L2/06-229

Dated at Aruppukottai (S.India) the 3rd July 2006.

From,

V.Ramasami,
(eMail ID <vgr_ramasami@sancharnet.in>),
(Ph: +91-4566-221303)
Retired Telecom Engineer,
38, PichakuttyNadar Street,
SBK Elementary School Road,
ARUPPUKOTTAI,
TamilNadu State,
South India,
PIN: 626101.

To,

The Unicode Consortium,
Attn. Magda Danish,
1065,L'Avenida Street,
MicroSoft Building 5,
Mountain View,
CA 94043,
USA.

Sirs,

Sub: Proposal to update the codepage for the Tamil Script, with downward compatibility.
Ref: Code Block from U+0B80 to U+0BFF.

I'm born and bred in the heartland of TamilNadu, with knowledge of Tamil from birth. I had my schooling upto the Secondary stage in Tamil, covering all subjects. Apart from English, I can also read, write and speak Hindi. I've sufficient knowledge about fonts, their rendering in VDUs and data entry. I've sufficiently understood the Tamil Code Page standardised by the Consortium. I've sufficient knowledge of other encodings and renderings of the Tamil Script. My work on "Tamil Type As you Write" Vutam font(s) is available at :
http://sarovar.org/projects/keytrans

The Consortium has taken much efforts to form two groups out of the Indic Scripts. Tamil Script is a member of the South Indian subset of scripts. To mitigate the deficiencies arising out of such groupings, the Consortium has also to make provisions to cater to the individual charecteristics that differentiate the set members. Provisions have to be made for the commonalities as well as the differentialities of the member scripts.

However, it can't be said that the effort has left no stones unturned. In the areas identified in the PROPOSAL (enclosed), improvements can still be done to cater to the dynamic evolving nature of the Ancient Classic Language, Tamil.

I'm enclosing a PROPOSAL for updating the Tamil Code Page by allotting additional codes for a number of precomposed charecters (consonants with vowel signs), so that Tamil can be input into the computing machine as it is normally written by millions and to mitigate the difficulties faced by software professionals.

A PROPOSAL SUMMARY FORM in the prescribed format is also enclosed, in line with the PROPOSAL mentioned above.

In the PROPOSAL, while a glyph is referred as G+ followed by a decimal number, a code point is referred as U+ followed by a hexadecimal number.

Thanking You,

Yours,

V.Ramasami.
:

Encl:
(1) Proposal : Coding of "precomposed" charecters & Declaration of canonical equivalance.
(2) Prescribed Proposal Summary Form.
(3) Many Tamil glyphs in a bmp file. Numbers indicate G+ num in font file.
(4) LATHA.OTF of MicroSoft for aiding recognition of code points and glyphs.

ENCLOSURE (1)
=============

PROPOSAL : Coding of "Precomposed" Charecters & Declaration of canonical equivalance.
==========

(1) Any well written document of any format has its own humble beginning with paper and pencil of the Author. At a decent level of presentability, the manuscript is again manually typed into an editor of a computing machine by a Data Entry Assistant with knowledge in that Language and script. From this time of data entry, the Consortium's effort in the form of standards comes into play. Documents made on the fly have their own levels of acceptability, say, as Mails.The touch tablets and text converting SW of scanners, for inputting data, also aid the writing system.

To improve the efficiency of the Data Entry Assistant, it is imperative to make him spend his time only on issues related to the symbol to symbol transfer of the manuscript into the keyboard. So it is essential that there is no variation between the script in the manuscript and the script that is typed in the keyboard. While it is laudable to make provisions in standards for a simple to learn, futuristic, logical and evolutionary trend set of keyboard inputting, such a provision should not be an impediment by being an alien way to the currently prevalent way of writing the script. Again, it is but normal for the Assistant to expect the same thing to appear before him in the VDU, whatever he has currently typed in the keyboard. If he is "used to accept" a different rendering, there is an impediment to his efficiency.

In Tamil Language that uses the Tamil Script, vowel-marks/signs occupy the post positions, pre positions and both pre&post positions relative to the concerned (base form) charecter, while writing manually. U+0BBE (G+88), U+0BBF (G+89), U+0BC0 (G+90), U+0BC1 (G+91), U+0BC2 (G+92) and U+0BCD (G+99) are post positioned vowel marks; U+0BC6 (G+93), U+0BC7 (G+94) and U+0BC8 (G+95) are pre positioned vowel marks; U+0BCA (G+96), U+0BCB (G+97) and U+0BCC (G+98) are pre&post positioned (surrounding) two-part vowel marks. In Unicode, it is essential to type ALL these posthumously, which is an impediment to the Data Entry Assisstant's efficiency, as explained in the previous paragraph.

