Sorting Tamil Words: Issues and Solutions

K.Ponmozhi
Assistant Professor, Department of Information Technology, Hajee Karuth Rowther Howdia College
Uthamapalayam, Theni, TamilNadu

Abstract* — There are approximately 65 million Tamils in India, and 80 million worldwide. It has been recognized as Classical Language. Collation is one of the most important features of a script. It determines the order in which a given culture indexes its characters. Unicode has become a world standard and many computer applications have provided Unicode support so that multilingual text can be handled. . It encodes glyphs which have no sound and are not characters in Tamil. The Unicode standard proposed for Tamil has not taken into consideration some of the important linguistic issues. This leads to problems in Language processing in Tamil, especially the sorting which is the basic operation for database applications and the like. Unicode does not treat uyir-mei characters as separate characters. So for one uyir-mei characters there may be two or more Unicode characters are to be combined. Sorting in Tamil is not character sorting as that of English. In order to do sorting, this disparity has to set right then only we can do comparisons. In this paper we have shown the detailed picture on encoding of tamil characters, their problems, an algorithm for sorting based on current Unicode encoding, and the recommendations of Tamilnadu government TACE-16. As of now, we need to do sorting in Two phases The recommendations of TACE-16 has shown best for Tamil character encodings.

Keywords — *Tamil Computing, sorting, collation.*

I. Introduction

There are approximately 65 million Tamils in India, and 80 million worldwide. Usage of Tamil in commercial transactions, birth/death certificates, certificates related to studies, and petitions etc are in the increase and may be in multiples of million every year. It has been recognized as Classical Language.

Computing with Tamil relates to the design and development of useful applications which permit interaction with the system in Tamil. Among the Indian Languages, Tamil perhaps has the simplest set of aksharas consisting of twelve vowels and eighteen consonants. However, six aksharas from Sanskrit have also become part of the set. For many years now, text in Tamil has been displayed on the web, thanks to the magazines which have gone on-line. Each publication has standardized the approach to displaying the text through designated fonts and there are quite a few of them. Besides the magazines, independent groups have proposed data entry schemes to be used with specific fonts which seem to cater to some sort of character coding for the characters. The multiplicity of the fonts seen has posed real problems in arriving at some uniformity in text display.

This was the main theme of discussions during the Tamilnet99 conference [2] held in Chennai during February 1999. At this conference, it was proposed that the placement of glyphs within a Tamil font would follow a recommended scheme. Both bilingual and monolingual schemes were standardized. The conference also arrived at a standard for data entry in Tamil. Three different keyboard layouts were arrived at for use by different sections of the users. The first relates to what is termed as the phonetic keyboard where data entry is affected through lower case keys alone for the basic text. The second scheme referred to as the Romanized keyboard, specifies data entry based on the Roman letter that comes close to the sound of the vowel or consonant. The third is the layout seen in standard Tamil typewriters.

Total number of Tamil characters including Grantha letters : 325 [3], Tamil Numerals: 13. Special characters 9 and thus a total of 347 code points are required to represent Tamil character set. In general the difficulties at various levels of analysing Tamil text are due to the large set of characters and encoding system [8]. Current Tamil Encodings are ISCII – 7 bits, TSCII/TAB – 7 bit, TAM – 8 bit, Unicode – 7 bit.

Unicode is widely used encoding which has many disadvantages. Section II focuses on Tamil language characters, section II specifies overview of sorting and issues of Unicode, next section specifies a possible way to get over Unicode problem, and the TACE-16 recommendations and its advantages, Finally conclusion.

