Inside Tamil Unicode

Sinnathurai srivas

1 Introduction

11. Document Purpose

Titling the document as “Inside Tamil Unicode”, the author intends to interpret, analyse and detail the inner assignment of character codes in Tamil Unicode Encoding and expand on how this understanding can be applied to the development of Tamil software and Unicode Tamil font. As the author comes from a background of Computing and Information Technology development coupled with a long term involvement in researching the ins and outs of Tamil Alphabet and phonology, this paper intends on detailing the technical and cultural merits of the current Tamil Unicode encoding and opens up thoughts on future encoding enhancements.

12. Audience

This paper together with a presentation on stage specifically targets Tamil Computing related software development, complex rendering processing, Tamil Unicode and Indic Unicode font development, speech recognition using Tamil alphabet system, Tamil pronunciation dictionary, spell and grammar check for Tamil, and any one interested in Tamil as a language. This document also touches on the authors’ long proclaimed interpretations of the definitions of Tamil alphabet/ezuththu and thoughts on Ezuththuch chiirmai.

13. Background

INFITT is the technical institution that works hand in hand with Unicode Consortium, Tamil Nadu government, Tamil IT related government organisations of the world, Tamil and other universities, and

Tamil computing related software and hardware developers, along side Unicode Consortium. This paper is intended for describing technical and cultural information to anyone who chooses to use it.

Unicode Consortium is the international institution tasked with standardising all languages of the world for enabling the standard international multilingual computing. Tamil Unicode is already a working entity in Computing and the works in progress are for enhancing and facilitating a totally empowering Tamil computing environment.

2 Abstracts

The following topics are covered in this presentation.

Unicode as a Linear Encoding

Introducing Fallback Unicode characters to match the linear encoding

Linear encoding as an extension of Tolkappiyam

The need for complex rendering and the displaying/printing of Tamil

Tamil in cut down versions of OS and domestic appliances

Simplifying sorting process and code point for inherent “a”.

Pulli vs Virama

Matra vs Matrai

Extra long vowels, kuTTiyal, Matrai

Deprecating the duplicate “au” marker

Eliminating the alternative aravu usage with combining ee and oo.

3 Inside tamil unicode

31. Tamil Grammar and Liner Unicode Encoding.

Ancient Tamil grammar Tolkappiyam, which is also the contemporary Tamil Grammar, defines 30 alphabet (ezuththu) and 3 markers as the entities used in Tamil writing system. It is important to note that unlike any other languages of the world, Tamil writing system names the “Places of Articulation/Birth” (Pirappidam) in human organs,

which generate speech sounds/phonemes, and defines those Places of Articulation (PoA) as alphabet. It is also important to note that each alphabet represents a spectrum of phonemes generatable by its PoA. The thirty alphabets are divided into two types as 18 consonants and 12 vowels. Consonants (mey/physique) are the covered physical organs of the PoA while vowels (uyr/soul) are the covered places of articulation. The vowel is believed to agitate the consonant to live for a suitable matrai/time-interval and vowel can also spring to live on its own to a required matrai. However, the consonant in general is thought incapable of springing to live on its own; hence in Tamil, consonant conjuncts are thought scientifically as non-existent. However, Tamil defines near-voiceless/kuRRiyal vowels as enabling factor for consonant-conjuncts. Unicode is yet to encode **kuTTiyal markers** and **matrai/timing markers** as defined in Tamil Grammar. In day-to-day usage, Tamil is believed to contain countless number of phonemes because of the scalable nature of each PoA. A request for the use of **diacritic-markers** to define phonemes into sub-spectrum ranges is yet to be made, probably by INFITT. This will be useful for activities such as pronunciation dictionaries and speech recognition interpreters.

Unicode encodes thirty alphabets and one marker (aytham) as the basic Tamil character set. This is called linear encoding. The philosophy of 30 characters is in sync with Tamil grammar. Additionally a few Tamil-Granta characters are also encoded within the Tamil Unicode code range. Though it is in day-to-day usage, Tamil-Granta however does break the principle that alphabets in Tamil represent scalable PoA and not phonemes.

