SCRIPT GRAMMAR FOR TAMIL LANGUAGE

Prepared by

Technology Development for Indian Languages (TDIL) Programme of DIT, GoI in association with

Centre for Development of Advanced Computing (C-DAC)

Table of Contents

0. INTRODUCTION

The term **script grammar** refers to the behaviour pattern of the writing system of a given language. Languages which have written representations do not use a haphazard manner of storing the information within the system, but use a coherent pattern which is similar to the linguistic grammar of a given language. With the help of specialists (not necessarily linguist) who work in the area of the written representation of the language, the manner in which the shapes of the characters of the language and the representation of the conjunct forms is provided. In other words the Script Grammar deals with the surface structure of the language and tries to provide the best possible "fit" for shapes and their representation. Since this is a highly subjective issue, the shapes provided here are recommendations at the best and conform to the perception of the mandating body/evaluators who consensually arrive at the "best possible fit" which is acceptable to a majority of users. An example will make the above clear. Although Marathi and Hindi share the same script Devanāgarī, not only do they not share the same character inventory but in addition the representation of certain characters is different. Thus the Marathi /la/ is different from the Hindi /la/ in so far as the placement of the stem is concerned Hindi ल Marathi ल. This ensures that the Script Grammar conforms to the language in question and provides the character shapes acceptable to a given user community. It should be noted that this does not mean monotony. The Hindi and Marathi /la/ can have a variety of forms once the intrinsic structure of the character is determined.

Script Grammar is the term used to define:

- the writing system used to inscribe a given language
- the history of the script and language (wherever available)
- the syllabic structure of the writing system of the language
- the rule ordering of the characters within the syllable (akshar)
- description of the syllabic clusters
- collation order of the characters: lexical / dictionary sorting order

1. OBJECTIVES OF SCRIPT GRAMMAR

The Objectives of the script grammar for each language can be divided into two major parts:

Societal:

- Provide a visual representation of shapes that are deemed to be in conformity with the perception of a given community

- Ensure thereby that this perception is safe-guarded

- Through wide-spread dissemination and creation of appropriate tools ensure that within the given linguistic community, all media tries to adopt the given shape.

Technical:

- Classify the language in terms of its ISO and also whether it belongs to the Abjad, Akshar (Alphasyllabary) class.

- Provide an inventory of the characters pertinent to the language and classify the same in terms of their taxonomy.

- As a corollary determine whether the inventory is in conformity to the Syllable formalism as stipulated in ISCII'91 and subsequently adopted by Unicode.

- Since Brahmi is written from left to right, and since certain characters do not follow the linear L to R order, provide an inventory of displaced catenators i.e. characters such as Mātrās that concatenate to the Consonant

- Propose the best shape representation of the individual characters as well as of the ligatures used within a given script. As a corollary request the expert(s) to identify the largest possible strings of such ligatures.

- Finally provide the collation order pertinent to that script/language, which would be of great utility to high-end NLP as well as to CLDR's in the pertinent language. The collation order for Marathi is different from Hindi although both languages share the same script. Thus in Marathi क्ष, ज्ञ are placed at the end of the consonant inventory i.e. after ह in the sort order. In Hindi क्ष is sorted along with क and ज्ञ with ज

2. END USERS FOR SCRIPT GRAMMAR

The script-grammar specific to a given language can be used by a large number of users.

- Most importantly it can be used by font developers desirous of developing a font which is compliant with the perception of the characters and ligatures of a language by its user community.

- Certain features of the script grammar such as the shapes can also be used for testing OCR and OHWR. Similarly information regarding Ligatures as well as collation order can help in high-end NLP work such as detecting invalid combinations, correct implementation of syllable structure, prediction routines to name a few. Information regarding collation and character sets can be also used for CLDR.

- They allow the font designer to design a font which is in compliance with the norms and standards of that particular script. A major problem which will be dealt with in the template is one of ligatures. The final list of ligatures defined by the script grammar allows the font designer to write specific rules for such glyphs.

- It permits the software developer to design and implement the keyboard and the input mechanism which will meet the requirement of the particular linguistic community.

- The collation or sort order as described in a Script Grammar permits the software developer to write software functions/ routines for sorting data in all applications.

- Script Grammars are equally important for keyboard design, especially when supplemented by frequency data from a corpus.

As can be seen the script grammar has a wide range of use and can be of utility to font developers, Indian language developers and linguists in the area of computation.

3. SCOPE

This script grammar document contains following information about the language and the script used for writing the language.

1. Name of the language and its representation in the 3 letter mnemonic as per ISO 639.1 & 639.3 standard.
2. Script used to inscribe the given language
3. The structure of the script used for writing the language
    - Rule ordering of the characters within the syllable formation is a language
    - Description of the syllabic clusters of the script
    - Collation order of the characters: lexical / dictionary sorting order
    - Compliance of the script with Unicode.

These will be treated within the relevant sections of the script grammar

4. TERMINOLOGY[1]

Abjad: A writing system in which each symbol always or usually stands for a consonant. The long vowels are indicated. However the short vowels are rarely marked and the reader needs to supply these. Example: Urdu written in Perso-Arabic Script is an example of this writing system.

