Unicode Expert committee - TVA
Minutes of the meeting held on 08-May-2020 through conference call

Subject: *Proposal for encoding Tamil characters RRA ఀ and LLLA ఁ into Telugu unicode as native characters*

Source: Proposal L2/20-119 in Unicode document register

Participants: Prof.Ponnavaiko, Dr.Krishnan Ramasami (a.k.a Iraamaki), Naga Elangovan, Dr.S.Rajavelu, Dr.Shriramana Sharma, Dr. James Anthony

The subject/issue was discussed in the meeting from the perspectives of History, Linguistics and Technical aspects in detail. Participants expressed various problems that this initiative may lead to not only for Tamil but also to the entire Indian multilingual context. Recommendations and solutions have also been discussed and agreed. Below are the Summary, Recommendation and complete Details of the discussion.

Summary and Recommendations:

i. It is technologically possible to represent the characters ఁ and ఀ without encoding them as new characters in Telugu block. As the author himself says, these natively Tamil letters may be added to ScriptExtensions.txt with appropriate font solution. There is no need to encode new characters.

ii. It is patently wrong to claim that Telugu RRA 'ṟ' (u+0C31) has a different phonetic realization to that of Tamil RRA 'ṟ' (0BB1). Even if different glyphs were seen in different inscriptions or in the same document, they are best treated as graphic/allophonic variants of the same phoneme.

iii. Pronunciation of ఄ among Tamil community varies between Srilankan Tamils and Tamilnadu Tamils. Srilankan Tamils pronounce ఄ properly while Tamilnadu Tamils pronounce it as ఩/ఀ which is wrong and continue to consider ఄ as phoneme joint. Unknowingly, the proposer of L2/20-119 wants to perpetuate a wrong pronunciation into Telugu, and he claiming an encoding for it is a simple fallacy.

iv. Duplicating native characters of one language into another in an Indian context will have long term negative impact on all the languages of India that has huge number of languages having close relations to one another. Transcribed texts with mixed

language charaters are the simplest cases that exist in every language of India with several versions. If Unicode Consortium start encoding every such mix, it will impact the Indian language structure adversely over the period of time.

Therefore, the proposed new characters ౝ and ౡ should not be encoded as native character of Telugu in Telugu block. Instead, it is recommended to go for the simple technology solution as discussed in here.

Minutes of the discussions of the expert committee

1) It is better to discuss this proposal document in detail along with earlier ones viz., L2/12-015, L2/12-016, L2/12-076, L2/L2-079, L2/13-121, L2/19-401, L2/19-405 and L2/20-085 in order to get the context clear to an uninitiated reader.

Historical and Epigraphical background:

2) First are a few preliminaries. The Telugu letter LLLA 'ఌ' was not in the Telugu or Kannada repertoire until it appeared in the proto Telugu-Kannada script through the Sālamkāyana inscription dated 5th century. Sālamkāyanas were vassals of the Pallavas of South India. Use of LLLA 'ఌ' continued in inscriptions till the time of Eastern Chalukya (Vēngi) dynasty king Rajaraja Narendra (11th century), the grandson of the Chozha Emperor Rajarajan. During this period, intense interaction existed between the South Indian Dynasties including the Western and Eastern Chalukya, Pallava, Chozha (both Renatu Chozha who were bilingual viz., Tamil and Telugu and Imperial Chozha who were Tamil speaking) and others. It was natural that the usage of LLLA 'ఌ' became prevalent in Telugu due to extensive communications in this age, but different glyphs were used in Telugu-Kannada and Tamil regions.
3) As for RRA 'ఋ', its use in the Telugu – Kannada territory also started around 5th century and continued rather long till 18th century, after which it was apparently discontinued in the Telugu speaking areas. Here also, there existed just one character but two different glyphs, one in Telugu-Kannada script region and the other in the Tamil script region.
4) However, both LLLA 'ఌ' and RRA 'ఋ' were not recorded in the earliest Telugu literature and grammar. This fact is learnt after going through the first Telugu Grammar "Andhra Shabda Chintamani." The great Adhikavi Nannaya Bhattaraka, who was the author of the first

mandala and the third mandala of the *Andhra mahabharatam*, a Telugu retelling of the *Mahabharata,* also wrote this Telugu grammar, but in Sanskrit language. There was no grammatical work in Telugu prior to it. [The first full-fledged grammar for Kannada, viz., Nagavarma II's Karnataka Basha Bhushana, was also written in Sanskrit around 12th century.]