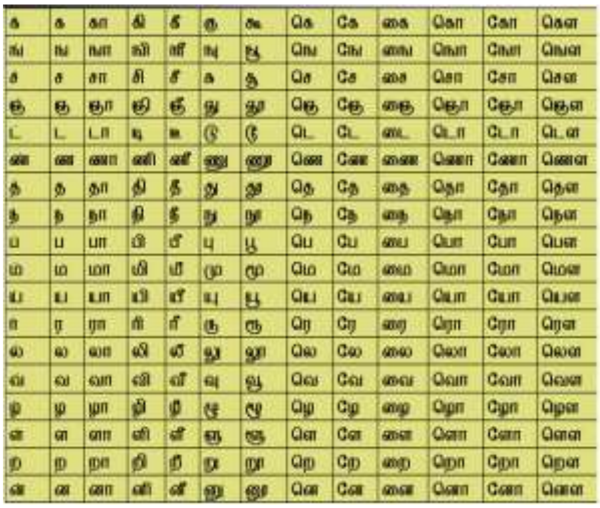
II. Overview of Tamil language writing system

- Tamil is an alphasyllabary with the akshar as its core. Its main features are:
- The consonant has an implicit vowel built-in.
- The inherent vowel can be modified by the addition of other vowels or muted by a diacritic.
- Vowels can be handled as full vowels with vocalic value
- Generally akshar is a single consonant or a consonant followed by a pulli.

A. *Characters in Tamil Language*

Character set for Tamil [5] comprises of the following:

- Vowels: அ,ஆ,இ,ஈ,உ,ஊ,எ,ஏ,ஐ,ஒ,ஓ,ஔ,ஃ.
- Consonants: க்,ங்,ச்,ஞ்,த்,ந்,ப்,ம்,ட்,ண்,ய்,ர்,ல்,ள்,வ்,ழ்,ற்,ன்.
- Vowel-consonants(uyir-mei): These are characters combined by vowels and consonants. They are tabulated in table 1.



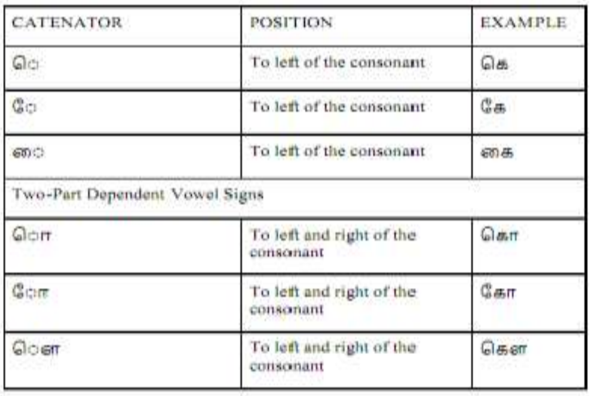
க்	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ
ங்	ங	ஙா	ஙி	ஙீ	ஙு	ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ
ச்	ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
ஞ்	ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
ட்	ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ
ண்	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
த்	த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
ந்	ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ
ப்	ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ
ம்	ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ
ய்	ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ
ர்	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ
ல்	ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	லௌ
வ்	வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ
ழ்	ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ
ள்	ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ
ற்	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
ன்	ன	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ

Table 1: Uyir-mei characters

- Matra set: It is vowel modifier set.

◌ா ◌ி ◌ீ ◌ு ◌ூ ◌ெ ◌ே ◌ை ◌ொ ◌ோ ◌ௌ

- Displaced catenators: Under normal circumstances vowel modifiers also known as catenations (since they concatenate to the preceding consonant) in scripts are written from left to right in linear order. However certain modifiers are placed to the left and right of the consonant to which they concatenate. These are termed in Unicode as Two-part Dependent vowel signs. Catenators with example is shown in table 2.



CATENATOR	POSITION	EXAMPLE
◌ெ	To left of the consonant	கெ
◌ே	To left of the consonant	கே
◌ை	To left of the consonant	கை
Two-Part Dependent Vowel Signs		
◌ொ	To left and right of the consonant	கொ
◌ோ	To left and right of the consonant	கோ
◌ௌ	To left and right of the consonant	கௌ

Table 2: Catenators in Tamil Unicode encoding

- Numerals : the used in Tamil language are:

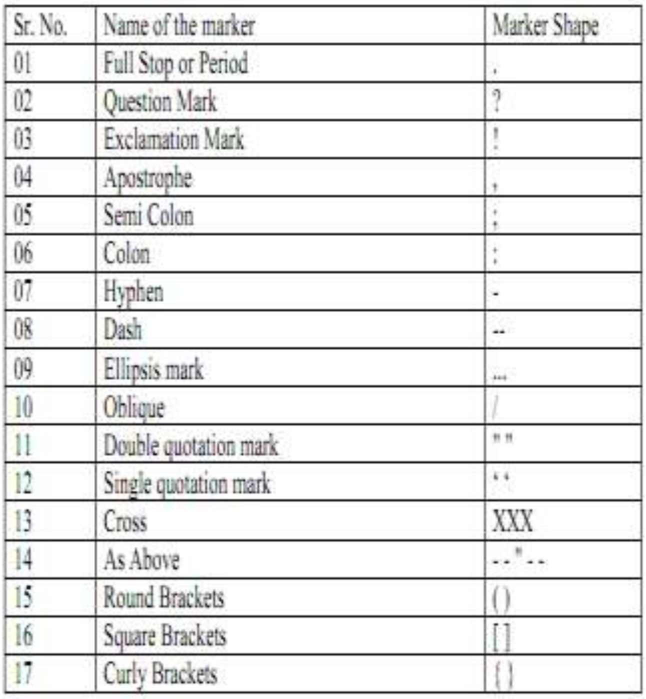
Numeral Shapes	Explanation
௦	Tamil Digit Zero
௧	Tamil Digit One
௨	Tamil Digit Two
௩	Tamil Digit Three
௪	Tamil Digit Four
௫	Tamil Digit Five
௬	Tamil Digit Six
௭	Tamil Digit Seven
௮	Tamil Digit Eight
௯	Tamil Digit Nine

Use of English numerals occurs in handwritten text as well as in all official documents and also in day to day use. Tamil numerals are rarely used.

- Numeric characters

௺	Tamil Number Sign
௰	Tamil Number Ten
௱	Tamil Number One Hundred
௲	Tamil Number One Thousand
௰௲	Tamil Number Ten Thousand[9]

- Punctuation marks.



Sr. No.	Name of the marker	Marker Shape
01	Full Stop or Period	.
02	Question Mark	?
03	Exclamation Mark	!
04	Apostrophe	'
05	Semi Colon	;
06	Colon	:
07	Hyphen	-
08	Dash	--
09	Ellipsis mark	...
10	Oblique	/
11	Double quotation mark	" "
12	Single quotation mark	' '
13	Cross	XXX
14	As Above	--"--
15	Round Brackets	( )
16	Square Brackets	[ ]
17	Curly Brackets	{ }

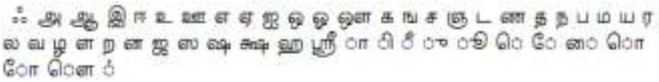
Though script shape has changed over centuries its syllabic characters and sound remains the same. Uyir-meys are not glyphs, not ligatures, not compound words. But are simple characters just like A, B, C, are characters to English speaking children க, கா, கி, கீ are characters to Tamil children.

B. *The collation order of Tamil*

The collation order refers to the order in which the characters is given language are sorted. One of the issues which has received much attention in respect of Indian languages and Unicode is the problem of sorting order (called collation by some experts). Collation is one of the most important features of a script. It determines the order in which a given culture indexes its characters. This is best seen in a dictionary sort where for easy search words are sorted and arranged in a specific order. Different scripts admit different sort orders and for all high-end Natural Language Processing applications, sort is a crucial feature to ensure that the applications index data as per the cultural perception of that community. Every Tamil child has been learning Tamil character set with the character order shown in table1.

Traditionally, the assignment of codes to the characters of a language took into consideration the order in which the letters of the alphabet would be arranged for purposes of creating lists which could be viewed easily and scanned quickly by a person. Almost all the classical sorting algorithms (including indexing of data bases) arrange the letters in the increasing or decreasing order of the assigned codes.