However, using a Keyboard driver that achieves even in a round about manner, what a simple such driver achieves in implementing the Transliteration Scheme, a conventional way of inputting can still provide the Unicode's requirements and hence is not an issue here.

The existing posthumus marking scheme is eminantly suitable to input data of ANY script, by the English Transliteration scheme, (obtaining a syllabic output, using the alphabetic input method) and hence is to be retained.

(2) It is every ones knowledge that Tamil is one of the most ancient Classic languages. Normally such Languages are rigid and unaccomodative. But Tamil Language's vibrant nature is to be seen to be believed !!

Tamil accomodated Sanskrit within it, by increasing its Base Form Consonant count from 18 to 23, which results in an actual increase of 5 X 12 = 60 consonant graphemes. It uses modern punctuation marks aplenty, as if they were its own. European numerals are part and parcel of it, much to the chagrin of the purists. The use of U+0B83 (G+53) to denote foreign sounds is increasing day by day. Its abbreviation (symbol) list from U+0BF3 to U+0BFA includes, apart from its own, Sanskrit (U+0BF5) (G+189) and English (U+0BFA) (G+200) ones. Even the modern Indian currency symbol (U+0BF9) (G+199) is included in it. A large number of symbols denoting fractions are not yet coded by the consortium. Very many vowel marked consonants in the form of graphemes consisting of allographs (vowel marks) and base forms, have been given individual shapes to save on "time and distance". Some single charecters have even been given a "face lift". The list will be endless if one wishes to include everything that is evolving in this language and script.

(i) Graphemes to single (precomposed) charecters: (glyphs in LATHA of MS)

(1) ALL 'U' (G+91) and 'UU' (G+92) vowel marked 18 Tamil consonants (not the 5 Tamilised Sanskrit ones).
(2) Tamilised Sanskrit consonant ksha (G+113) and the Sanskrit symbol (abbreviation) srii (G+188).
(3) U+0BB1 (G+80), U+0BA9 (G+75) and U+0BA3 (G+72) with U+0BBE (G+88) 'AA' vowel marked consonants.
(4) U+0BB2 (G+81), U+0BB3 (G+82), U+0BA9 (G+75) and U+0BA3 (G+72) with U+0BC8 (G+95) 'AI' vowel marked consonants.
(5) U+0B9F(G+71) and U+0BB0 (G+79) with U+0BBF (G+89) 'I' vowel marked consonants.
(6) U+0B9F(G+71) and U+0BB0 (G+79) with U+0BC0 (G+90) 'II' vowel marked consonants.
(7) U+0B9F(G+71) and U+0BB0 (G+79) with U+0BCD (G+99) virama (dot or pulli) ie., the Halant form..

(ii) Aesthetically "Face Lifted" charecters:

U+0BA8 (G+74), U+0BAF (G+78) and U+0BB1 (G+80).

U+0BA8 : Tamil Letter NNA. Modified one (G+208) looks like a tail trimmed dog.
U+0BAF : Tamil Letter YA. Modified one (G+209) looks like a tea kettle with a trunk like pour-out.
U+0BB1 : Tamil Letter RRA. Modified one (G+207 or G+177) has a constriction in its right shoulder.

In the above list, EXCEPT (i) (1), (2) (3) & (4), the rest can be considered as "minor variant forms" of existing "combining charecter sequences".

(i) (3) & (4) can be considered as having equivalent alternate writing forms.

The PROPOSAL is for allotting independant code points for the precomposed charecters that come under (i) (1) & (2) only.

JUSTIFICATION FOR (i) (1)
===================

a) These are neither ligatures nor digraphs; these are "substitutions" in LATHA.TTF of MicroSoft, since their presentation forms do not come anywhere near their "combining charecter sequence(s)". These are rather "major variant forms". These 36 consonants presently require Uniscribe for their rendering as substitutions.

b) In the day to day writing system, these are written in "one go" as against 2 or 3 distinct efforts for writing the other consonants with either the "one part" vowel sign or the "two parts" vowel sign, respectively. Thus forcing the Data Entry Assistant to type twice, first for the base fom and second for the vowel sign, to represent a "one go" charecter & glyph, is a serious impediment to his efficiency. In fact the two vowel signs concerned in (i) (1) (U+0BC1/G+91 and U+0BC2/G+92) are used in the writing system only for their five Tamilised Sanskrit base form consonants (G+69, 85, 86, 87, 113), and never for the 18 Tamil base form consonants (G+66, 67, 68, 70 to G+84). These are nothing but processing issues in the writing system.

