A Brief Report on the Evaluation of Unicode Standards for Tamil

1. Introduction

There has been a uniform feeling in the minds of the Tamil computing persons in the world that the present Unicode standard for Tamil is not adequate for efficient and effective usage of Tamil in computers, particularly, for applications like Database Design with search and sort facilities and Natural Language Processing. While considering an alternative to the present Unicode Standard for Tamil, there are two different views : One, the scheme that incorporates only the vowels and consonants into the encoding scheme, leaving the other characters to a rendering process. Two, the scheme that incorporates all the Tamil Characters including vowels, consonants and vowel-consonants as characters in the standard.

For quite some time, there is a continued debate on these schemes, for and against. But until recently, no one had carried out any investigation on the merits of the different schemes. The problem was discussed at the Ministry of Information Technology, Government of India and at the State Level IT Task Force, Govt. of Tamil Nadu and a decision was taken to test these schemes systematically through TSDF. Accordingly, to overcome these difficulties and to take an appropriate decision on the most efficient encoding scheme for Tamil, the TVU took the initiative of getting the three schemes tested thoroughly. The testing responsibility was assigned to a software company, CADGRAF through a contract. The studies were completed and the test results are consolidated in this report.

2. Schemes considered for evaluation

The following three different Unicode encoding schemes were considered for evaluation :

(i) Present Unicode Standard for Tamil. (Fig. 1)
(ii) Vowel and Consonants for Tamil. (Fig. 2)
(iii) All Character Encoding Schemes. (Annexure 3A)

1) Scheme 1 (Existing Unicode ver 3.0)

Scheme 1 is the existing Unicode ver.3.0. The additional characters like symbols, fractions, archival characters etc. which are included in the Scheme 2 and Scheme 3 are not included in this scheme.

2) Scheme 2 (Pure Consonant - Vowel Scheme)

In Scheme 2 the 23 consonants (mei characters) (कखगघङचछजझञटठडढणत थदधनऩपफ) are defined as consonants in the range of 0915 to 092B. The 12 vowel characters (□□□□□□□□□□□□) and the aytham क़ are defined in the range of 093E to 0949 and 0958.

Further, in Scheme 2 the following 38 additional characters are also added which are not available in the current Unicode scheme (Scheme 1).

i) Archival characters - 7 numbers
(□, □, □, □, □, □, □ in the range E 001 - E007)

ii) Tamil symbol characters - 14 numbers
(□, □, □, □, □, □, □, □, □, □, □, □, □, □ in the range E180 - E18D)

iii) Tamil fraction characters - 16 numbers
(E190 - E19F □, □, □, □, □, □, □, □, □, □, □, □, □, □, □, □ in the range E180 - E18D)

iv) Tamil numeral zero (□ in E1A0)

The consonants and vowels in Scheme 2 are placed inside the Indic Devanagiri locations and the additional characters are placed in the Private Use Area from E000 to E1AF. But the actual Encoding Scheme (Fig. 2 to be implemented is given in the Fig.4)

3) Scheme 3 (All Character Scheme)

In Scheme 3 all Vowels, Consonants, Vowlised consonants and Additional symbol characters (as mentioned in scheme 2) are defined as independent characters and allotted separate locations.

Scheme 3 characters are defined in the Private Use Area in the range of E000 - E17F.

In Scheme 3 a Null character is also defined in the location E000.

4. Editor / Interface for inputting Unicode Characters

To input Unicode Tamil Characters of anyone of the above schemes, an interface package was developed incorporating the TamilNet 99 keyboard. Two fonts for each scheme were developed, bearing the names xxxxx_TVU_BARATHI and xxxxx_TVU_PARI where xxxxx stands for TVU_1 / TVU_2/ TVU_3.

Wordpad application and TVU-Uni-Editor were used for the investigations carried out.

5. Investigations

The following investigations were made :

(i) Space occupation of the Encoding Schemes.
(ii) Efficiency of Text Processing.
(iii) Efficiency of Database Application
(iv) Efficiency of Morphological Analysis.

5.1 Space Occupation

The storage spaces occupied by the three encoding schemes are given below, with and without the special characters added to schemes 2 &3.

		File size in kB Without additional Symbols	*File size in kB With additional Symbols*
Scheme 1	-	56 kB	56 kB
Scheme 2	-	70 kB	82 kB
Scheme 3	-	65 kB	77 kB

5.2 Text Processing

Five different Source files of varying sizes were used for testing. The following parameters were assessed.

(i) File Size
The results are presented in Table –1
Scheme 3 requires almost 30% less space for storing a data file than the other two schemes.

(ii) File Reading to Memory and Memory to Display
The results are presented in Table –2
It takes just 35% of the time taken by the other schemes to read from disk to memory and to display from memory.

(iii) Time for Search and Replacement
The results are presented in Table –3
To find a word and replace it by another scheme 3 takes only about 30 to 40 percent of the time taken by the other schemes.

(iv) File Compression size and time
The results are presented in Table –4
In terms of compression, scheme 3 does not show any special advantage. In all the three schemes, the compressed file size is almost the same.

(v) Time to File Copy from Disk to Disk
The results are presented in Table –5
For copying a file from disk to disk, scheme 3 takes only 33% less time than the other schemes.