Abugida: also called an alphasyllabary, is a segmental writing system in which consonant–vowel sequences are written as a unit: each unit is based on a consonant letter, and vowel notation is obligatory but secondary[2]

Akshar: see **Abugida**

Allographs: Variants of the representation of a character. Thus ae and æ [U+00E6] in Latin alphabet are allographs.

Allo-Script: The term relates to languages which share a common script. Thus Devanāgarī is used to write 9 official languages. However these languages do not use the same set of characters. Thus Marathi uses the retroflex lla ळ [U+ 0933] which Hindi does not use. Flaps used in Hindi ड़ [U+095C] ढ़ [U+095D] are not used in Konkani. These sub-sets of scripts based on a single "matricial" script are termed as allo-scripts.

Alphabet: A set of letters used in writing a language. Example: The English Alphabet.

Aspirated consonant: A consonant which is pronounced with an extra puff of air coming out at the time of release of the oral obstruction. This has a sound of an extra "h".

Basic alphabet: The minimal set of letters which can be used for uniquely encoding every word of a language. The basic alphabet for English consists of only the upper-case letters A-Z

Catenators: Also termed as Concatenators are characters which are concatenated to another character. In the Brahmi script these are the Mātrās or Vowel modifiers which are adjoined to the consonant and add a vocalic value to the consonant.

Conjunct: The Indic scripts are noted for a large number of consonant conjunct forms that serve as orthographic abbreviations (ligatures) of two or more adjacent letterforms. This abbreviation takes place only in the context of a consonant cluster….Under normal circumstances, a consonant cluster is depicted with a conjunct glyph if such a glyph is available in the current font. In the absence of a conjunct glyph, the one or more dead

---

[1] As in the case of the BIS Document, in order to make the terminology accessible for all readers, examples have been chosen from English/Latin scripts, wherever possible. Some definitions have been excerpted from the BIS ISCII91 document and suitably modified where necessary.

2 Wikipedia definition

consonants that form part of the cluster are depicted using half-form glyphs. In the absence of half-form glyphs, the dead consonants are depicted using the nominal consonant forms combined with visible virama signs.[3]

Consonant: A letter representing a speech sound in which the breath is at least partly obstructed,

Diacritic:A mark added to a letter which distinguishes it from the same letter without a mark, usually having a different phonetic value or stress.

Displaced Catenator: (see Catenator) Within the Brahmi script, the writing system is linear and moves from left to right. However in the case of some catenators this rules is not observed and the catenator (wholly or partially) is placed to the right of the consonant to which it relates. The short vowel I in Devanāgarī is an example of a displaced catenator.

Display composing: The process of organizing the basic shapes available in a font in order to display (or print) a word.

Display rendition: The process by which a string of characters is displayed (or printed). In this process several consecutive characters may combine with each other on the screen. The sequence of display of the characters may become different.

Eyebrow repha: (See **Eyelash ra)**

Eyelash ra: The eyelash ra is used in Konkani, Nepali and Marathi. It is treated as different from the र् (repha) by certain linguists. While the former is treated as a flap, the latter is a continuant trill (*cf.*, Kalyan Kale and Anjali Soman. 1986). There are cases in Marathi of minimal pairs such as: आचार्यास "to the teacher" vs. आचाऱ्यास "to the cook" or दर्या /darya/ "ocean" vs. दऱ्या /darya/ "valleys".

Font: A set of symbols used for display or printing of a script in a particular style.

International numerals: The conventional 0 to 9 digits used in English for denoting numbers. these are also known as Indo-Arabic numerals (to differentiate them from the Roman numerals like IX for 9).

Latin alphabet: The alphabet used for writing the language of ancient Rome. Also known as the Roman alphabet. The alphabet is used today for writing English and European languages.

Letter: A character representing one or more of the simple or compound sounds used in speech. It can be any of the alphabetic symbols.

---

[3] Unicode 6.0 Chapter 9.0 pp 6-7

Ligature: (see **Conjunct**)

Nasal consonant: A consonant pronounced with the breath passing through the nose. Example *m n* in English.

Nasalized vowel: A vowel pronounced with the breath passing both through the nose and the mouth. In Indian scripts this is denoted by a Chandrabindu and gives the vowel/vowel modifier over which it placed a nasal value. Example: जाँच

Phonetic alphabet: An alphabet which has direct correspondence between letters and sounds Example: The International Phonetic Alphabet..

Pure consonant: A consonant which does not have any vowel implicitly associated with it.

Rafar: A special case of a ligature constituted by the adjunction of ra followed by a halanta to consonant. The resultant combination places the ra on top of the consonant to which it is adjoined. In case the consonant itself is adjoined to another consonant, the rafar is placed above the consonant e.g. र्+क र्क , र्+घ्+य र्घ्य

Rakar: A special case of a ligature constituted by the adjunction of a consonant followed by a halanta to ra. In a large number of Brahmi scripts the ra is adjoined to the stem of consonant to which it relates. In the case of consonants which have no stem such as the dental retroflexes in Devanāgarī, the rakar is placed below the consonant to which it relates.