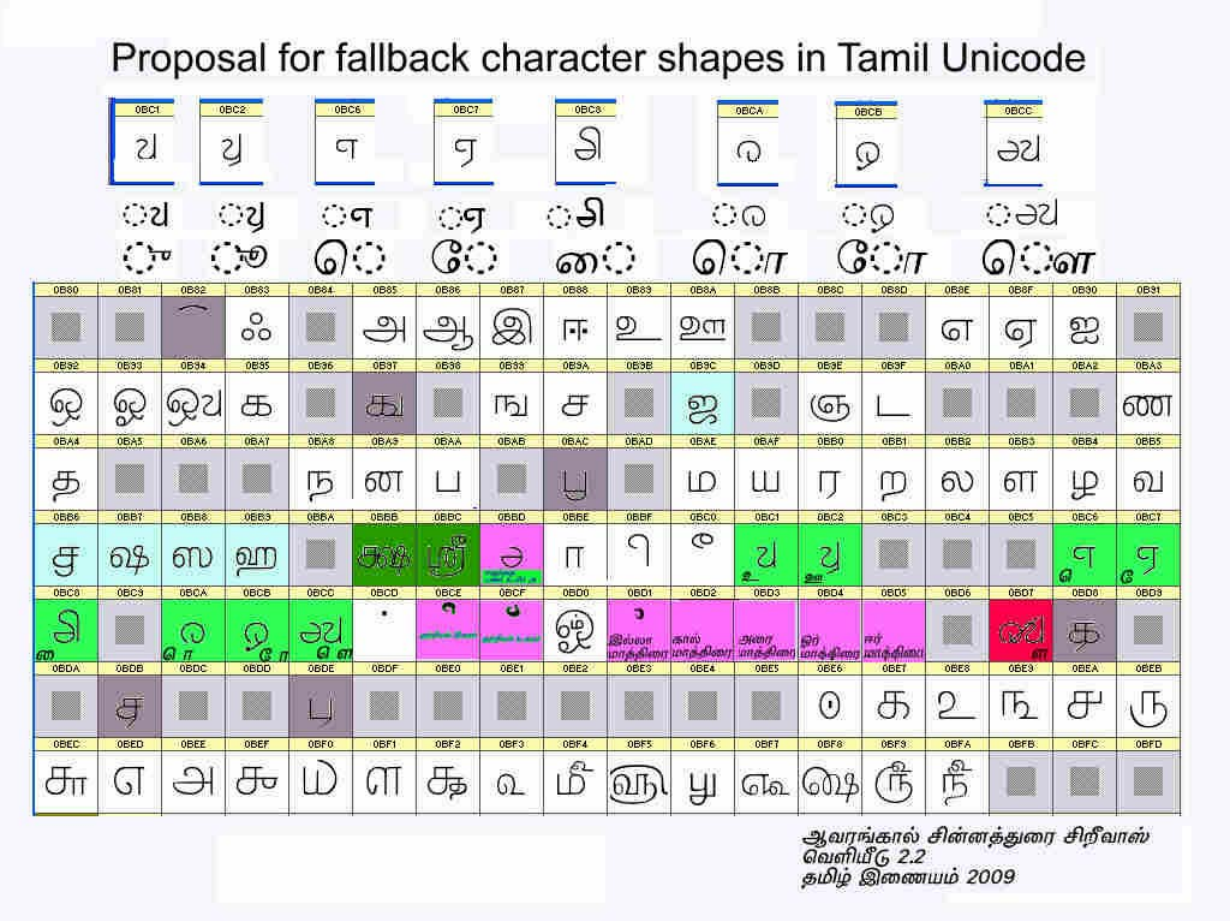
In theory the processing of Tamil data is achievable at linear Plain-1 level. This would imply, for example sorting of Tamil text is processed at 16bit Plain-1 without consideration for any complex processing of 32bit and beyond. In theory, for Tamil, only display and printing should be processed using complex rendering at 32bit and beyond. However, due to present state of operating system design shortcomings and shortcomings within Unicode (additions) definitions, complex processing is also required for a very few instances of processing in Tamil software,

in addition to print and display processing. A fix for these shortcomings cannot be expected in the immediate future. Therefore, when architecting software solutions, one need to decide if linear processing requirements and complex processing requirements can be modularised so that **about 99%** of the development tasks can be simplified to 16bit processing. Some of the shortcomings in Unicode definitions can be listed as “X as the only one complex conjunct in Tamil”, “ duplicate au-marker”, and clearance for the use of duplicate “combu-markers”, which cause unnecessary software development initiatives, when Tamil could be developed simply at 16bit level. Inherent “a” is currently not encoded for any of the Indic languages. For now, developers can assign their own code point to inherent “a” so that software routines for sorting Tamil can be made virtually a simple task. All of the shift and pull operations required to sort in complex requirement, because of the lack of inherent “a” code point, will become a simple task once a code point is assigned for inherent “a”. There are written materials that talk about the discussions took place in the ancient time about the pros and cons of the use of inherent “a” and visible “combing “a”. The debate still goes on and that requires the facility to express the various theories on the pros and cons. It is therefore necessary to **encode inherent “a”** with an alternative visible “a” form and its usage be permitted for research and software processing purposes.

