

9. Tamil

Tamil is a Southern Dravidian language [51]. It is currently spoken by about 77 million people around the world with 68 million speakers residing in India mostly in the state of Tamil Nadu. It is one the official language in India, Sri Lanka and Singapore.

Tamil language is written in Tamil script which descends from South Brahmi script and dates back to 500 BC [8, 52]. It is a syllabic writing system, like other Indic systems, written without a top-line characteristic of South Brahmi scripts and different from the North Brahmi scripts.

9.1. Writing System

9.1.1. Character Set

Tamil has fewer characters, a total of 18 consonants, 12 independent vowels and 11 dependent vowels (schwa, the twelfth vowel, is implied with each consonant and thus not written explicitly). These are given in Figures 9.1 and 9.2.

அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ ஓ ஔ
Independent Vowels
ா ி ி ா ஶ ே ை ொ ோ ௌ
Dependent Vowels

Figure 9.1. Tamil Vowels

க ங ச ஞ ட ண த ந ப ம ய ர ல வ ழ ள ற ன

Figure 9.2. Tamil Consonants

Tamil borrows five special consonants to represent Sanskrit loan words. These are known as Grantha characters [52] and are shown below.

ஐ ஸ ஷ ஹ க்ஷ

Figure 9.3. Additional Tamil Consonants (For Loan Words)

Tamil also has two special characters, Virama and Aytam. Virama is used to cancel the implicit vowel with each consonant. Aytam causes spirantization, turning [p] into [f] and [j] into [z]. Their use is shown in Table 9.1.

Table 9.1. Virama and Aytam Characters in Tamil

Name	Glyph	Usage
Virma	◌̣	◌̣
Aytam	◌̣◌̣	◌̣◌̣

Tamil has its own set of numerals but these are rarely used, and normally 0, 1, 2... 9 are used.

௦ ௧ ௨ ௩ ௪ ௫ ௬ ௭ ௮ ௯

Figure 9.4. Tamil Numerals

In addition, Tamil has special characters as multipliers for 10, 100 and 1000, shown in Figure 9.5. Thus, ௩, ௧௩, ௩௧, ௩௧௩ represent 3, 13, 30 and 33 respectively [5].

௧௦ ௧௦௦ ௧௦௦௦

Figure 9.5. Tamil Multipliers for 10, 100, and 1000

Moreover, there are some special symbols for day, month, year, debit, credit, as above, rupee and numeral in Tamil shown in Figure 9.6 below [4].

௨ மீ (௨௦) ௫ ௬ ௭ ௮ ௯ ௧௦ ௧௧ ௧௨

Figure 9.6. Special Signs

9.1.2. Script Details

Tamil is written from left to right. Letters are uncased and are grouped based on their place and manner of articulation, like other Indic scripts.

9.1.2.1. Consonants and Vowels

Tamil has a syllabic writing system. Vowels and consonants are not represented as an individual unit, rather as syllabic units in which consonant has an inherent [a] vowel, if not otherwise specified. For example Tamil Letter KA க has [ka] sound. In case the consonant is to be articulated without a vowel sound, e.g. in a cluster or at the end of a word, Virama is placed at top of the consonant to cancel the [a] sound. So க̣ has [k] sound.

Independent vowels are used for syllables which do not have an onset consonant and thus start with a vowel. For all syllables which have an onset consonant, dependent vowels attach with this

consonant. If the consonant is followed by a vocalic sound different from [a], the appropriate dependent vowel mark is attached before, after, above or below the consonant (though it always logically follows the consonant). In some cases the vowel splits into two halves and is placed around the consonant. Table 9.2 below shows these cases.