Repha: (see **Rafar)**

Roman script: The script based on the ancient Roman alphabet, with the letters A-Z and additional diacritic marks. Used for writing a language which is not usually written in the Roman alphabet.

Script: A distinctive and complete set of characters used for the written form of one or more languages.

Script numerals: The 0 to 9 digits in a script, which have shapes distinct from their international counterparts.

Syllable: A unit of pronunciation uttered without interruption, forming whole or part of a word, and usually having one vowel or diphthong sound optionally surrounded by one or more consonants

Transliteration: Representation of words with the closest corresponding letters in an alphabet of a different language.

Vowel: A letter representing a speech sound made with the vibration of the vocal cords, but without audible obstruction

Vowel sign: A graphic character associated with a letter, to indicate a vowel to be associated with that character (Mātrā in Hindi).

5. PHILOSOPHY AND UNDERLYING PRINCIPLES

The script grammar is based on the following principles:

1. The Grammar aims to depict the surface grammar of the written language: the manner in which characters as well as conjuncts are depicted

2. Where a given script admits many languages, it is pre-suppose that such languages will prescribe different representations for a given shape or conjunct according to the perception of the native users of that language

3. Corollary to the above the result is a script and allo-scripts i.e. a given script shared by many languages is not uniformly deployed across all the languages but is subject to variations and modulations.

4. The term Grammar is used here in a non-normative sense: what is prescribed is in the form of recommendations provided by experts who visualize the shape of the given script in their mother tongue in a specific manner. Subjective variations may occur[4]

5. The Grammar is limited to its synchronic use i.e. the manner in which a given language as of today admits a character set within the script used to write it. It is not diachronic or historical in nature and does not study the evolution of the given script across centuries.

[4] It is recommended that such variations be culled by placing the Grammars of different scripts in public review.

6. SCRIPT GRAMMAR STRUCTURE

The script grammar provided below has the following parts.

Part 6.1. deals with peripheral elements such as the ISO of the language, the writing system used: (Alphasyllabic) Abugida or Abjad.

Part 6.2. treats of the syllabic structure. It verifies whether the character set of the language complies with the ISCII syllabic structure and if not which cases are not compliant.

Part 6.3 is the script grammar proper and describes the character set as well as the conjunct shapes of the given script along with the collation order

6.1. PERIPHERAL ELEMENTS OF THE SCRIPT GRAMMAR

These constitute the elements that are peripheral to the Script Grammar. The main parameters considered are the mnemonic and name of the language (needed for CLDR and also for language tags), the writing system used to inscribe the language and wherever possible a short history of the language.

6.1.1. Name of the language and its representation in the 3 letter mnemonic as per ISO 639.1. & 639.3

Name of the Language: TAMIL
ISO Mnemonics: *tam*
This refers to a one line description of the language and its mnemonic representation as per the ISO. In the case of Tamil, the above information is pertinent.

6.1.2. Identification of the writing system(s) used to inscribe the given language

Tamil is written using the Tamil script. It is an alphasyllabary with the akshar as its core.
This is a one line description of the script used to write the language. However in case the language uses more than one script, all the scripts in question are specified, provided these constitute the official language of the given state.

All scripts derived from Brahmi are Abugidas i.e. syllabary driven systems. The main features of Abugidas are as under:

- The consonant has an implicit vowel built-in which is normally the schwa.
- The inherent vowel can be modified by the addition of other vowels or muted by a diacritic termed as a Virama or Halanta
- Vowels can be handled as full vowels with a vocalic value
- In Tamil the akshar is generally a Single consonant or a Consonant followed by a Pulli (Halanta), but in rare cases in Grantha script, when two or more consonants join together they form ligatures. These are only two in number: ஸ்ரீ, க்ஷ
- Abugidas/Alphasyllabaries because of their syllabic structure require a special description which is the subject of the discussion in 6.2. below.

6.1.3. Amendments needed in Unicode for Tamil language

The experts do not feel the need for addition of any characters to the Unicode Code-block..

6.2. CONFORMITY TO THE SYLLABLE STRUCTURE

Tamil language complies with the syllable (akshar) structure described above. It can admit up to 3 consonant clusters.

Alphasyllabaries are determined by the notion of the syllable or the Akshar. The compositional grammar of the syllable determines it well-formedness. This is through a series of formal constraints based on a Backus-Naur Formalism which is given below. The syllable (akshar), first defined in the ISCII document (1991), identifies the following character 'sub-sets' for the purposes of identifying the syllable (akshar). In what follows the syllable analysis will be restricted to Tamil, which is a special case.

(C) Consonants

க			ங	
ச		ஜ	ஞ	
ட			ண	
த			ந	ன
ப			ம	
ய	ர	ற	ல	ள
வ	ழ	ஷ	ஸ	ஹ

(V) Vowels

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ

(M) Mātrās or Vowel Modifiers

ா	ி	ீ	ு	ூ	ெ	ே	ை	ோ	ொ	ௌ

(D) Diacritics

ஃ Ayutham	Ayutham, denoted by two dots placed above the other. For example,

(H):Pulli /Halanta ◌் - Halant used in most writing systems of the Indian subcontinent to signify the lack of an inherent vowel.

