Notes on the 156 Tamil “characters” captured in the dataset

Tamil, like most of the other Indic scripts, is defined as a ``syllabic alphabet" in that the unit of encoding is a syllable. In general, these syllabic units are the smallest units of isolated writing in the Indic scripts, and hence the nearest thing to isolated characters.

These syllabic units correspond to

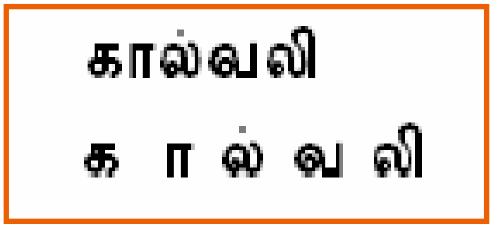
- isolated vowels (e.g., ‘u’, represented as உ)
- isolated consonants (with inherent neutral vowel a) (e.g, ‘ka’, represented as க)
- CV combinations where a consonant has been modified by a vowel and is indicated by a vowel diacritic (e.g. ‘ku’, represented as கு)
- Clusters of 2 or more consonants modified by vowels (CCV, CCCV, and so forth) (e.g. ‘kshū’, represented as க்ஷூ)

Most Indic scripts have the order of 600 CV units and as many as 20,000 CCV ones in theory, although only a much smaller subset (especially of CCV units) is used in practice. In general, the graphic units corresponding to these syllables show distinctive internal structure and a constituent set of graphemes. However The V diacritics and ligatures for C clusters are not standardized in some scripts. Sometimes the C clusters are represented as new opaque graphemes. In the online HWR scenario, the beginnings of most graphemes are marked by pen-lifts, but not always. In particular, certain V diacritics may be fused inseparably with the underlying C grapheme. Different V diacritics may be visually similar and differ only in how they attach to the C grapheme.

The present-day Tamil script is simpler than other Indic scripts because of the use of the lack of separate graphemes for voiced, voiceless and aspirated Cs and the use of vowel muting to unravel C clusters into linear sequences of C graphemes. In addition, some of the vowel diacritics are written in Tamil as distinct symbols to the left and/or right of the C grapheme.

This results in Tamil being written linearly as a sequence of visually discrete symbols, which we will refer to as characters, for lack of a better term. However as mentioned earlier, these characters do not have a consistent linguistic interpretation, nor is it claimed that they represent the only or best way to separate Tamil writing into a linear sequence of symbols.

The set of 156 “characters” represented in this collection includes in addition to independent V and C graphemes CV combinations where the vowel diacritics attach above or below the base C grapheme or are otherwise difficult to segment, and those vowel diacritics that occur as distinct characters to the left or right of the base C. The set also includes selected C cluster ligatures and their CV combinations, for a total of 156 “characters”. The adjoining figure shows a Tamil word split into “characters” by this definition of Tamil characters.



கால்வலி

க ா ல் வ லி

Like all Indic scripts, there is no tradition of writing Tamil in boxes; however informal observation of several native Tamil writers revealed that they could write “characters” as defined here, in boxes consistently with no or minimal training.

The entire set of 156 characters is shown in Table 1. Table 2 provides the sequences of Unicode characters corresponding to the 156 characters used in this dataset.