(i) (1) addresses the following 18 base form consonants, that are followed by the 'U' (G+91) and 'UU' (G+92) vowel signs, requiring 36 new code points for their precomposed forms :

U+0B95 (G+66); U+0B99 (G+67); U+0B9A (G+68); U+0B9E (G+70); U+0B9F (G+71); U+0BA3(G+72); U+0BA4 (G+73); U+0BA8 (G+74); U+0BA9 (G+75); U+0BAA (G+76); U+0BAE

(G+77); U+0BAF (G+78); U+0BB0 (G+79); U+0BB1 (G+80); U+0BB2 (G+81); U+0BB3 (G+82); U+0BB4(G+83) & U+0BB5 (G+84).

The existing way of typing these in two parts may also remain for donward compatibility.

Glyphs: ALL are available in LATHA (Enclosure 3).

G+66	G+91	-> G+148	(Requires its own code point)
G+66	G+92	-> G+149	(DITTO)
G+67	G+91	-> G+150	(DITTO)
G+67	G+92	-> G+151	(DITTO)
G+68	G+91	-> G+152	(DITTO)
G+68	G+92	-> G+153	(DITTO)
G+70	G+91	-> G+154	(DITTO)
G+70	G+92	-> G+155	(DITTO)
G+71	G+91	-> G+158	(DITTO)
G+71	G+92	-> G+159	(DITTO)
G+72	G+91	-> G+160	(DITTO)
G+72	G+92	-> G+161	(DITTO)
G+73	G+91	-> G+162	(DITTO)
G+73	G+92	-> G+163	(DITTO)
G+74	G+91	-> G+164	(DITTO)
G+74	G+92	-> G+165	(DITTO)
G+75	G+91	-> G+167	(DITTO)
G+75	G+92	-> G+168	(DITTO)
G+76	G+91	-> G+169	(DITTO)
G+76	G+92	-> G+170	(DITTO)
G+77	G+91	-> G+171	(DITTO)
G+77	G+92	-> G+172	(DITTO)
G+78	G+91	-> G+173	(DITTO)
G+78	G+92	-> G+174	(DITTO)
G+79	G+91	-> G+175	(DITTO)
G+79	G+92	-> G+176	(DITTO)
G+80	G+91	-> G+178	(DITTO)
G+80	G+92	-> G+179	(DITTO)
G+81	G+91	-> G+180	(DITTO)
G+81	G+92	-> G+181	(DITTO)
G+82	G+91	-> G+182	(DITTO)
G+82	G+92	-> G+183	(DITTO)
G+83	G+91	-> G+184	(DITTO)
G+83	G+92	-> G+185	(DITTO)
G+84	G+91	-> G+186	(DITTO)
G+84	G+92	-> G+187	(DITTO)

Naming the 'U' and 'UU' marked 36 precomposed consonants: Remove "A" from the base form's name, and instead add the vowel sign's name.

Eg.: Base form's name : Tamil Letter KA. Vowel sign's name: Tamil Vowel Sign U.
Proposed charecter name: Tamil Letter KU.

Codes: To be allotted in the existing vacant code points of the Tamil Block, by the consortium.

JUSTIFICATION FOR (i) (2)
===================

(i) (2) is concerned about Tamil base form consonant ksha (G+113) and Tamilsed Sanskrit symbol srii (G+188), which are presently coded sequences, ksha (G+113) = U+0B95(G+ 66);

U+0BCD (G+99) ;U+0BB7 (G+85) and srii (G+188) = U+0BB8 (G+86); U+0BCD (G+99); U+0BB0 (G+79); U+0BC0 (G+90).