5) Exactly similar to Telugu, Kannada literary works also employed the letters ಱ u+0C31 (transliterated 'r̲' or 'rh') and ೞ u+0C34 (transliterated 'l̲', 'lh' or 'zh'), whose manner of articulation most plausibly could be akin to those in present-day Malayalam and Tamil. The letters dropped out of use in the Kannada region around 12th and 18th centuries, respectively. Later Kannada works replaced 'rh' and 'lh' with ರ u+0CB0 (RA, 'r') and ಳ u+0CB3 (LLA, 'l̲') respectively.
6) Nannaya Bhattaraka ascribed to Telugu 36 tatbhava (indigenous) letters along with 19 imported letters to describe tatsama and foreign words. However, he did not include LLLA 'l̲', RRA 'r̲' and NNNA 'ṉ' at all in his collection. Even though the people in the region were aware of such sounds and the corresponding glyphs prevalent in their region, the scholars somehow did not include them in the literature repertoire, perhaps to emphasize priority to Sanskrit.
7) Prof Bhadriraju Krishnamurti in his book "The Dravidian Languages" (Cambridge University Press. 2003, on page 81) has recognized RRA 'r̲' as part of Telugu script but has excluded LLLA 'l̲' in its scope.

Various proposals submitted to UC for Telugu LLLA, RRRA and nukta:

8) However, this fact of medieval usage of LLLA 'l̲' and RRA 'r̲' in Telugu inscriptions/documents and later dropping them from use is of merit. It is perfectly legitimate and proper that they be encoded in Telugu to record historical information. RRA 'r̲' ఱ was encoded in the very first edition of Unicode Ver.1.0, while LLLA 'l̲' ఴ was encoded later in 2012 through the efforts of few individuals (including the present proposer of L2/20-119) raised through the proposal L2/12-015.
9) Subsequently, around Jan 2012, another proposal L2/12-016 was raised to establish an encoding of a letter called RRRA 𑌰𑍍𑌰 (u+0C5A) which was just a glyph variant of RRA 'r̲' ఱ. While the proposers presented the glyph through inscriptional evidence, they did not

corroborate the phonetic value. Had they obtained assistance from Telugu epigraphists, they could have averted a confusion. In page 2 of their proposal L2/12-016, it is stated:

'*Thus ఱ is a distinct character worthy of encoding. It is however clearly phonologically related to ఐ RRA/'r̲a' as seen in its usage in words like "mūnr̲u" (old Dravidian "three")* '.

10) "Mūnr̲u (three)" is also there in Tamil. On pursuing through the inscription along with an epigraphist by ourselves, it is clearly seen as a graphic variant of RRA ఐ. To claim that these phones (RRA ఐ and RRRA ఱ) are different phonologically, is not correct. Even in Tamil with words like "mūnr̲u (மூன்று)", "mur̲r̲am (முற்றம்)" and "mur̲ai (முறை)", not all Rs may sound the same on recording through an audiograph. Alveolar r̲ in "mūnr̲u" will sound a little voiced. r̲ in "mur̲r̲am" will sound a little aspirate. r̲ in "mur̲ai" will sound a little fricative. All these phones are allophones of the same phoneme. They are treated as a single phoneme in Tamil.
11) It is patently wrong to claim that Telugu RRA 'r̲' (u+0C31) has a different phonetic realization to that of Tamil RRA 'r̲' (0BB1). For all the three phones (plosive, aspirate, and fricative) of r̲ mentioned above one need not have opted for different characters in Telugu and hence new codepoints. Even if different glyphs were seen in different inscriptions or in the same document, they are best treated as graphic/allophonic variants of the same phoneme. In language computing, if there is a need, the usual practice to take care of the graphic variants of the same phoneme was through separate font development for each of the glyph with different phonetic values, and not through separate encoding.
12) It is quite surprising to know that an unknown phonetic value (RRRA) is encoded by UC. This raises confusion over the conventions of whether glyph or character encoding is practiced.
13) Out of the two glyph variants of RRRA shown in L2/12-016, one was recommended by its proposers. On reading these documents mentioned in the previous page with the assistance of an epigraphist, it appears to us that RRRA was superfluous and it was just a glyph variant of RRA. The matter could have been settled with further search in inscriptions and linguistics. The present reviewers are of the opinion that RRRA should have merited only a mention under RRA and not be given a separate codepoint 0C5A.