(N)[5] Nukta ◌़ - is Not used in Tamil

Each of these sub-types has its restrictions in terms of what can precede or follow it within a syllable (akshar), as shown in the table below:

PRECEDED BY	SUBTYPE	FOLLOWED BY
-, H	**C**	M,D,H
-	**V**	D
C	**M**	D
C,V,M	**D**	-
C	**H**	C

C can be preceded by H or no subtype and followed by any one of the following: M,D,H
V can be preceded by no subtype and followed by D but not by another sub-type.
M can be preceded by C and followed by D.
D can be preceded by C, V, M and followed by no other subtype. It closes the syllable (akshar).
H can be preceded by C alone and followed only by C and no other sub-set.

6.2.1. Syllable (akshar) Types

The formalism defines the syllable (akshar) in terms of both what can constitute a syllable (akshar) and what cannot. A valid syllable (akshar) as per this definition can be of only two types:

1. A vowel syllable (akshar): a full vowel.
2. A consonant syllable (akshar): a full consonant (having a weak vowel or a mātrā)

The three other subsets viz. Mātrās, Diacritics, Halanta cannot constitute a syllable (akshar) by themselves or in combination among themselves.

6.2.1.1. The Vowel syllable (akshar) is of the following types:

6.2.1.1.1. A pure vowel all by itself: அ, /a/ ஆ /ā/ etc.
6.2.1.1.2. A vowel followed by a modifier i.e. Ayutham: ஃ/எஃகு/

[5] The nukta is a small dot placed under a character in Northern scripts to show that they are flapped or for deriving 5 other consonants required for representing loaned characters from Urdu क़,ख़,ग़,ज़,फ़

6.2.1.2. The Consonant syllable (akshar) can be of the following types:

6.2.1.2.1. A full consonant (with or without Nukta) i.e. with the inherent vowel : க : /ka/

6.2.1.2.2. A consonant followed by a mātrā i.e. the inherent vowel being substituted by another vowel: கீ/kii/

6.2.1.2.3. A consonant followed by a mātrā and a modifier: முஃடீது /**muk**tiithu[6]/.

6.2.1.2.4. A consonant cluster i.e. a dead or half consonant (Consonant+Halanta/Pulli) followed by a full consonant followed optionally by a mātrā.In Tamil this is not exists but in Grantha loans in Tamil only two characters are admissible: க்+ஷ = க்ஷ kSha, ஸ்+ரீ= ஸ்ரீ /sri/

The above permutations and combinations result in major syllable (akshar) types.

This means that theoretically the following forms can be postulated:

1. Vowel Set: With the Vowel as the node.
   V VD (Only Ayutham)
2. Consonant set: With the Consonant as the node (an implicit or modified vowel is pre-implied).

Node	Mātrā	Modifier	Mātrā+Modifier
C	CM	CD	-
CHC	CHCM	-	-
CHCHC	CHCHCM	-	-
-	-	-	-

A total number of 7 theoretical syllables is therefore possible. It will be seen that the written syllable (akshar) is not very different in structure from the phonetic syllable and that the movement from the written to the spoken levels is made feasible by application of certain rules.

This formal structure of the syllable (akshar) explained above is common to all Brahmi based scripts (with a few variations). It will form the basis of an exhaustive description of the characters as well as their ligatural representations.

[6] This is an example from archaic Tamil.

6.3 SCRIPT GRAMMAR PROPER

This section lays down in detail the different parameters of the Script Grammar for Tamil. These are:

6.3.1. The Character Set of Tamil.
6.3.2. The Consonant mātrā combinations of Tamil.
6.3.3. The Ligature Set of Tamil.
6.3.4. Collocation Order of Tamil
6.3.5. Cardinal Numbers used in Tamil.

6.3.1. The Character Set of Tamil.

This section provides detailed information about the characters in the language and the list of the same and also more importantly shows the manner in which the character is to be written. Each subsection comprises therefore two parts: the basic character set and the shape each character should have, as mandated by the experts who have designed the script grammar of Tamil.

This comprises the following:
6.3.1.1. The Consonant Set
6.3.1.2. The Vowel Set
6.3.1.3. The Mātrā Set
6.3.1.4. Displaced Catenators
6.3.1.5. Shape of the combination of ra (rakar,repha)
6.3.1.6. The Set of Diacritics
6.3.1.7. Numerals
6.3.1.8. Punctuation marks
6.3.1.9. Other symbols
Each of these will be analysed in detail:

6.3.1.1. The Consonant Set

The Consonant set of Tamil comprises the following characters:

Basic Consonant inventory arranged as per their Vargas.