Table 9.2. Dependent Vowels with the Consonant [h]

Consonant + Dependent Vowel	Joined Form	Comment
ஹ + ெ	ஹெ	Connects to the left of consonant
ஹ + ா	ஹா	Connects to the right of consonant
ஹ + ீ	ஹீ	Connects to the top of consonant
ஹ + ஶ	ஹூ	Connects to the base of consonant
ஹ + ெள	ஹௌ	Wraps around the consonant

Only one vowel can occur in a syllable, thus only a single dependent vowel can attach with a consonant and the dependent vowels can not occur with independent vowels.

9.1.2.2. *Conjunct Consonants and Consonantal Vowel Ligatures*

Tamil, unlike other Indic based languages do not form conjunct consonants except for the case of letter KSA which is formed as letter KA + Virama + SA (க+ஷ = க்ஷ). However, Tamil frequently forms consonant-vowel ligatures. Some vowels might form variety of different shapes when combined with different consonants. This is shown in the figure below as different consonants combine with the long vowel ஶ. For a more comprehensive list, see [52].

க + ஶ = க்ஷ
 ச + ஶ = ச்ஷ
 த + ஶ = த்ஷ
 ப + ஶ = ப்ஷ
 ம + ஶ = ம்ஷ

Figure 9.7. Consonant Vowel Ligatures in Tamil

9.2. *Collation*

This section elaborates on this collation sequence for Tamil and an algorithmic implementation using UCA [2].

All characters have primary level significance for collation purposes. Numerals and currency symbols are given smallest weight. These are followed by modifiers, independent vowels, consonants and finally dependent vowels. However, before collation can be applied, some text processing is required. These details are also given below.

9.2.1. Text Processing

9.2.1.1. Reordering

As shown in Table 9.1 above, independent vowels combine with consonants in different ways. In hand-written orthography, old type-writers and early proprietary Tamil fonts, the vowels that append to the left are written first followed by a consonant. The vowels that append to the right, above or below are written after the consonant. However, the logical order in both cases is the same, i.e. the consonant is followed by the vowel. The more recent font formats and fonts based on Unicode follow the logical order of typing. However, for the legacy fonts, re-ordering will be required before string comparisons can be performed for collation.

9.2.1.2. Virama

Virama is implicitly an integral part of most Indic scripts; however, its explicit use is sometimes optional. In case Virama is not written, a native speaker can still use the knowledge of the language to correctly recognize and pronounce the words. However, it is much more consistently used in Tamil. Consonants with Virama are lighter than the same consonant without it. This is not possible to do if separate collation elements are assigned to consonants and Virama, as per Unicode collation algorithm. A solution is to define a contraction corresponding to each consonant with a collation element with lesser value compared to the consonant without the Virama, as given in the collation table later.

9.2.1.3. Normalization

Many Tamil vowels are formed with two parts, one part attaching before and other after the following consonant. These and some other dependent forms of vowels can be encoded in more than one way in Unicode. As they are equivalent to each other for text processing, they have to be equated or normalized into the same composed or decomposed form. Some examples are illustrated in Table 9.3 below.

Table 9.3. Normalization in Tamil

Decomposed Form	Unicode of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form
-----------------	----------------------------	--------------------------	--------------------------

ெ + ா	0BC6 + 0BBE	ொ	0BCA
ே + ா	0BC7 + 0BBE	ோ	0BCB
ெ + ள	0BC6 + 0BD7	ௌ	0BCC
ஔ + ள	0B92 + 0BD7	ௗ	0B94

9.2.1.4. Contraction

In Tamil, a few sequences of encoded characters map onto contracted linguistic units, which have distinct role in collation, different from their constituents. These contractions need to be assigned appropriate collation elements. These include the KSA character and the consonants with Virama (as discussed earlier). Examples for these two cases are given in Table 9.4 below.