5.3 Database Application

(i) Creation of Unicode database and file size
The results are presented in Table –6
When creating a Database, scheme 3 takes 11 to 13 percent less time and occupies 14 to 20 percent less space than the other two schemes.

(ii) Indexing of Unicode database and indexing time & indexing file size
The results are presented in Table –7
Investigation on Indexing shows that scheme 3 takes only about 56% of the time taken by the other schemes and the index size is only about 65% of the size required by the other schemes.

(iii) Sorting of Unicode database
The results are presented in Table –8
In the case of database sorting, Scheme 3 takes just about 60% of the time taken by the other schemes.

(iv) Searching in Unicode Database
The results are presented in Table –9
For searching a given word in a database, scheme 3 takes just 5% less time than the other two schemes. Thus, no appreciable difference in performance.

5.4 Morphological Analysis

Search using Morphological Analysis

Basic Assumptions :

To search for words, at the end of a sentence, which end with a masculine suffix (an, An)/feminine suffix (aL, AL)/plural suffix (ar, Ar). Each one of them has to found independently, in test files of the 3 Unicode versions. Same text matter is to be used for all search.

Solution methods for the searches :

The full stop is used to find only the words at the end of a sentence.

Unicode 1

Search for “an”
Step 1 – search for the string “na+MeiPulli+”.”. (Length of search string = 3)
Step 2 – check whether the previous character is in the range of ka to na.

Search for “An”
Step 1 – search for the string ThunaiEzhuthu+”na”+pulli+”.”.(Length of search string =4)

Unicode 2

Search for “an”
Step 1 – search for the string “a”+”n”+”.”.(Length of search string = 3)

Search for “An”
Step1 – search for the string “A”+”n”+”.”. (Length of search string = 3)

Unicode 3

Search for “an”
Step 1 – search for the string “n”+”.”. (Length of search string = 2)
Step 2 – check whether the previous character has the value 1 in the last 4 bits.

Results

Time (in milliseconds) taken by repeating the Search five times in the file Search Sample.
Computer used – 400MHz Celeron and the results are given in Table 10.

Scheme 3 seems t be the best scheme for morphological analysis. It takes just 38% of the time taken by scheme 1 and 45% of the time taken by Scheme 2. At the outset it may seem to be not believable. But it is because of the following reasons. Tamil text is mostly with vowel consonants. Each vowel consonants occupies a 32 bit space in scheme 3. Hence the time taken for searching a word over entire text space is less in Scheme 3 due to the above reason. Further, the end character is known by checking only the last 4 bits of character space in Schemes 3.

Conclusion

As could be seen from the results tabulated in Table 1-10, the discussion presented above the Scheme 3 proves to be efficient in terms of the resource consumption and execution time. Hence, Scheme 3 is recommended for incorporation in Unicode Standard for Tamil.

Table – 1

Investigation of File Size

	TAB File Size	Unicode		% on Scheme 1 (File Size)
		File Size	Size on Disk	
Sample I				
	32 KB			
Scheme 1		67 KB	72 KB	**100**
Scheme 2		66 KB	72 KB	**99**
Scheme 3		44 KB	48 KB	**66**
Sample II				
	68 KB			
Scheme 1		143 KB	144 KB	**100**
Scheme 2		139 KB	144 KB	**97**
Scheme 3		97 KB	104 KB	**68**
Sample III				
	159 KB			
Scheme 1		338 KB	344 KB	**100**
Scheme 2		327 KB	328 KB	**97**
Scheme 3		229 KB	232 KB	**68**

Fig. 1

SCHEME 1 ENCODING - CODE CHART

(font used : TAU_1_TVU_BARATHI)

	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0		ஐ		ர	ீ			ௐ
1				ற	ு			௑
2	்	ஒ		ல	ூ			௒
3	ஃ	ஓ	ண	ள				
4		ஔ	த	ழ				
5	அ	க		வ				
6	ஆ				ெ			
7	இ			ஷ	ே	ௗ	க	
8	ஈ		ந	ஸ	ை		உ	
9	உ	ங	ன	ஹ			ஙூ	
A	ஊ	ச	ப		ொ		சூ	
B					ோ		டூ	
C		ஜ			ௌ		கூ	
D					்		எ	
E	எ	ஞ	ம	ா			அ	
F	ஏ	ட	ய	ி			கூ	

Fig. 2

SCHEME 2 ENCODING - CODE CHART

(font used : TAU_2_TVU_BARATHI)

	091	092	093	094	095	E00	E18	E19	E1A
0		ठ		இ					0
1		ड		ஈ					
2		ढ		உ					
3		ण		ஊ					
4		त		எ					
5	क	थ		ஏ					
6	ख	द		ஐ					
7	ग	ध		ஒ					
8	घ	न		ஓ	क़				
9	ङ	ऩ		ஔ					
A	च	प							
B	छ	फ							
C	ज								
D	झ								
E	ञ		அ						
F	ट		ஆ						

Table 10

Morphological Analysis

	Scheme 1	Scheme2	Scheme 3	No of Words found
Ar	230	230	150	1555
Ar	1933	1573	540	2945
An	105	105	65	410
An	135	120	85	855
AL	135	135	90	825
AL	175	155	105	1140
Time in ms	2713	2318	1035	
% on Scheme 1	100%	85.44%	38.15%	