	-voiced -aspirated	-voiced +aspirated	+voiced -aspirated	+voiced +aspirated	Nasal
Velar	க	-	-	-	ங
Palatal	ச,ஜ	-	-	-	ஞ
Retroflex	ட	-	-	-	ண
Dental	த	-	-	-	ந
B-labial	ப	-	-	-	ம

Other consonants

ய	ர	ற	ல	ள	ன	வ	ழ	ஷ	ஸ	ஹ	ஜ[7]	

The exact shapes as desired by the experts are provided in the table below:

	-voiced -aspirated	+voiced -aspirated	Nasal
Velar	க		ங
Palatal	ச,ஜ		ஞ
Retroflex	ட		ண
Dental	த		ந
B-labial	ப		ம

Other consonants

ய	ர	ற	ல	ள	ன
வ	ழ	ஷ	ஸ	ஹ	

6.3.1.2. The Vowel Set

The Vowel set of Tamil is as under:

அ	TAMIL LETTER A
ஆ	TAMIL LETTER AA
இ	TAMIL LETTER I
ஈ	TAMIL LETTER II
உ	TAMIL LETTER U
ஊ	TAMIL LETTER UU
எ	TAMIL LETTER E
ஏ	TAMIL LETTER EE
ஐ	TAMIL LETTER AI
ஒ	TAMIL LETTER O

[7] Normally considered a Grantha Consonant

ஓ	TAMIL LETTER OO
ஔ	TAMIL LETTER AU

As per expert recommendations the character set should be written as under:

அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஓ ஒ ஔ

6.3.1.3. The Mātrā Set

The Mātrā (Vowel Modifier Set) of Tamil is as under:

Mātrās Names	Mātrās Sign	Where is it used ?	Consonant Shapes formed
1. Tamil sign AA	ா	ஆ	க் + ஆ = கா
2. Tamil sign I (stands to the left of the consonat)	ி	இ	க் + இ = கி
3. Tamil sign II	ீ	ஈ	க் + ஈ = கீ
4. Tamil sign U	ு	உ	க் + உ = கு
5. Tamil sign UU	ூ	ஊ	க் + ஊ = கூ
6. Tamil sign E	ெ	எ	க் + எ = கெ
7. Tamil sign EE	ே	ஏ	க் + ஏ = கே
8. Tamil sign AI	ை	ஐ	க் +ஐ = கை
9. Tamil sign O	ொ	ஒ	க் + ஒ = கொ
10. Tamil sign OO	ோ	ஓ	க் + ஓ = கோ
11. Tamil sign AU	ௌ	ஔ	க் + ஔ = கௌ

As per expert recommendations the character set should be written as under:

ா ி ீ ு ூ ெ ே ை ோ ொ ௌ

The vowel modifier /ayutham ஃ is used in few words only: எஃகு , அஃது , இஃது

6.3.1.4. Displaced Catenators

Under normal circumstances Vowel Modifiers also known as catenators (since they concatenate to the preceding consonant) in Brahmi based scripts are written from left to right in linear order (with the exception of Consonant stacks). 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

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

6.3.1.5. Shape of the combination of ra (rakar,repha)

In Tamil the concept of rafar/repha and rakar does not exist

6.3.1.6. Diacritics

These are as under in the case of Tamil:

ஃ - Visarga/Aytham

6.3.1.7. Numerals

Following are the numbers used in Tamil language. 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.

Numeral Shapes	Explanation
௦	Tamil Digit Zero
க	Tamil Digit One
உ	Tamil Digit Two
௩	Tamil Digit Three
௪	Tamil Digit Four
௫	Tamil Digit Five
௬	Tamil Digit Six
௭	Tamil Digit Seven
அ	Tamil Digit Eight
௯	Tamil Digit Nine

6.3.1.7.1 Numerics[8]

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

The shapes of some of these numbers vary depending upon the user.

<u>Tamil Fraction</u>

க ¼ ஒன்று

ஏ ¾ முக்கால்

[8] As mentioned in Unicode
[9] Not mentioned in Unicode needs to be proposed

வ ¼ கால்

½ அரை

8.1.8. OTHER SYMBOLS (religious, currency markers etc. included in Unicode)

Tamil symbols

0BF3 ௳ TAMIL DAY SIGN
0BF4 ௴ TAMIL MONTH SIGN
0BF5 ௵ TAMIL YEAR SIGN
0BF6 ௶ TAMIL DEBIT SIGN
0BF7 ௷ TAMIL CREDIT SIGN
0BF8 ௸ TAMIL AS ABOVE SIGN

6.3.1.8. Punctuation Markers

Tamil uses punctuation markers from the Latin set. such as . , ; : “ ‘ ( ) [ ] etc.
A list of punctuations is provided below:

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	{ }

6.3.1.9 Other Symbols

These are religious, currency markers etc. included in Unicode:

0BD0 ௐ TAMIL OM (as written in Tamil)[10]

[10] Integrated in Unicode 6.0

₹: Rupee Sign as mandated by Government of India.

6.3.2. Consonant Mātrā Combinations.

These refer to the shapes generated when a Mātrā is adjoined to the Consonant. The layout of these is in the shape of a matrix where the first horizontal row refers to the active consonant and the first vertical column refers to the vowel-modifier.

Due to constraints of space and also clarity, for each class 2 tables are provided.