Table 9.4. Contraction to Single Collation Element from Multiple Unicodes

Glyph	Unicodes of Decomposed Form	Composed Form	Collation Element	Name
க + ஃ + ற	0B95 + 0BCD + 0BB7	கற	13B6 0020 0002	TAMIL LETTER KSA
ம + ஃ	0BAE + 0BCD	ம	1392 0020 0002	TAMIL LETTER M

9.2.1.5. Consonantal Vowel Ligatures

As discussed, in some cases when vowels combine with consonants, they form a conjoined shape which is different from simple concatenation. The formation of these consonantal vowel ligatures is a visual phenomenon and does not change the encoding or the linguistic entities in any complex way. Thus, it is not relevant for collation process.

9.2.2. Unicode Collation Elements

In order to realize Tamil collation, following collation elements need to be assigned. The UCA algorithm proposed in [2] may be applied for sorting. The realized sequence is same as recommended by [53, 54].

Table 9.5. Tamil Collation Elements

Glyph	Unicode	Collation Elements	Unicode Name
← Numerals and Various Signs →			
௦	0BE6	0E29 0020 0002	TAMIL DIGIT ZERO
௧	0BE7	0E2A 0020 0002	TAMIL DIGIT ONE

௨	0BE8	0E2B 0020 0002	TAMIL DIGIT TWO
௩	0BE9	0E2C 0020 0002	TAMIL DIGIT THREE
௪	0BEA	0E2D 0020 0002	TAMIL DIGIT FOUR
௫	0BEB	0E2E 0020 0002	TAMIL DIGIT FIVE
௬	0BEC	0E2F 0020 0002	TAMIL DIGIT SIX
௭	0BED	0E30 0020 0002	TAMIL DIGIT SEVEN
௮	0BEE	0E31 0020 0002	TAMIL DIGIT EIGHT
௯	0BEF	0E32 0020 0002	TAMIL DIGIT NINE
௧௦	0BF0	0E33 0020 0002	TAMIL NUMBER TEN
௧௧	0BF1	0E34 0020 0002	TAMIL NUMBER ONE HUNDERED
௧௨	0BF2	0E35 0020 0002	TAMIL NUMBER ONE THOUSAND
௧௩	0BF3	0E36 0020 0002	TAMIL DAY SIGN
௧௪	0BF4	0E37 0020 0002	TAMIL MONTH SIGN
௧௫	0BF5	0E38 0020 0002	TAMIL YEAR SIGN
௧௬	0BF6	0E39 0020 0002	TAMIL DEBIT SIGN
௧௭	0BF7	0E3A 0020 0002	TAMIL CREDIT SIGN
௧௮	0BF8	0E3B 0020 0002	TAMIL AS ABOVE SIGN
௧௯	0BF9	0E3C 0020 0002	TAMIL RUPEE SIGN
௨௦	0BFA	0E3E 0020 0002	TAMIL NUMBER SIGN
ஃ	0BCD	1350 0020 0002	TAMIL SIGN VIRMA
ஃ	0B83	1353 0020 0002	TAMIL SIGN VISARGA
←Independent Vowels Primary Level→			
அ	0B85	1356 0020 0002	TAMIL LETTER A
ஆ	0B86	1359 0020 0002	TAMIL LETTER AA
இ	0B87	135C 0020 0002	TAMIL LETTER I
ஈ	0B88	135F 0020 0002	TAMIL LETTER II
உ	0B89	1360 0020 0002	TAMIL LETTER U
ஊ	0B8A	1363 0020 0002	TAMIL LETTER UU
எ	0B8E	1366 0020 0002	TAMIL LETTER E
ஏ	0B8F	1369 0020 0002	TAMIL LETTER EE
ஐ	0B90	136C 0020 0002	TAMIL LETTER AI
ஓ	0B92	136F 0020 0002	TAMIL LETTER O
ஔ	0B93	1370 0020 0002	TAMIL LETTER OO
ஔள	0B94	1373 0020 0002	TAMIL LETTER AU
ஔ ள	0B92 0BD7	1373 0020 0002	TAMIL LETTER AU
←Consonants→			
க ஃ	0B95 0BCD	1375 0020 0002	TAMIL LETTER K
க	0B95	1376 0020 0002	TAMIL LETTER KA