Software developers need to understand that in Tamil Unicode there are no “combukaL (ு, ூ)”; there is no “sangkili kombu (ை)”; there are no “uhara-mey nor uuhaara-mey (கு, கூ, … னு, னூ)”. The uyr-meykaL that appear on print and on screen are not really there. They are only software illusions. Once a software developer understands this phenomenon, the development cycle would become a walk through exercise. However, erroneously encoded secondary “au marker U+0bd7” in Unicode is real and this can cause havoc if its behaviour is not understood properly. Note that the primary “au marker U+0bcc” behaves normally like any other combining vowels and it is fully fit for purpose. As depicted in the image below, all the real combining vowel markers only combine with consonants from the right. This is in line with Tamil Grammar, while the

contemporary written Tamil combining vowels misleadingly combines with consonants from left or right or top or bottom or left & right and even manufactures new shapes (u-haram, uu-haaram). For software development purposes misleading appearances can be discarded and grammatical definitions of alphabet can be utilised. To graphically represent the linear nature of combining vowels in Unicode/Tolkappiyam the letters' shapes "ா ி ீ ு ூ ெ ே ை ொ ோ ௌ" are recommended. All of these shapes combine with consonants, visibly on the right side, as logically coded in Unicode/Tolkappiyam. When there is a lack of complex rendering OS interface, Tamil need to use these linear forms of the combining vowels. Simple electronic devices would always lack complex rendering facility. Displaying the deformed combukaL in these instances would be an utter degrading experience. Some examples of utter degrading Unicode displays are கெ=க◌, கொ=க◌, கே=க◌, கோ=க◌, கை=க◌. As the simple electronic devices are always going to be in existence, Unicode standard should allow the Fllback characters of combining vowels as ◌ா ◌ி ◌ீ ◌ு ◌ூ ◌ெ ◌ே ◌ை ◌ொ ◌ோ ◌ௌ and not as degrading கெ=க◌, கொ=க◌, கே=க◌, கோ=க◌, கை=க◌.

The picture below provides a conscious guide to what is encoded in Unicode, what need to be encoded in Unicode, what need to be deprecated out of Unicode, what need to be clarified as external to Tamil and also gives a head start for Tamil character reform in line with Tholkappiyam.



A long standing claim by me is that CombukaL in Tamil are a highly illogical, disruptive and retarding influencers that were introduced in recent times by Viramamuni, because of his lack of understanding of Tamil, even though intentions were good. So the combukaL related linear representation can be made part of Tamil character reform. It is scientific, logical and simple if all combining vowels take position on the right side of consonant, naturally. For this reason, proposed linear characters can be classed as potentially an important item in the Tamil character reinstatement/reform agenda. The withering of u-hara meykaL and uu-haara meykaL will also come natural, if Unicode fallback characters are made part of any reform agenda.

Numerous phonemes/sounds exist and are used in day to day language. Each PoA generates a number of phonemes. Diacritic markers are essential to document and communicate the use of these phonemes and also to publish pronunciation dictionaries. New Unicode code point

ranges need to be allocated for Tamil diacritics or the feasibility of using the existing diacritic markers in Unicode for Tamil with the defined diacritic glyph shape and positions need to be considered. (Unicode code range U+0300 to U+036f. The PoA/pirappidam "த" for example, generates multiple phonemes such as அத்தகு, அந்த, அகன். The phoneme த is not proved to exist in any other languages. Though some say some use this phoneme த in the word father rather than the phoneme த. The PoA/pirappidam "அ" for example, generates multiple phonemes such as அம்மா, அன்னை and வகை, விரை. The effect of diacritics when combining before or after a Tamil character also need thorough investigation, because of the nature of disruptive combining vowels in Tamil, joining the consonants in every direction in a non-uniform fashion.

In Unicode, the character Virama for other Indic languages was erroneously assumed to have similar properties to the puLLi character in Tamil. However, there are considerable differences between Virama and puLLi. Because Unicode names the puLLi also as Virama (with recent annotation), there is a danger that developers may assume the general characteristics of Indic Virama as the characteristics of puLLi and indulge in wasted development activities. Similarly Aytham in Tamil is wrongly named and defined (with recent annotation) as Visarga. Visarga has totally unrelated properties to Aytham; hence the properties of Aytham are wrongly defined in Unicode. Again, it is the responsibility of developers not to indulge in wasted efforts by the lead-believe, that the Visarga and Aytham as having the same/similar properties. The Aytham in Tamil act as a vibrationary modulator for glotalising and other purposes of speech. Example formation of Aytham are "◌ஃக, ◌ஃ◌ஃத, பஃ, பஃ◌ஃ, கஃ◌ஃப" and the recent usage of "◌ஃப" to clearly identify the phoneme "f=ஹ". For example, the single character KHA in Devanagari translates as KAH=கஃ=க ◌ஃ. Matrai in Tamil grammar defines the timing mechanism involved with generating spoken sounds, while Matra in Unicode is used to denote the combining vowels (puNar Uyr/Pin uyr ezuththukkaL அ ஆ ◌ி ◌ீ ◌ு ◌ூ ◌ெ ◌ே ◌ை ◌ொ ◌ோ ◌ௌ ◌்).

4 References

Tamil Unicode Chart: http://www.unicode.org/charts/PDF/U0B80.pdf

Combining Diacritics U+0300 to U+036f: http://www.unicode.org/charts/PDF/U0300.pdf

Tamil grammar “Tolkappiyam”

Document No: Arai/2u **Version: 2.2** **Date:** Saturday, September 19, 2009 2009-09-17T16:35:00Z