Table 1: க ங ச ஞ ட ண த ந ன ஜ
Table 2: ப ம ய ர ற ல ள வ ழ ஷ ஸ ஹ

Wherever there is an X it implies that the combination does not exist. For the font developer this is an indication that for this particular combination which is not possible in the language but needs to be accommodated in the font table, a simple linear combination be provided.
//e.g. Although the combination of ங+ா is possible (in certain places in earlier Tamil) it needs to be handled at the font level in the anticipation that a user could type this combination. The font would show the following: ங

The classes are as under:
6.3.2.1. refers to a simple concatenation of Consonant and Mātrā combinations.
6.3.2.2. refers to a concatenation of Consonant and Mātrā + Nasal marker combinations.
Other diacritics such as avagraha and visarga have been avoided, since these are linear in nature, are adjoined to the combination and do not in any way modify the structure of the shapes.

6.3.2.1 Consonant and Mātrā combinations.

This set refers to a simple concatenation of Consonant and Mātrā.

Consonant and Mātrā combinations Set 1

	க	ங	ச	ஞ	ட	ண	த	ந	ன	ஜ
ா	கா	ஙா	சா	ஞா	டா	ணா	தா	நா	னா	ஜா
ி	கி	ஙி	சி	ஞி	டி	ணி	தி	நி	னி	ஜி
ீ	கீ	ஙீ	சீ	ஞீ	டீ	ணீ	தீ	நீ	னீ	ஜீ
ு	கு	ங	சு	ஞு	டு	ணு	து	நு	னு	ஜு
ூ	கூ	ங	சூ	ஞூ	டூ	ணூ	தூ	நூ	னூ	ஜூ
ெ	கெ	ஙெ	செ	ஞெ	டெ	ணெ	தெ	நெ	னெ	ஜெ
ே	கே	ஙே	சே	ஞே	டே	ணே	தே	நே	னே	ஜே
ை	கை	ஙை	சை	ஞை	டை	ணை	தை	நை	னை	ஜை
ொ	கொ	ஙொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	னொ	ஜொ
ோ	கோ	ஙோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	னோ	ஜோ
ௌ	கௌ	ஙௌ	சௌ	ஞௌ	டௌ	ணௌ	தௌ	நௌ	னௌ	ஜௌ

Consonant and Mātrā combinations Set 2

This set is in continuation of set 1 which shows consonant and Matra combinations.

	ப	ம	ய	ர	ற	ல	ள	வ	ழ	ஷ	ஸ	ஹ
ா	பா	மா	யா	ரா	றா	லா	ளா	வா	ழா	ஷா	ஸா	ஹா
ி	பி	மி	யி	ரி	றி	லி	ளி	வி	ழி	ஷி	ஸி	ஹி
ீ	பீ	மீ	யீ	ரீ	றீ	லீ	ளீ	வீ	ழீ	ஷீ	ஸீ	ஹீ
ு	பு	மு	யு	ரு	று	லு	ளு	வு	ழு	ஷு	ஸு	ஹு
ூ	பூ	மூ	யூ	ரூ	றூ	லூ	ளூ	வூ	ழூ	ஷூ	ஸூ	ஹூ
ெ	பெ	மெ	யெ	ரெ	றெ	லெ	ளெ	வெ	ழெ	ஷெ	ஸெ	ஹெ
ே	பே	மே	யே	ரே	றே	லே	ளே	வே	ழே	ஷே	ஸே	ஹே
ை	பை	மை	யை	ரை	றை	லை	ளை	வை	ழை	ஷை	ஸை	ஹை
ொ	பொ	மொ	யொ	ரொ	றொ	லொ	ளொ	வொ	ழொ	ஷொ	ஸொ	ஹொ
ோ	போ	மோ	யோ	ரோ	றோ	லோ	ளோ	வோ	ழோ	ஷோ	ஸோ	ஹோ
ௌ	பௌ	மௌ	யௌ	ரௌ	றௌ	லௌ	ளௌ	வௌ	ழௌ	ஷௌ	ஸௌ	ஹௌ

6.3.2.2 Consonant and Mātrā +Nasal combinations.

No such combination exists in Tamil

6.3.3. The Ligature Set of Tamil.

Since the structure of modern Tamil is purely linear, no ligatural constructs exist. However in Tamil using Grantha two ligature exist.

e.g : ஸ்+ரீ= ஸ்ரீ, க்+ஷ = க்ஷ

The expert has however provided along with examples all possible combinations of Consonant +Pulli+Consonant as well as the few instances of Consonant + Pulli + Consonant + Pulli + Consonant

These are as under

6.3.3.1. CHC (combination of two consonanats)

The following set shows a combination of two Tamil consonants.

ப்ப்	கப்பல்
த்த்	எடுத்து
ற்ற்	இயற்றல்
ட்ட்	ஊட்டல்
ச்ச்	அச்சம்
க்க்	ஆக்கல்
ம்ம்	தும்மல்
ந்ந்	முந்நீர்
ண்ண்	பண்ணை
ய்ய்	எய்யாமை
ய்ஞ்	அய்ஞ்நூறு
ல்ல	முல்லை
ள்ள்	தள்ளாடு

ர்த்	உணர்தல்
ர்க்	அவர்கட்கு
ர்ம்	கூர்மை
ர்வ்	ஆர்வம்
ழ்ப்	உறழ்பு
ழ்த்	சூழ்தல்
ழ்வ்	கதழ்வு
ல்ப்	இயல்பு
ல்க்	அல்குல்
ள்ப்	கொள்ப
ல்வ்	கல்வி
ற்ப்	ஏற்பின்
ற்ச்	அகற்சி
ற்க்	ஏற்கும்
ட்க்	கேட்க