ங்	0B99 0BCD	1378 0020 0002	TAMIL LETTER NG
ங	0B99	1379 0020 0002	TAMIL LETTER NGA
ச்	0B9A 0BCD	137B 0020 0002	TAMIL LETTER C
ச	0B9A	137C 0020 0002	TAMIL LETTER CA
ஞ்	0B9E 0BCD	137F 0020 0002	TAMIL LETTER NY
ஞ	0B9E	1380 0020 0002	TAMIL LETTER NYA
ட்	0B9F 0BCD	1382 0020 0002	TAMIL LETTER TT
ட	0B9F	1383 0020 0002	TAMIL LETTER TTA
ண்	0BA3 0BCD	1385 0020 0002	TAMIL LETTER NN
ண	0BA3	1386 0020 0002	TAMIL LETTER NNA
த்	0BA4 0BCD	1388 0020 0002	TAMIL LETTER T
த	0BA4	1389 0020 0002	TAMIL LETTER TA
ந்	0BA8 0BCD	138B 0020 0002	TAMIL LETTER N
ந	0BA8	138C 0020 0002	TAMIL LETTER NA
ப்	0BAA 0BCD	138F 0020 0002	TAMIL LETTER P
ப	0BAA	1390 0020 0002	TAMIL LETTER PA
ம்	0BAE 0BCD	1392 0020 0002	TAMIL LETTER M
ம	0BAE	1393 0020 0002	TAMIL LETTER MA
ய்	0BAF 0BCD	1395 0020 0002	TAMIL LETTER Y
ய	0BAF	1396 0020 0002	TAMIL LETTER YA
ர்	0BB0 0BCD	1398 0020 0002	TAMIL LETTER R
ர	0BB0	1399 0020 0002	TAMIL LETTER RA
ல்	0BB2 0BCD	139A 0020 0002	TAMIL LETTER L
ல	0BB2	139B 0020 0002	TAMIL LETTER LA
வ்	0BB5 0BCD	139C 0020 0002	TAMIL LETTER V
வ	0BB5	139D 0020 0002	TAMIL LETTER VA
ழ்	0BB4 0BCD	139F 0020 0002	TAMIL LETTER LLL
ழ	0BB4	13A0 0020 0002	TAMIL LETTER LLLA
ள்	0BB3 0BCD	13A2 0020 0002	TAMIL LETTER LL
ள	0BB3	13A3 0020 0002	TAMIL LETTER LLA
ற்	0BB1 0BCD	13A5 0020 0002	TAMIL LETTER RR
ற	0BB1	13A6 0020 0002	TAMIL LETTER RRA
ன்	0BA9 0BCD	13A8 0020 0002	TAMIL LETTER NNN
ன	0BA9	13A9 0020 0002	TAMIL LETTER NNNA
ஜ்	0B9C 0BCD	13AB 0020 0002	TAMIL LETTER J
ஜ	0B9C	13AC 0020 0002	TAMIL LETTER JA
ஸ்	0BB8 0BCD	13AE 0020 0002	TAMIL LETTER S