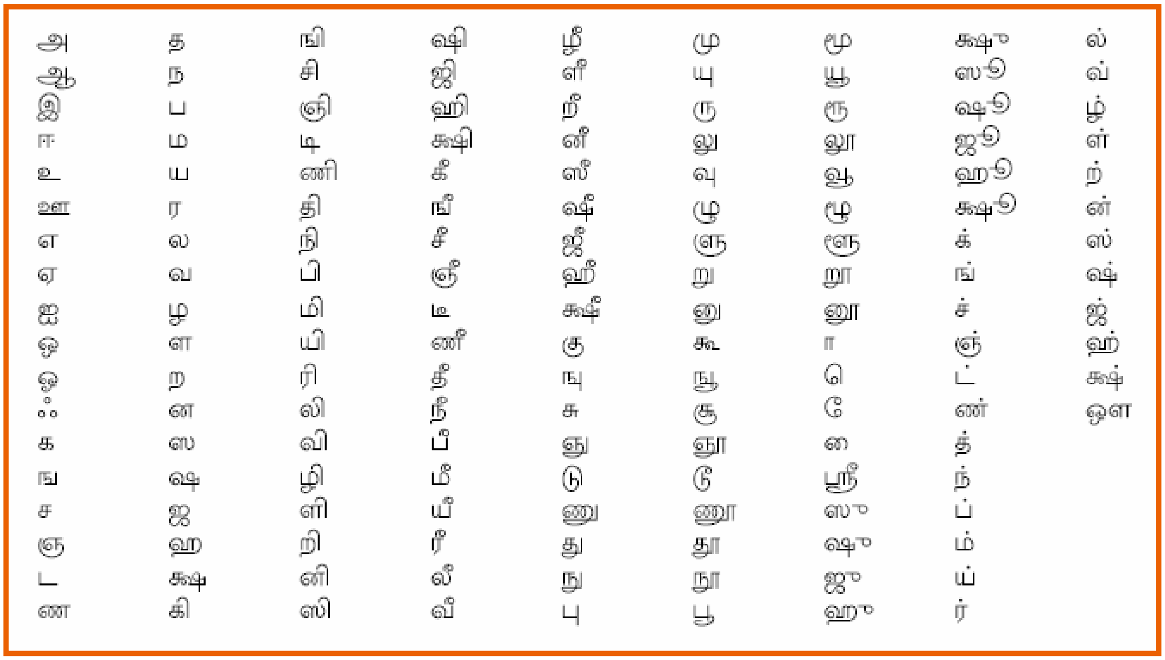
References

Experiences in Collection of Handwriting Data for Online Handwriting Recognition in Indic Scripts, Ajay S Bhaskarabhatla and Sriganesh Madhvanath, 4th Intl Conf. Linguistic Resources and Evaluation (LREC 2004), Lisbon, Portugal, May 26-28, 2004

Florian Coulmas, ``The Blackwell Encyclopedia of Writing Systems," First Ed. pp. 229-230, 1999

Omniglot – A Guide to Written Language. http://www.omniglot.com/writing/tamil.htm

Table 1: 156 characters in Tamil dataset



அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ ஓ ஃ க ங ச ஞ ட ண
த ந ப ம ய ர ல வ ழ ள ற ன ஸ ஷ ஜ ஹ க்ஷ கி
ஙி சி ஞி டி ணி தி நி பி மி யி ரி லி வி ழி ளி றி னி ஸி
ஷி ஜி ஹி க்ஷி கீ ஙீ சீ ஞீ டீ ணீ தீ நீ பீ மீ யீ ரீ லீ வீ
ழீ ளீ றீ னீ ஸீ ஷீ ஜீ ஹீ க்ஷீ கு ஙு சு ஞு டு ணு து நு பு
மு யு ரு லு வு ழு ளு று னு கூ ஙூ சூ ஞூ டூ ணூ தூ நூ பூ
மூ யூ ரூ லூ வூ ழூ ளூ றூ னூ ா ெ ே ை ஸ்ரீ ஸு ஷு ஜு ஹு
க்ஷு ஸூ ஷூ ஜூ ஹூ க்ஷூ க் ங் ச் ஞ் ட் ண் த் ந் ப் ம் ய் ர்
ல் வ் ழ் ள் ற் ன் ஸ் ஷ் ஜ் ஹ் க்ஷ் ஔ

Table 2: Unicode sequences corresponding to 156 characters

Class Id	Tamil Character	Unicode
0	அ	0B85
1	ஆ	0B86
2	இ	0B87
3	ஈ	0B88
4	உ	0B89
5	ஊ	0B8A
6	எ	0B8E
7	ஏ	0B8F
8	ஐ	0B90
9	ஒ	0B92
10	ஓ	0B93
11	◌ஃ	0B83

12	க	0B95
13	ங	0B99
14	ச	0B9A
15	ஞ	0B9E
16	ட	0B9F
17	ண	0BA3
18	த	0BA4
19	ந	0BA8
20	ப	0BAA
21	ம	0BAE
22	ய	0BAF
23	ர	0BB0
24	ல	0BB2
25	வ	0BB5
26	ழ	0BB4
27	ள	0BB3
28	ற	0BB1
29	ன	0BA9
30	ஸ	0BB8
31	ஷ	0BB7
32	ஜ	0B9C
33	ஹ	0BB9
34	க்ஷ	0B95 0BCD 0BB7
35	கி	0B95 0BBF
36	ஙி	0B99 0BBF
37	சி	0B9A 0BBF
38	ஞி	0B9E 0BBF
39	டி	0B9F 0BBF
40	ணி	0BA3 0BBF
41	தி	0BA4 0BBF
42	நி	0BA8 0BBF
43	பி	0BAA 0BBF
44	மி	0BAE 0BBF
45	யி	0BAF 0BBF
46	ரி	0BB0 0BBF
47	லி	0BB2 0BBF
48	வி	0BB5 0BBF