14) Around February 2012, Govt. of Andhra Pradesh (GoAP) through its agency AP Society for Knowledge Networks had accepted the proposal for LLLA 'ḷ' ఴ (u+0C34), and said that the proposal L2/12-015 is most welcome and this character should have been encoded in the Telugu block quite long time ago, but it remained as a desideratum. The letter L2/12-076 further said:

> *"The requirement of the letter is sometimes circumvented by using diacritics, sometimes obsolete Telugu symbol or The Tamil symbol which is readily available."*

15) The underlying meaning of the above statement was that the encoding of LLLA 'ḷ' ఴ (u+0C34) with native glyph is welcome and that GoAP does not want to continue using diacritics and Tamil symbol for this character in Telugu. Unfortunately, this valuable advice of GoAP was not duly considered by the proposer of the proposal L2/20-119.
16) Government of India indicated in L2/12-079 that it recommends the inclusion of the Character LLLA 'ḷ' ఴ (u+0C34) as proposed in L2/12-015.
17) As for RRRA u+0C5A, while GoAP in its letter L2/13-121 indicated the unknown phonetic quality, it acknowledged that the probable use of RRRA is like RRA. They also suggested a new name for RRRA as Telugu Alternate Alveolar Consonant. The probable/plausible intention of GoAP was to use RRRA as a graphic variant of RRA.
18) Under L2/19-401, a nukta proposal was also presented to the UC for use along with several consonant characters to represent additional foreign sounds. We do not have to discuss this proposal here except for its relevance to RRA 'ṛ' and LLLA 'ḷ', as it is an internal matter of the language community.

Analysis of the proposal L2/20-119:

19) Now limiting ourselves to LLLA 'ḷ' and RRA 'ṛ', let us first consider LLLA 'ḷ'. the Telugu language community has proposed 3 glyphs for LLLA 'ḷ'. [There was still another conjoint combination of u+0C32 along with u+0C30 ఀ used infrequently in later inscriptions to indicate LLLA 'ḷ'. This was also adopted in a few English transcriptions. An "rl" combination appearing in "swirl" will sound like LLL.]

a) One glyph was a native one in the Telugu-Kannada inscription from the 5th century to 11th century. This was considered through based on the proposal L2/12-015 and a codepoint u+0C34 ఴ was allotted to it.

b) Second was a Nukta solution affixing u+0C3C to u+0C33.

c) Third was the present new proposal to import Tamil LLLA, viz. u+0BB4 and create a new codepoint u+0C5B in the Telugu block.

20) It is quite surprising to know that the proposer wants to import u+0BB4 and be given a codepoint 0C5B, when Telugu already has 3 ways to accommodate LLLA 'ḻ'. We are unable to understand the underlying rationale for this request. We find it to be utterly unconvincing. One could have just limited to u+0C34 and popularized the same. Just to say that it is virtually unknown to the modern populace outside of the epigraphic domain is rather untenable. A language cannot have one LLLA 'ḻ' for epigraphy domain, another for religious text, another for literature and yet another for plain text (To show a parallel example, not all the characters in the Tamil block are known to public. A certain section of the block is known to only experts. It is only with prolonged use that the public comes to know and start using it.) It does not take much time or effort with the latest technology to develop Telugu fonts exhibiting different glyphs for RRA 'ṟ' and LLLA and popularize it with the public accommodating the requirements of Srivaishnava Telugu community. Font 1 can have Telugu u+0C34 LLLA 'ḻ' . Font 2 can have Tamil 0BB4 LLLA 'ḻ' along with shape engine modifications. The horizon is wide open. (Again, we bring a Tamil parallel here. The combination of Tamil NNA u+0BA3 + AI u+0B90 had a different glyph representation (ஜ) prior to 1975. GoTN brought out a Government Order for Tamil script reformation subsequently. The glyph of u+0BA3 u+0BC8 combination got completely changed as a result of this Order. Thus the Tamil consonant vowel NNAI changed from ஜ to ணை The present UC Tamil block shows only the latest glyph. Mr, Vinod Rajan himself has brought out an input software than can show pre-1975 glyph for the combination u+0BA3 u+0BC8, if anyone wants it. What is to be observed here is that when there two glyphs in Tamil, IT practitioners there did not go for an encoding solution. They went for a font development solution. This is precisely the practice that we would like to recommend to UC on LLLA and RRA issue discussed in L2/20-119.

21) Of course, there is an old solution of using the conjoint with a combination u+0C32 and u+0C30.