ஸ	0BB8	13AF 0020 0002	TAMIL LETTER SA
ஷ்	0BB7 0BCD	13B0 0020 0002	TAMIL LETTER SS
ஷ	0BB7	13B1 0020 0002	TAMIL LETTER SSA
ஹ்	0BB9 0BCD	13B2 0020 0002	TAMIL LETTER H
ஹ	0BB9	13B3 0020 0002	TAMIL LETTER HA
க்ஷ்	0B95 0BCD 0BB7 0BCD	13B5 0020 0002	TAMIL LETTER KS
க்ஷ	0B95 BCD 0BB7	13B6 0020 0002	TAMIL LETTER KSA
ஸ்	0BB6 0BCD	13B8 0020 0002	TAMIL LETTER SH
ஸ	0BB6	13B9 0020 0002	TAMIL LETTER SHA
← Dependent Vowels →			
ா	0BBE	13C0 0020 0002	TAMIL VOWEL SIGN AA
ி	0BBF	13C3 0020 0002	TAMIL VOWEL SIGN I
ீ	0BC0	13C6 0020 0002	TAMIL VOWEL SIGN II
ு	0BC1	13C9 0020 0002	TAMIL VOWEL SIGN U
ூ	0BC2	13CA 0020 0002	TAMIL VOWEL SIGN UU
ெ	0BC6	13CC 0020 0002	TAMIL VOWEL SIGN E
ே	0BC7	13D0 0020 0002	TAMIL VOWEL SIGN EE
ை	0BC8	13D3 0020 0002	TAMIL VOWEL SIGN AI
ொ	0BCA	13D6 0020 0002	TAMIL VOWEL SIGN O
ொ	0BC6 0BBE	13D6 0020 0002	TAMIL VOWEL SIGN O
ோ	0BCB	13D9 0020 0002	TAMIL VOWEL SIGN OO
ோ	0BC7 0BBE	13D9 0020 0002	TAMIL VOWEL SIGN OO
ெள	0BCC	13DC 0020 0002	TAMIL VOWEL SIGN AU
ெள	0BC6 0BD7	13DC 0020 0002	TAMIL VOWEL SIGN AU
்ள	0BD7	13DF 0020 0002	TAMIL AU LETTER MARK

9.2.3. Results

The collation elements were applied to sort a random set of strings of Tamil. The input and corresponding output is given in Table 9.6.

Table 9.6. Input and Corresponding Sorted Output for Tamil

Input		Output	
வெள	ஓமம்	அக்கறை	கேழல்
வெள	ஓஷ	அககுள்	கைராசி
வெளவு	ஐயர்	அககுள்	கைராட்டு

யக்தம்	ஐயர	ஆன்	கொத்து
யக்தி	எஃகு	இன்	கொத்து
யகம்	கூட்டம்	ஈசல்	கொத்தூ
முந்து	கெடு	உனற	கோரி
முந்தை	கெடுதி	ஊழ்	கோரி
பிரமன்	கேழ்	எஃகு	கோல்
பிரமை	கேழல்	எற்று	கௌ
நீக்கம்	கைராசி	ஏர்	கௌ
நீக்கல்	கைராட்டு	ஐயர்	செட்டி
நீங்கு	கொத்து	ஐயர	செட்டு
தூற்று	கொத்தூ	ஓமம்	செடி
தூறல்	கொத்து	ஓஷ	ஞாலம்
தூறு	கோரி	ஓளவ	ஞாழல்
ஞாலம்	கோரி	ஓளவை	தூற்று
ஞாழல்	கோல்	கஃசு	தூறல்
செட்டி	கௌ	கக்கம்	தூறு
செட்டு	கௌ	கக	நீக்கம்
செடி	ஓளவை	காந்தல்	நீக்கல்
கஃசு	ஓளவ	காந்தள்	நீங்கு
கக்கம்	ஏர்	கிங்கரன்	பிரமன்
கக	எற்று	கிங்கிணி	பிரமை
காந்தல்	ஊழ்	குலவு	முந்து
காந்தள்	உனற	குலாலன்	முந்தை
கிங்கரன்	ஈசல்	குலாவு	யக்தம்
கிங்கிணி	இன்	கூட்டடி	யக்தி
குலவு	ஆன்	கூட்டம்	யகம்
குலாலன்	அக்கறை	கெடு	வௌ
குலாவு	அககுள்	கெடுதி	வௌ
கூட்டடி	அககுள்	கேழ்	வௌவு

9.3. Conclusion

Tamil has single level of collation, like other Indic languages. All characters are sorted at primary level. The sorting process requires some text processing to decompose the characters and contract multiple characters onto single collation elements. However after the mapping, the collation algorithm discussed in the second chapter is applied in a regular manner for eventual collation.