tamil writing system has a different approach for script. But Unicode 3.0 treats Tamil writing system in the same way as it does with the other Indian scripts and has introduced errors. The errors on these percepts have been listed below:

a) It allows a character (Tamil Consonants) to get formed with a non-character pulli encoded as (TAMIL SIGN VIRAMA) .
b) Unlike in other Indian scripts, Tamil script does not have Virama (pulli) nor does it have any functionality in the Tamil writing system. It doesn't function as Vowel omission sign as in the case of other Indian scripts while forming Orthographic syllables with consonant clusters. Hence, the encoded TAMIL SIGN VIRAMA is not recognized as a character.
c) Tamil writing system employs the consonants in full form as against the other Indian languages where Orthographic syllables formed for the consonant clusters.

Now, it is worth observing that this policy, as it stands would not accommodate even the genuine requirement that has arisen in this case.

But, implementation of the current proposal alongside the already existing Unicode would introduce multiple spellings. This aspect has to be taken care of by the process of Normalisation, which if not accomodated would result in the proposal qualifying the criteria for rejection. This is an ambiguous situation which has to be solved by appropriately accommodating the Normalisation of these characters.

Submission of proposals for inclusion of new characters is governed by the ISO document N2352R titled "Principles and Procedures for Allocation of New Characters and Scripts and handling of Defect Reports on Character Names". On the one hand, at section G.1, 'Criteria', the 'positive criteria' for inclusion of a precomposed letter to Unicode, it records the criteria for inclusion of new characters. It records that the new character has to be an established or an official alphabet.

On the other, Unicode has set to itself a policy of not allowing Normalisation of any new characters that would get added to Unicode standard after Unicode version 3.0 for reasons of stability. This policy is also listed in the same document. This policy, as it stands would not accommodate even the genuine requirement that would arise thereafter even if based on sound linguistic and/or on technical percepts and leading to introduction of multiple spellings. Such a situation would need Normalisation, lest the proposal would qualify the criteria for rejection. This rigidity in policy would continue to promote errors and imperfectness as no proposals for new characters would be accepted.

3. Deprecate the use of TAMIL SIGN ANUSVARA.

This recommendation is being made on the basis that this character does not exist in Tamil. It is wrongly encoded on the basis of ISCII, which is a common script standard for all Indian scripts.

4. Encode the grantha letter 'ksha' which is being formed out of sequence of encoded characters.

This recommendation is being made on the following basis:

a) 'ksha' exhibits the characteristics of a character. 'ksha' combines with allographs of vowels like 'i' to form 'kshi' as in 'Meenakshi'.
b) Tamil writing system does not have any Ligature formation for consonant clusters.

5. Encode 'Shri' and 'Om' as symbols.