Notes on the 166 Telugu symbols captured in the dataset

Telugu, 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 ఉ)
- Isolated consonants (with inherent neutral vowel a) (e.g, 'ka', represented as క)
- CV combinations where a consonant has been modified by a vowel and is indicated by a vowel diacritic (e.g. 'k$\bar{a}$', represented as కా)
- Clusters of 2 or more consonants modified by vowels (CCV, CCCV, and so forth) (e.g. 'kro', represented as క్రొ)

Telugu script has 18 vowels and 36 consonants of which thirteen vowels and thirty five consonants are in common usage. Of all the Indic scripts, the Telugu script has the largest number of vowels and consonants. Theoretically, the number of syllabic units is $O(10^4)$ though a much smaller subset is used in practice. From recognition point of view, considering each syllabic unit as a pattern class is practically impossible. The approach used here is to identify a smaller subset of basic units sufficient to cover the entire set of syllabic units. These basic units need not always be the constituent graphemes of the syllabic units because of the structural complexities. For example, certain vowel 'maatra' can be fused inseparably with the underlying consonant as in the syllable 'ని' in Telugu. In such a case, recognition based on considering the consonant and vowel as basic units would have to deal with segmentation problems. In practice, the basic units are determined by various factors like ease of segmentation of characters into these units and not just by linguistic criteria.

Telugu script is generally non-cursive in style and hence pen-up usually separates the basic graphemes though not always. So, the basic graphemes of the script i.e. independent vowels, consonants, vowel diacritics and consonant modifiers are included in the symbol set. Also included are some consonant-vowel units which cannot be easily segmented. In addition, the symbol set also contains some symbols which do not have linguistic interpretation but have stable pattern across writers and help reduce the total number of symbols to be collected.

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

The complete symbol set contains a total of 150 unique symbols, shown in Table 1. Some of the maatras are collected more than once at different possible relative positions w.r.t. the underlying consonant. Including these, the total number of symbols is 166. The sequences of Unicode characters for the entire set of symbols are provided in Table 2.