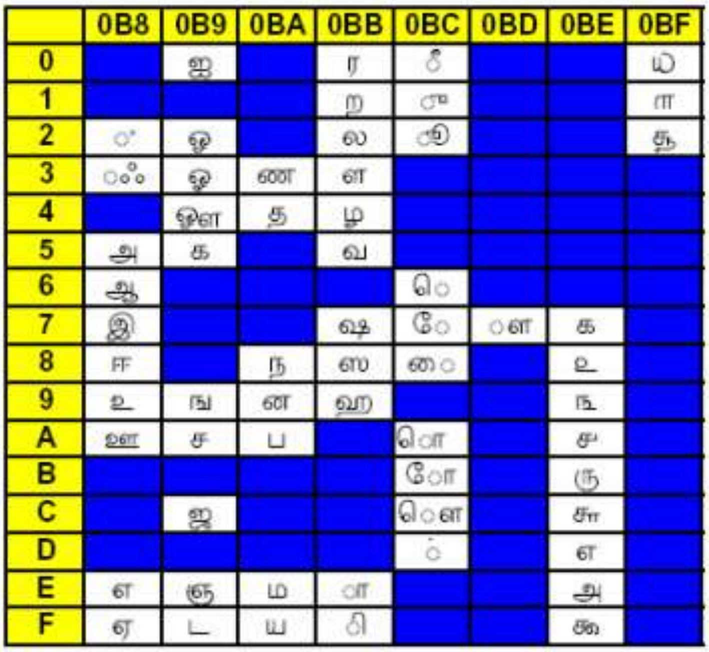
With respect to Tamil there are four main coding and collation standards. They are Unicode, TSCII, TAB and TAM. Of these Unicode and TSCII are widely used in the internet and Tamil softwares. The sort order provided by Unicode



ஃ அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ ஓ ஔ க ங ச ஞ ட ண த ந ப ம ய ர
ல வ ழ ள ற ன ஜ ஸ ஷ க்ஷ ஹ ஸ்ரீ ா ி ீ ு ூ ெ ே ை ொ
ோ ௌ ்

III. UNICODE FOR TAMIL

Unicode standard is the universal character encoding scheme for written characters and text[6].Unicode[1] has become a world standard and many computer applications have provided Unicode support so that multilingual text can be handled. Unicode Tamil encode is shown in figure 1. 16-bit space that is 64,536 code points are available. Uses only 128 code point block and that too is mostly empty. It encodes glyphs which have no sound and are not characters in Tamil. The Unicode standard proposed for Tamil has not taken into consideration some of the important linguistic issues. Also, even among the professionals, there seems to be considerable difference of opinion in respect of the adequacy of Unicode.



	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0		ஐ		ர	ீ			௰
1				ற	ு			௱
2	ஂ	ஒ		ல	ூ			௲
3	ஃ	ஓ	ண	ள				
4		ஔ	த	ழ				
5	அ	க		வ				
6	ஆ				ெ			
7	இ			ஷ	ே	ௗ	௧	
8	ஈ		ந	ஸ	ை		௨	
9	உ	ங	ன	ஹ			௩	
A	ஊ	ச	ப		ொ		௪	
B					ோ		௫	
C		ஜ			ௌ		௬	
D					்		௭	
E	எ	ஞ	ம	ா			௮	
F	ஏ	ட	ய	ி			௯	

Figure 1. Unicode Tamil Encoding

A. ***Issues of Unicode in Tamil***

- 7/8 bit is insufficient to represent all Tamil characters. By using them a maximum of 256 characters only (2^8 = 256).
- They are Hinders to Natural Language Processing including parsing, searching, sorting etc.

- Inefficient to store, transmit and retrieve. There should be complex algorithms for each of those functions.
- Need Normalization for string comparison.

It is clearly known that Unicode has not taken into account the required lexical ordering of the aksharas in any of the Indian scripts. This is understandable, for Unicode was essentially derived from ISCII where the ordering was based on similar sounding aksharas rather than the actual ordering conventions and this applied mainly to the Southern Languages. ISCII gave a uniform set of codes for all the languages however and perhaps on account of this no one really raised the issue. Unicode made a departure by assigning language specific (actually script specific) codes to our aksharas but in essence retained the basic structure of ISCII.

The two ர, ற of Tamil are placed together though they are separated by four consonants in the conventional order U0BB1(ர), U0BB2(ற) . The two "U0BA3 ண " “U0BA9 ன ” in Tamil are placed together where as they are separated by nine consonants. The very soft "ன " in Tamil actually comes at the end. The consonants in our languages are also grouped together linguistically and it will be necessary to keep this in mind when attempting any sort of Linguistic Text processing.