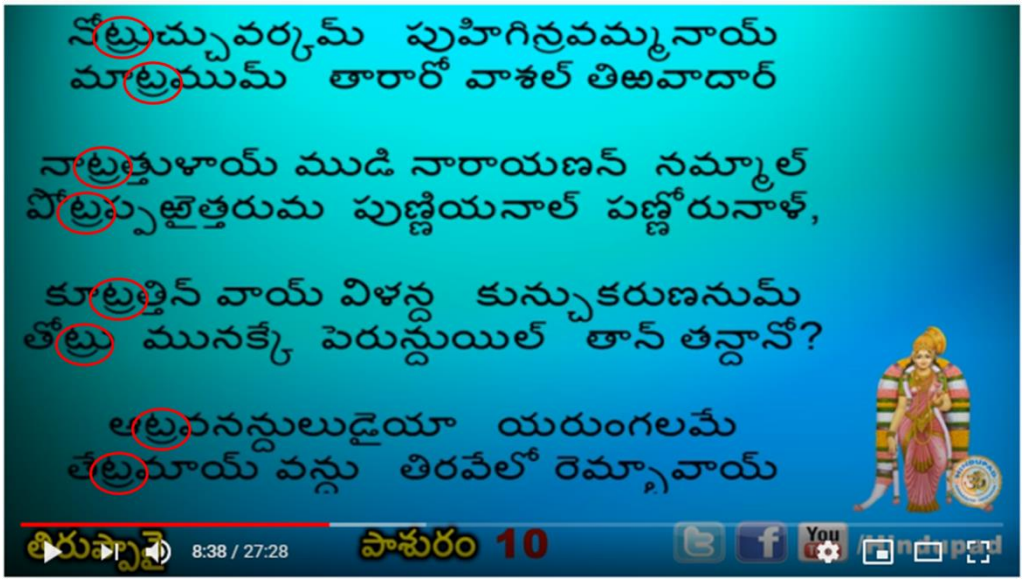
22) If one wants to still use Tamil u+0BB4, there are other technical ways of doing it. We ourselves have tried it through and it works. It is very much possible to combine u+0BB4 along with Telugu vowel mathras using the shape engine modules, perhaps with a few kerning adjustments. The author of proposal L2/20-119 is concerned about the rendering engines treating the Tamil characters correctly in the attested combination along with Telugu characters, and font vendors including these characters in Telugu fonts.

23) Now coming to the opening statement in L2/20-119 that they are proposing Telugu u+0C5B ళ as parallel to Tamil u+0BB4 ள so as to phonetically render Srivaisnava Tamil liturgical songs properly in Telugu transcription, this is a simple fallacy. As for as we can see, the Telugu-RRA (u+0C31) is phonologically equivalent to Tamil-RRA (u+0BB1). So, doubling them and adding an "a" sound would beget ற்ற. To say that they are not equivalent to Tamils' way of pronouncing is mere lack of observation. Perhaps the proposer of L2/20-119 is not aware that ற்ற is pronounced properly only by Srilankan Tamils, and wrongly by Tamilnadu Tamils (ட்ர). Unknowingly, the proposer wants to perpetuate a wrong pronunciation. Srilankan Tamils themselves do not mind the variation in pronunciation by Tamilnadu Tamils. Both Tamils are aware of the status. And hence continue to consider ற்ற as a simple phoneme joint, irrespective of the pronunciation difference. That is all.

24) In case if it is still required to keep TN pronunciation, we can suggest an alternative which we learnt from some of the Telugu scholars. They are of the opinion that the phonetic realization of Tamil ற்ற can be achieved by using native Telugu glyph combination like ఱ్ఱ(ట్ర) using u+0C1F and u+0C31, instead of importing foreign characters. UC and the proposer of L2/20-119 may want to have a discussion on this with the Telugu scholars to understand more on this aspect.

25) We have located another telugu source of the same āṇṭāḷ song with the occurrence of a Telugu character involving u+0C1F throughout the song for ற்ற. The picture is enclosed

below for reference. The verse and song can be seen and heard at https://www.youtube.com/watch?v=9ZGHDeEy_9I from 8.3 minutes (10th song of tiruppāvai). This only shows that the transcription of those text/verse in reference has multple versions with different characters. When this source with the native Telugu character involving u+0C1F providing the required phonetie realization for ற்ற, why would one go for encoding the foreign characters?



Conclusions:

26) As shown in the above, there are various reasons as to why encoding of Tamil RRA and Tamil LLLA into Telugu block is not advisable. Alternate solutions do exist. We do understand that the attestations shown by the author do exist. However it is technologically possible to represent these characters without encoding new characters. When the existing technology itself can represent the texts, there is no need to encode new characters. The spaces in the Telugu block may be used for future encoding of other native Telugu Characters.

27) As the author himself says, these natively Tamil letters may be added to ScriptExtensions.txt to allow the usage of the Tamil letters outside of Tamil block. Concerned users should request the appropriate updates to ScriptExtensions.txt, rendering engines software and their desired fonts to use the existing Tamil Characters.

28) Since these characters are not required for native Telugu writing but only for specialized transcription purposes, only specialized fonts will need to include these characters and font vendors may not want to add these characters to their general-purpose Telugu fonts.

29) Appropriately designed fonts can be used to provide the desired stylistic shapes of these characters in the particular contexts and there is no requirement to and it is also not possible to represent stylistic differences in plain text.

30) To preserve the identity of these characters as native Tamil letters, we recommend that these new characters should not be encoded, and the existing technology can be used with appropriate technical updates to represent these texts as discussed above.

……end of document……