ட்ச்	மாட்சி
ம்ப்	அம்பல்
ம்க்	எம்கய்யர்
ந்த்	ஏந்து
ன்ப்	என்பு
ன்ற்	என்றல்
ன்ம்	என்மனார்
ண்ட்	கண்டு
ண்ம்	எண்மை
ங்ச்	அஞ்ச
வ்க்	அவ்கி
ய்ப்	செய்பு
ய்த்	செய்த
ய்க்	செய்கை
ய்ம்	சேய்மை
ய்வ்	காய்வு
ர்ப்	உணர்ப

6.3.3.2 CHCHC (combination of three consonants)

Only three consonants can exists along with two consonants.

e.g : ழ்,ய்,ர்

ய்ம்ம்	மெய்ம்மை
ய்ந்ந்	செய்ந்நன்றி
ய்ந்த்	பாய்ந்தான்
ய்ங்க்	வேய்ங்குழல்
ய்ப்ப்	வாய்ப்பு
ய்க்க்	வாய்க்கால்
ய்ச்ச்	பாய்ச்சு
ய்த்த	வாய்த்தது
ர்ந்த்	சேர்ந்தது
ர்ப்ப்	வார்ப்பு
ர்த்த்	பார்த்தல்
ர்க்க்	சேர்க்கை
ர்ச்ச்	வளர்ச்சி
ழ்ந்த்	சூழ்ந்து
ழ்ங்க்	பாழ்ங்கிணறு
ழ்ப்ப்	காழ்ப்புணர்ச்சி
ழ்த்த்	வாழ்த்து
ழ்க்க்	வாழ்க்கை

ழ்ச்ச்	வீழ்ச்சி

6.3.3.3.CHCHCHC (Combination of four Consonanats)

This combination is not found in Tamil

6.3.4 The Collation Order of Tamil.

The collation order refers to the order in which the characters in a given language are sorted. In the case of Tamil the following is the traditional sort order as determined by the experts. The order as given below is pertinent to sorting by a computer program and is compliant with CLDR as laid down by Unicode and W3C.

Collation is one of the most important features of a script grammar. 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. Within a given script, each allo-script may have a different sort-order. Thus in Devanagari the conjunct glyph क्ष is sorted along with क, since the first letter of that conjunct is क and on a similar principle ज्ञ is sorted along with ज. In Marathi, the two conjunct glyphs are given at the end of the sort order. Different scripts admit different sort orders and for all high-end NLP applications, sort is a crucial feature to ensure that the applications index data as per the cultural perception of that community. In quite a few States, sort order is clearly defined by the statutory bodies of that state and hence it is crucial that such sort order be ascertained and introduced in the script grammar.

(Modern Tamil sorting order as prescribed in dictionaries)

Ayutham: ஃ

vowels : அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ ஓ ஔ

consonants : க ங ச ஞ ட ண த ந ப ம ய ர ல வ ழ ள ற ன

granthas : ஜ ஸ ஷ க்ஷ ஹ ஸ்ரீ

Matras : ா ி ீ ு ூ ெ ே ை ொ ோ ௌ

Sort order for the Computer

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 ா ி ீ ு ூ ெ ே ை ொ ோ ௌ ்

7. REFERENCES

1. http://www.unicode.org
2. ISCII'91

8. ANNEXURES

Annexure 1: Names of experts who have contributed to the script grammar

Annexure 2: Unicode Table of Tamil

0B80 **Tamil** **0BFF**

	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0		ஐ		ர	ீ	ஒ		ய
1				ற	ி			ர
2	்	ஒ		ல	ு			ற
3	ஃ	ஓ	ண	ள				வ
4		ஔ	த	ழ				ம்
5	அ	க		வ				வௌ
6	ஆ			ஶ	ொ		௦	ய
7	இ			ஷ	ோ	ௗ	க	வை
8	ஈ		ந	ஸ	ை		உ	வ்வ
9	உ	ங	ன	ஹ			ந	ரீ
A	ஊ	ச	ப		ொ		ச	ஜ
B					ோ		ரு	
C		ஜ			ௌ		சு	
D					்		எ	
E	எ	ஞ	ம	ா			அ	
F	ஏ	ட	ய	ி			கூ	

 79

Based on ISCII 1988

Various signs

0B82 ஂ TAMIL SIGN ANUSVARA
* not used in Tamil

0B83 ஃ TAMIL SIGN VISARGA
= aytham

Independent vowels

0B85 அ TAMIL LETTER A
0B86 ஆ TAMIL LETTER AA
0B87 இ TAMIL LETTER I
0B88 ஈ TAMIL LETTER II
0B89 உ TAMIL LETTER U
0B8A ஊ TAMIL LETTER UU
0B8B ▨ <reserved>
0B8C ▨ <reserved>
0B8D ▨ <reserved>
0B8E எ TAMIL LETTER E
0B8F ஏ TAMIL LETTER EE
0B90 ஐ TAMIL LETTER AI
0B91 ▨ <reserved>
0B92 ஒ TAMIL LETTER O
0B93 ஓ TAMIL LETTER OO
0B94 ஔ TAMIL LETTER AU
= 0B92 ஒ 0B07 ௗ