In modern society, one's success in life is measured by his economic success, rightly or wrongly. So the Goddess of Wealth preoccupies every mind. This is reflected in the names one chooses for the wards and for the business enterprises. In this aspect, Tamil society is no different from the rest. Thus, every alternate girl bears the name of the Goddess of wealth, called "Lakshmi" in Tamil, and every alternate enterprise bears the name of the Goddess of wealth, also known as "Srii" in Tamil. Girls and boys with beatiful eyes require this 'ksha' in their names like "Meenaakshi", "Kaamaakshi", "Pankajaakshan" etc.. Ruling class identify themselves as "Kshatriya". There are many other words and proper names requiring this ksha and srii. "ksha" and "srii" presently being a sequence of charecters to be substituted as a single glyph, require Uniscribe for their rendering.

It should be noted here that ksha is a base form consonant in Tamil and written in one go rather than being written as a three charecter sequence. It takes ALL 12 vowel signs of Tamil, including its Halant form. srii being a symbol is written in two goes, the first go for the symbol without its 'II' vowel sign and the second go for the 'II' vowel sign. Since being not a charecter, the first part has no meaning (as understood by its users) without the second part. The first part does not take any other vowel or other sign for any other representation. Thus for all purposes of the writing system, srii is also a one go symbol, deserving its own coding as a precomposed charecter.

Glyphs: Available in LATHA (Enclosure 3).

ksha (G+113)	(Requires its own code point)
srii (G+188)	(DITTO)

Coding & Naming: It is suggested that U+0BBA be assigned for the base form consonant "ksha" (named Tamil letter KSHA) and U+0BFB be assigned to "srii", being an abbreviation (named Tamil SRII symbol). Existing substitution scheme also to remain, for sake of downward compatibility.

(3) NOTES:
=====

Note 1):

Evolutions that have occurred in the Tamil Language without affecting its script have not been brought out in the PROPOSAL, as being grammatical.

Note 2):

In the discussions in the Tamil Yahoo and Google groups, opinion is for providing individual code for every one of the 23 X 12 = 276 Tamil consonants and six alternade codes for the prepositioned vowel signs, for the following advantages:

1) Saves on storage space. Use of CGJ will further increase the bulk.
2) Avoids wrong searches.
3) Counting the number of charecters and reversing the charecter string is much easier.
4) Restoring correct data from corrupted data, corrupted in storage due to efflux of time, will give reliable results.
5) Easy sorting possible.
6) Natural way of writing the script is maintained, with all its advantages in the evolution process.

However, taking into account the WG2's default stand on the issues concerned, PROPOSAL is limitted to its present form.

If finding code points is the only issue in the way of the WG2, the off beat idea of keeping a common pool of codes (canonical block) for such purposes for a group of scripts could be considered, in which a single code point is allotted to all similar "phoneme representing graphemes" of the member scripts. This will however render Unicode un-unique; font files with multi-scripts of a family or group may not be possible to be created. However, those with multi-linguistic interest in specified groups will welcome this off beat suggestion of canonical allottment of codes.

Note 3):

Proving these 276 graphemes as "charecters" is quite easy. Just note how these are written vertically from top to bottom, by its users. These are always written a grapheme as a whole, and never as base forms and allographs in two or more pieces. As an example note how "directer's view" in Tamil is written in the popular SUN TV program. Other examples are "tiffin ready" in Tamil, in front of hotels, written vertically in narrow planks, etc.

In English, an equivalent example is, writing INN vertically, from top to bottom, using only marks, (||\||\|), like :

```
                    |
                    |
                    \
                    |
                    |
                    \
                    |
```

compared to :

```
                |
                |\|
                |\|
```

Vertical writing makes sense to its users only when the alphabet or the symbol charecters as the case may be, are written as a whole, and not in piecemeal, proving their status as charecters, irrespective of a theoretical definition otherwise!!

END OF PROPOSAL

```
================
```

ENCLOSURE (2)
=============

PROPOSAL SUMMARY FORM
========================

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from
http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** Coding of precomposed charecters & Declaration of canonical equivalance.
2. Requester's name: V.Ramasami.
3. Requester type (Member body/Liaison/Individual contribution): Individual contribution.
4. Submission date: 3rd July 2006.
5. Requester's reference (if applicable): eMail ID <vgr_ramasami@sancharnet.in>
6. Choose one of the following:
This is a complete proposal: YES
(or) More information will be provided later: ________ *NO*

This is a complete proposal.