Only 10% of the Tamil characters are provided code space in the present Unicode Tamil. 90% of the Tamil characters are the vowel consonants of these only the following vowel consonants are encoded க,ங,ச,ஞ,ட,ண,த,ந,ப,ம,ய,ர,ல,ள,வ,ழ,ற,ன

Other vowel consonants need to be rendered using the following vowel consonants and the vowel signs encoded in the standard through a specially designed rendering engine which will change the shape of the consonants[4]. Same character can be formed by two different sets of code points leading to ambiguity for example

0B9A (ச) + 0BC6(□) +0BBE (□) = சொ

0B9A(ச) + 0BCA(□) = சொ

In the first place, it is a difficult proposition indeed to write any text processing application which has to work with multiple characters to arrive at a linguistic quantum, namely the syllable, which is central to all the Indian languages. If Unicode had concentrated on the linguistic content alone and had not prescribed rendering rules, the situation would be a little better. This is not the case however and linguistic processing with Unicode will require very complex algorithms to actually infer the context in which each character appears by examining the characters appearing before as well as those appearing after it.

Consider the situation in respect of the Matras. The matra itself is not a proper linguistic unit but a representation of a medial vowel, i.e., a vowel occurring in a syllable in the middle or end of a word. Matras have been assigned codes so that a computer program can quickly identify a syllable boundary in a text string .For Example the in the name மணிவண்ணன் even Tamil child could say there are only 6 characters, but in Unicode there are nine characters as follows:

ம ண □ வ ண □ ண ன □

So the number of characters is 9. This will lead to problems in parsing.

The vowel consonants are not glyphs. They are characters. Consonant + Vowel = Vowel Consonants

For example க் + இ = கி

Unicode provides the rendering. □ + □ = கி

This has no meaning in Tamil. This type of rendering does not help simple character parsing. And thus the current implementations of Unicode support seem to concentrate mainly on data entry and not really any text processing.

IV. OVERVIEW OF SORTING

Even though you might think you understand all the issues involved with sorted lists, users of world-ready applications might have very different expectations of what constitutes a "sorted" list. Not only does alphabetical order vary among languages, but conventions for sequencing items in dictionaries and phone books can also be quite different.

In Swedish, for example, some vowels with an accent sort after "Z," whereas in other European countries the same accented vowel comes right after the nondiacritic vowel. Languages that include characters outside the Latin script have special sorting rules. The Asian languages have several different sort orders depending on phonetics, radical order, number of pen strokes, and so on.

String sorting and comparison are language-specific. Even within languages based on the Latin script, there are different composition and sorting rules. Thus do not rely on code points to do proper sorting and string comparison.

Some features of linguistic sorting are:

- A language's writing system will determine what influences the sort order of the language. For example, a sort order for Russian would be based on Cyrillic letters and possibly diacritics, but a sort order for Japanese might be based on the number of strokes it takes to draw a character.
- Linguistic sort orders are different than the Unicode code point order.
- Languages that use the same script often have different linguistic sort orders.
- A sorting element (such as a character) can be the combination of more than one Unicode code point. For example, □□□□(க+ □ + □□ + □) it takes four characters, but it should be a single sorting element in Tamil.

A. *SORTING TAMIL WORDS*

Lexical ordering according to the specified order of the letters of Tamil has been a major issue and this specific problem has not been given sufficient attention by those developing standards for Tamil. The Unicode assignment for Tamil is a hopelessly mangled set of the letters but the claim is that Unicode does not purport to preserve lexical ordering!

There is some confusion about the correct lexicographic ordering of the vowels as seen from the ordering given in different dictionaries.

In some cases, the two support vowels "am" and "aha" are placed before the first vowel "a". The generic consonant (a consonant without a vowel) is sometimes placed ahead of its combinations with the vowels. As per sorting Tamil alphabetical sorting order is concerned vowel comes first and followed by a pure consonant and vowel consonant [7]. As to which scheme is correct is a debatable issue.
Algorithm for tamil word sorting

Step 1: Create collator for Tamil.