Table 1: The 150 unique (out of the total 166) symbols in the Telugu Dataset

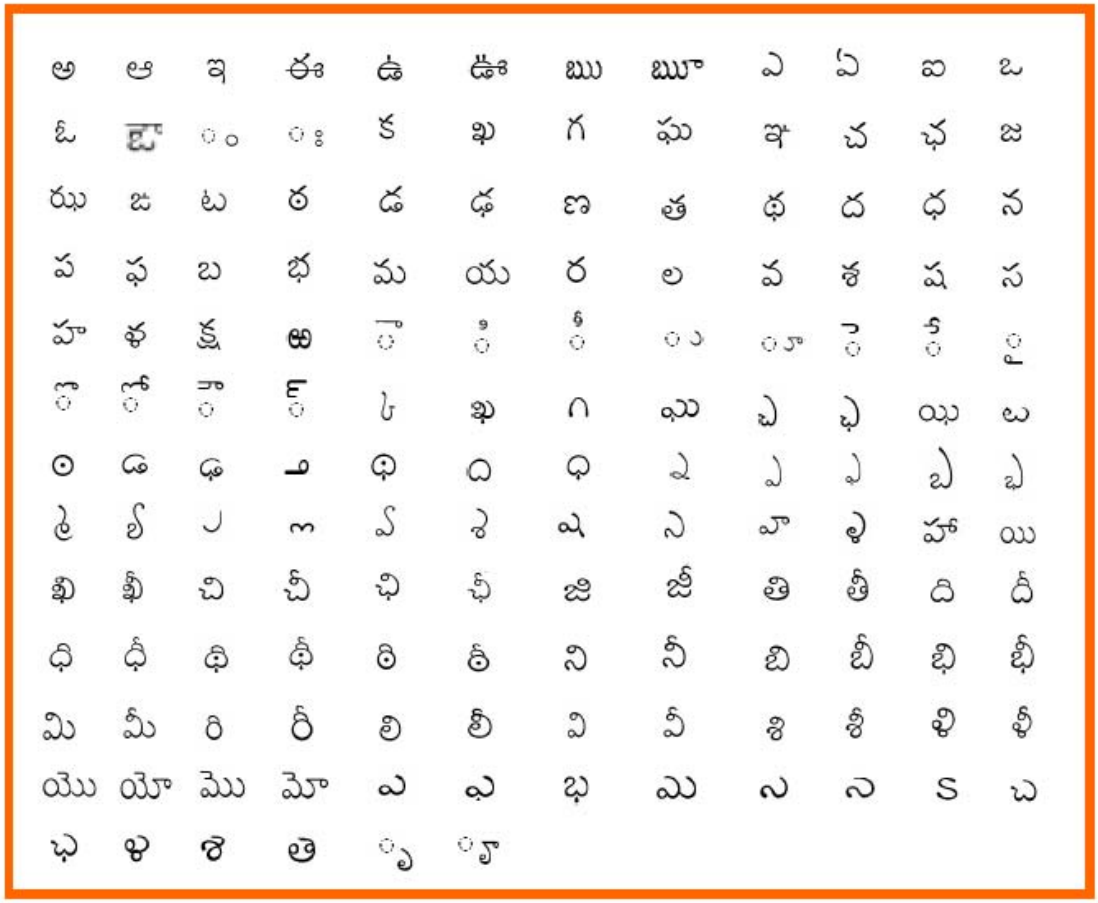

Table 1: Unicode sequences corresponding to 166 symbols

Class Id	Telugu Symbol	Unicode
0	అ	0C05
1	ఆ	0C06
2	ఇ	0C07
3	ఈ	0C08
4	ఉ	0C09
5	ఊ	0C0A
6	ఋ	0C0B
7	ౠ	0C60
8	ఎ	0C0E
9	ఏ	0C0F
10	ఐ	0C10
11	ఒ	0C12
12	ఓ	0C13
13	ఔ	0C14
14	◌ం	0C02
15	◌ః	0C03
16	క	0C15
17	ఖ	0C16
18	గ	0C17
19	ఘ	0C18
20	ఞ	0C1E
21	చ	0C1A
22	ఛ	0C1B
23	జ	0C1C
24	ఝ	0C1D
25	ఙ	0C19
26	ట	0C1F
27	ఠ	0C20
28	డ	0C21
29	ఢ	0C22
30	ణ	0C23
31	త	0C24
32	థ	0C25
33	ద	0C26
34	ధ	0C27
35	న	0C28

36	ప	0C2A
37	ఫ	0C2B
38	బ	0C2C
39	భ	0C2D
40	మ	0C2E
41	య	0C2F
42	ర	0C30
43	ల	0C32
44	వ	0C35
45	శ	0C36
46	ష	0C37
47	స	0C38
48	హ	0C39
49	ళ	0C33
50	క్ష	0C15 0C37
51	ఱ	0C31
52	◌ా	0C3E
53	◌ి	0C3F
54	◌ీ	0C40
55	◌ు	0C41

56	◌ూ	0C42
57	◌ృ	0C46
58	◌ౄ	0C47
59	◌ె	…..
60	◌ై	0C4A
61	◌ొ	0C4B
62	◌ో	0C4C
63	ఌ	…..
64	ఁ	…..
65	ం	…..
66	ఄ	…..
67	ఽ	…..
68	ఀ	…..
69	ఴ	…..
70	ఀ	…..
71	ఀ	…..
72	ఀ	…..
73	ఀ	…..
74	ఀ	…..
75	ఀ	…..

76	ఁ	…..
77	ఀ	…..
78	ఄ	…..
79	ఌ	…..
80	఍	…..
81	ఎ	…..
82	ఏ	…..
83	ఐ	…..
84	఑	…..
85	ఒ	…..
86	ఓ	…..
87	ఔ	…..
88	ఘ	…..
89	ఙ	…..
90	ఛ	…..
91	ఝ	…..
92	ఞ	…..
93	హా	0C39 0C3E
94	యి	0C2F 0C07
95	వి	0C35 0C3F

96	వీ	0C35 0C40
97	జి	0C1C 0C3F
98	జీ	0C1C 0C40
99	ఖి	0C16 0C3F
100	ఖీ	0C16 0C40
101	భి	0C2D 0C3F
102	భీ	0C2D 0C40
103	మి	0C2E 0C3F
104	మీ	0C2E 0C40
105	లి	0C32 0C3F
106	లీ	0C32 0C40
107	శి	0C36 0C3F
108	శీ	0C36 0C40
109	తి	0C24 0C3F
110	తీ	0C24 0C40
111	చి	0C1A 0C3F
112	చీ	0C1A 0C40
113	ఛి	0C1B 0C3F
114	ఛీ	0C1B 0C40
115	ని	0C28 0C3F

116	నీ	0C28 0C40
117	ళి	0C33 0C3F
118	ళీ	0C33 0C40
119	ఠి	0C20 0C3F
120	ఠీ	0C20 0C40
121	రి	0C30 0C3F
122	రీ	0C30 0C40
123	ది	0C26 0C3F
124	దీ	0C26 0C40
125	ధి	0C27 0C3F
126	ధీ	0C27 0C40
127	థి	0C25 0C3F
128	థీ	0C25 0C40
129	బి	0C2C 0C3F
130	బీ	0C2C 0C40
131	యొ	0C2F 0C4A
132	యో	0C2F 0C4B
133	◌ె	0C43
134	◌ే	0C44
135	◌్	0C4D

136	మొ	0C2E 0C4A
137	మో	0C2E 0C4B
138	ఎ	…..
139	ఫ	…..
140	భ	…..
141	ఱ	…..
142	ఌ	…..
143	ఌ	…..
144	◌ా(టా)	0C3E
145	◌ా(పా)	0C3E
146	◌ి(టి)	0C3F
147	◌ీ(టీ)	0C40
148	◌ు (పు)	0C41
149	◌ూ(పూ)	0C42
150	◌ు (జు)	0C41
151	◌ూ(జూ)	0C42
152	◌ొ(బొ)	0C4A
153	◌ొ(పొ)	0C4A
154	◌ో(బో)	0C4B
155	◌ో(పో)	0C4B

156	ె(టె)	0C46
157	ే(టే)	0C47
158	ొ (టొ)	0C4C
159	ో (పో)	0C4C
160	ఽ	…..
161	ఌ	…..
162	ఌ	…..
163	ఴ	…..
164	ఀ	…..
165	ఀ	…..