49	ழி	0BB4 0BBF
50	ளி	0BB3 0BBF
51	றி	0BB1 0BBF
52	னி	0BA9 0BBF
53	ஸி	0BB8 0BBF
54	ஷி	0BB7 0BBF
55	ஜி	0B9C 0BBF
56	ஹி	0BB9 0BBF
57	க்ஷி	0B95 0BCD 0BB7 0BBF
58	கீ	0B95 0BC0
59	ஙீ	0B99 0BC0
60	சீ	0B9A 0BC0
61	ஞீ	0B 9E 0BC0
62	டீ	0B9F 0BC0
63	ணீ	0BA3 0BC0
64	தீ	0BA4 0BC0
65	நீ	0BA8 0BC0
66	பீ	0BAA 0BC0
67	மீ	0BAE 0BC0
68	யீ	0BAF 0BC0
69	ரீ	0BB0 0BC0
70	லீ	0BB2 0BC0
71	வீ	0BB5 0BC0
72	ழீ	0BB4 0BC0
73	ளீ	0BB3 0BC0
74	றீ	0BB1 0BC0
75	னீ	0BA9 0BC0
76	ஸீ	0BB8 0BC0
77	ஷீ	0BB7 0BC0
78	ஜீ	0B9C 0BC0
79	ஹீ	0BB9 0BC0
80	க்ஷீ	0B95 0BCD 0BB7 0BC0
81	கு	0B95 0BC1
82	ஙு	0B99 0BC1
83	சு	0B9A 0BC1
84	ஞு	0B9E 0BC1
85	டு	0B9F 0BC1

86	ணு	0BA3 0BC1
87	து	0BA4 0BC1
88	நு	0BA8 0BC1
89	பு	0BAA 0BC1
90	மு	0BAE 0BC1
91	யு	0BAF 0BC1
92	ரு	0BB0 0BC1
93	லு	0BB2 0BC1
94	வு	0BB5 0BC1
95	ழு	0BB4 0BC1
96	ளு	0BB3 0BC1
97	று	0BB1 0BC1
98	னு	0BA9 0BC1
99	கூ	0B95 0BC2
100	ஙூ	0B99 0BC2
101	சூ	0B9A 0BC2
102	ஞூ	0B9E 0BC2
103	டூ	0B9F 0BC2
104	ணூ	0BA3 0BC2
105	தூ	0BA4 0BC2
106	நூ	0BA8 0BC2
107	பூ	0BAA 0BC2
108	மூ	0BAE 0BC2
109	யூ	0BAF 0BC2
110	ரூ	0BB0 0BC2
111	லூ	0BB2 0BC2
112	வூ	0BB5 0BC2
113	ழூ	0BB4 0BC2
114	ளூ	0BB3 0BC2
115	றூ	0BB1 0BC2
116	னூ	0BA9 0BC2
117	◌ா	0BBE
118	◌ெ	0BC6
119	◌ே	0BC7
120	◌ை	0BC8
121	ஸ்ரீ	0BB8 0BCD 0BB0 0BC0
122	ஸு	0BB8 0BC1

123	ஷு	0BB7 0BC1
124	ஜு	0B9C 0BC1
125	ஹு	0BB9 0BC1
126	க்ஷு	0B95 0BCD 0BB7 0BC1
127	ஸூ	0BB8 0BC2
128	ஷூ	0BB7 0BC2
129	ஜூ	0B9C0BC2
130	ஹூ	0BB9 0BC2
131	க்ஷூ	0B95 0BCD 0BB7 0BC2
132	க்	0B95 0BCD
133	ங்	0B99 0BCD
134	ச்	0B9A 0BCD
135	ஞ்	0B9E 0BCD
136	ட்	0B9F 0BCD
137	ண்	0BA3 0BCD
138	த்	0BA4 0BCD
139	ந்	0BA8 0BCD
140	ப்	0BAA 0BCD
141	ம்	0BAE 0BCD
142	ய்	0BAF 0BCD
143	ர்	0BB0 0BCD
144	ல்	0BB2 0BCD
145	வ்	0BB5 0BCD
146	ழ்	0BB4 0BCD
147	ள்	0BB3 0BCD
148	ற்	0BB1 0BCD
149	ன்	0BA9 0BCD
150	ஸ்	0BB8 0BCD
151	ஷ்	0BB7 0BCD
152	ஜ்	0B9C 0BCD
153	ஹ்	0BB9 0BCD
154	க்ஷ்	0B95 0BCD 0BB7 0BCD
155	ஔ	0B94