Consonants

0B95 க TAMIL LETTER KA
0B96 ▨ <reserved>
0B97 ▨ <reserved>
0B98 ▨ <reserved>
0B99 ங TAMIL LETTER NGA
0B9A ச TAMIL LETTER CA
0B9B ▨ <reserved>
0B9C ஜ TAMIL LETTER JA
0B9D ▨ <reserved>
0B9E ஞ TAMIL LETTER NYA
0B9F ட TAMIL LETTER TTA
0BA0 ▨ <reserved>
0BA1 ▨ <reserved>
0BA2 ▨ <reserved>
0BA3 ண TAMIL LETTER NNA
0BA4 த TAMIL LETTER TA
0BA5 ▨ <reserved>
0BA6 ▨ <reserved>
0BA7 ▨ <reserved>
0BA8 ந TAMIL LETTER NA
0BA9 ன TAMIL LETTER NNNA
0BAA ப TAMIL LETTER PA
0BAB ▨ <reserved>
0BAC ▨ <reserved>
0BAD ▨ <reserved>
0BAE ம TAMIL LETTER MA
0BAF ய TAMIL LETTER YA
0BB0 ர TAMIL LETTER RA
0BB1 ற TAMIL LETTER RRA
0BB2 ல TAMIL LETTER LA
0BB3 ள TAMIL LETTER LLA
0BB4 ழ TAMIL LETTER LLLA
0BB5 வ TAMIL LETTER VA
0BB6 ஶ TAMIL LETTER SHA
0BB7 ஷ TAMIL LETTER SSA
0BB8 ஸ TAMIL LETTER SA
0BB9 ஹ TAMIL LETTER HA

Dependent vowel signs

0BBE ா TAMIL VOWEL SIGN AA
0BBF ி TAMIL VOWEL SIGN I
0BC0 ீ TAMIL VOWEL SIGN II
0BC1 ு TAMIL VOWEL SIGN U
0BC2 ூ TAMIL VOWEL SIGN UU
0BC3 ▨ <reserved>
0BC4 ▨ <reserved>
0BC5 ▨ <reserved>
0BC6 ெ TAMIL VOWEL SIGN E
* stands to the left of the consonant
0BC7 ே TAMIL VOWEL SIGN EE
* stands to the left of the consonant
0BC8 ை TAMIL VOWEL SIGN AI
* stands to the left of the consonant

Two-part dependent vowel signs

These vowel signs have glyph pieces which stand on both sides of the consonant; they follow the consonant in logical order, and should be handled as a unit for most processing.

0BCA ொ TAMIL VOWEL SIGN O
= 0BC6 ெ 0BBE ா

0BCB ோ TAMIL VOWEL SIGN OO
= 0BC7 ே 0BBE ா

0BCC ௌ TAMIL VOWEL SIGN AU
= 0BC6 ெ 0B57 ௗ

Various signs

0BCD ் TAMIL SIGN VIRAMA
0BCE ▨ <reserved>
0BCF ▨ <reserved>
0BD0 ▨ TAMIL OM
0BD1 ▨ <reserved>
0BD2 ▨ <reserved>
0BD3 ▨ <reserved>
0BD4 ▨ <reserved>
0BD5 ▨ <reserved>
0BD6 ▨ <reserved>
0BD7 ௗ TAMIL AU LENGTH MARK

Reserved

For viram punctuation, see the generic Indic 0964 and 0965.

0BE4 ▨ <reserved>
→ 0964 । devanagari danda

0BE5 ▨ <reserved>
→ 0965 ॥ devanagari double danda

Digits

0BE6 ௦ TAMIL DIGIT ZERO
0BE7 ௧ TAMIL DIGIT ONE
0BE8 ௨ TAMIL DIGIT TWO
0BE9 ௩ TAMIL DIGIT THREE
0BEA ௪ TAMIL DIGIT FOUR
0BEB ௫ TAMIL DIGIT FIVE
0BEC ௬ TAMIL DIGIT SIX
0BED ௭ TAMIL DIGIT SEVEN
0BEE ௮ TAMIL DIGIT EIGHT
0BEF ௯ TAMIL DIGIT NINE

Tamil numerics

0BF0 ௰ TAMIL NUMBER TEN
0BF1 ௱ TAMIL NUMBER ONE HUNDRED
0BF2 ௲ TAMIL NUMBER ONE THOUSAND

Tamil symbols

0BF3 ௳ TAMIL DAY SIGN
0BF4 ௴ TAMIL MONTH SIGN
0BF5 ௵ TAMIL YEAR SIGN
0BF6 ௶ TAMIL DEBIT SIGN
0BF7 ௷ TAMIL CREDIT SIGN
0BF8 ௸ TAMIL AS ABOVE SIGN

Currency symbol

0BF9 ௹ TAMIL RUPEE SIGN

Tamil symbol

0BFA ௺ TAMIL NUMBER SIGN