B. Technical - General

1. Choose one of the following:
a. This proposal is for a new script (set of characters): NO ________
Proposed name of script: ________
b. The proposal is for addition of character(s) to an existing block: The proposal is for addition of charecters to an existing block.
YES
Name of the existing block: TAMIL
2. Number of characters in proposal: ________
18 + 18 + 2 = 38. (Thirty eight only) ________
3. Proposed category (select one from below - see section 2.2 of P&P document):
A-Contemporary.
A A-Contemporary ____ B.1-Specialized (small collection) ____ B.2-Specialized (large collection) ____
C-Major extinct ____ D-Attested extinct ____ E-Minor extinct ____
F-Archaic Hieroglyphic or Ideographic ____ G-Obscure or questionable usage symbols ____
4. Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document):Level 3. ________
Is a rationale provided for the choice? YES.
If Yes, reference: Vide the PROPOSAL writeup.
5. Is a repertoire including character names provided? Yes. Vide PROPOSAL writeup. ________
a. If YES, are the names in accordance with the "character naming guidelines"
in Annex L of P&P document? YES ________
b. Are the character shapes attached in a legible form suitable for review? Vide LATHA.TTF of MicroSoft, Enclosed.
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? ________ MicroSoft's LATHA .TTF
If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools

used: FCP4.5 of M/S High-Logic MS's LATHA or VUTAM from http://sarovar.org/projects/keytrans/vutam.zip

7. References:

a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? NO

b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? NO

8. Special encoding issues:

Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?Keying, Uniscribe, Transliteration, Searching and Restoration aspects dealt in text. YES

9. Additional Information: Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

If within my resources.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?
NO

If YES explain

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? YES

If YES, with whom? Yahoo and Google Tamil group members

If YES, available relevant documents:
By eMails.

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?
NO

Reference:

4. The context of use for the proposed characters (type of use; common or rare) Very common.

Reference:

5. Are the proposed characters in current use by the user community? YES

If YES, where? Reference: ALL printed and written documents already use these. Presence of the concerned glyphs in MicroSoft's LATHA.TTF is ample proof.

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely
in the BMP?
YES

If YES, is a rationale provided?

If YES, reference: Within the vacant points in the Tamil block.

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? Not necessary.

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?
YES

Yes, in text. If YES, is a rationale for its inclusion provided?

If YES, reference: Present way of writing requires these as individual charecters.

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? NO

If YES, is a rationale for its inclusion provided?

If YES, reference:

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? NO.

If YES, is a rationale for its inclusion provided?

If YES, reference:

11. Does the proposal include use of combining characters and/or use of composite sequences?NO

NO If YES, is a rationale for such use provided?

If YES, reference:

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? NO

If YES, reference:

12. Does the proposal contain characters with any special properties such as control function or similar semantics?

NO If YES, describe in detail (include attachment if necessary)

13. Does the proposal contain any Ideographic compatibility character(s)?
NO If YES, is the equivalent corresponding unified ideographic character(s) identified?

If YES, reference:

148 கு 149 கூ 150 ஙு 151 ஙூ 152 சு 153 சூ
154 ஞு 155 ஞூ 156 டி 157 டீ 158 டு 159 டூ 160 ணு
161 ணூ 162 து 163 தூ 164 நு 165 நூ 167 னு
168 னூ 169 பு 170 பூ 171 மு 172 மூ 173 யு 174 யூ
175 ரு 176 ரூ 178 று 179 றூ 180 லு 181 லூ
182 ளு 183 ளூ 184 ழு 185 ழூ 186 வு 187 வூ
188 ஸ்ரீ 99 ் 89 ி 90 ீ 91 ு 92 ூ 88 ா 95 ை 113 க்ஷ
53 ஃ 93 ெ 94 ே 189 ஹு 199 ஸ்ரீ 66 க 67 ங 68 ச
69 ஜ 70 ஞ 71 ட 72 ண 73 த 74 ந 75 ன 76 ப
77 ம 78 ய 79 ர 80 ற 81 ல 82 ள 83 ழ 84 வ
113 க்ஷ = 66 க + 99 ் + 85 ஷ
188 ஸ்ரீ = 86 ஸ + 99 ் + 79 ர + 90 ீ
148 கு = 66 க + 91 ு ; 149 கூ = 66 ,

200 நீ 96 ொ
97 ோ 98 ௌ

202 ஞு 166 ணு 177 ஙு 190 ரி 191 ரீ 203 லூ
204 னூ 205 ளூ 206 ணூ 207 ஙூ 208 ரு 209 ய