Collator is a class that is used to specify the lexicographic ordering (i.e. alphabetical order) of the alphabets in Tamil. In java the package text has class Collator, and RuleBasedCollator. The Interface Comparator which permits to select RuleBasedComparator in which the sort order for Tamil can be provided so that the needed order can be used.

Such as

String rule = "< '\u0be6' < '\u0be7' < '\u0be8' < '\u0be9' < '\u0bea' < '\u0beb' < '\u0bec' < '\u0bed' < '\u0bee' < '\u0bef' "+ "< '\u0bf0' < '\u0bf1' < '\u0bf2' < '\u0bf3' < '\u0bf4' < '\u0bf5' < '\u0bf6' < etc...

These are the Unicode equivalent for tamil characters. The class Collator has a function compare to compare two strings.

Step 2: Create comparator.

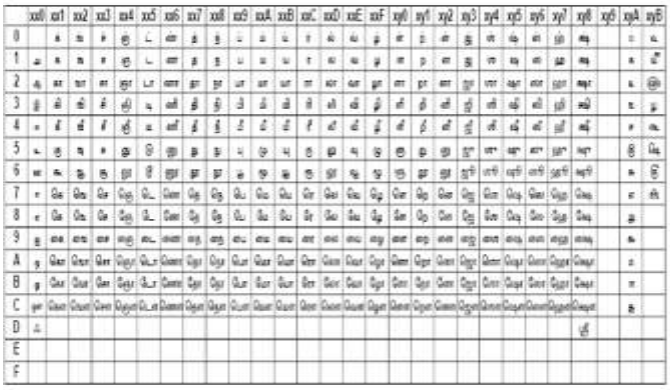
Comparator is a class that is used to specify whether a given word is less than, equal or greater than another word using Edit distance. The difficulty is with the non-uniform multi-character representation of an alphabet. E.g. the Tamil alphabet '□□' takes two characters (க + □) while 'க' is just one character and □□□□(க+ □ + □□ + □) takes 4 characters. As a consequence, if two words differ by say, one letter it may differ by one to 4 characters. So, you need to find alphabet boundaries and then compare the two alphabets instead of blindly comparing characters. The comparator for Tamil can created based on the Collator created by already. By using the below statement

Col = new RuleBasedComparator(rule);
where rule is a string which specifies the sort order of characters. This will return the collector object which can be used for sorting like Col.compare(strinr1,string2).

V. SOLUTIONS PROPOSED BY TAMILNADU GOVERNMENT

16-bit Tamil All Character Encoding(TACE_16) is proposed by TACE-16 ask force[4].

Which is shown in Figure



Advantage of this coding

- The encoding encompasses all characters that are found in Tamil language.
- The encoding is very efficient to parse for example

- By using simple arithmetic operation the characters can be parsed.
- Sorting and searching is very simple since the collation is sequential in accordance with code value.
- The encoding is unambiguous.

VI. CONCLUSIONS

Tamil a classical language has a very large population. Coding characters of any language has to take its culture, structure and heritage. The Unicode specification for Tamil does not considered the language's important features. Coding is the basis for all application generations based on the specific language. This paper discussed the issues of sorting words of Tamil language. Also shows the specification recommendations by Tamil Nadu government and its advantages.

REFERENCES

[1] *http://www.unicode.org*

[2] The second Tamil Internet Conference held in Febrary 1999.

[3] http://srinix.wordpress.com/2007/08/29/tutorial-how-to-store-utf8-indian-language-data-in-mysql/

[4] *http://www.hotfrog.in/companies/C-DAC-Gist-PACE-Mulitlingual-Computer-Training*

[5] Script grammar for tamil language developed by C-DAC

[6] *C. Sureshkumar and T. Ravichandran, Handwritten Tamil Character Recognition and Conversiom using Neural Network, International Journal on Computer Science and Engineering, vol. 02, No. 07, 2010, 2261-2267.*

[7] Tamil Style Guidedownload.microsoft.com/download/5/0./tam-tam-StyleGuide.pdf

[8] K. Rajan, M.Ganesan and V.Ramalingam, Tamil Text Analyse, Tamil Net 2003, Chnnai, Tamilnadu, India, pp: 38-44.