TOWARDS UNICODE STANDARD FOR URDU - WG2 N2413-1/SC2 N35891

Dr. Khaver ZIA
Director
Beaconhouse Informatics Computer Institute
Lahore. Pakistan
E-mail: kzia@informatics.edu.pk

ABSTRACT

This paper is an update on the progress made in standardization of Urdu in Pakistan. The compatibility of Standard character Set of Urdu with Unicode is analyzed. Inclusion of 25 Urdu Characters and ligatures in the Unicode standard is proposed.

KEYWORDS

Multilingual Processing, Standardization, Unicode, Urdu

1. INTRODUCTION

Urdu language and its characteristics have been discussed in detail in earlier papers [1] [2]. The code table of Urdu referred to in these papers was approved by the Government of Pakistan in August 2000. In the current paper an analysis is done with a view to make the Urdu character set compatible with Unicode.

2. ANALYSIS OF URDU CHARACTER CODES

The Unicode standard which is fully compatible with ISO/IEC 10646 specification encodes characters in a 16-bit code. This enables 65,535 unique characters to be encoded. The advantages of Unicode include uniform character width and ability to include all national standards. [3] [4].

On going through the encoding of characters in Unicode, it is found that Arabic and its associated languages have been allocated 1,200 code points. These code points range from 0600h to 06FFh (256 code points) and then from FB50h to FEFFh (944 code points). These code points comprise basic characters of the Arabic family of languages along with innumerable glyphs and ligatures.

An exercise was done to identify the Urdu characters in the Arabic block and draw up a table of comparison. The result is given in Table 1. After the exercise was completed it was found that 25 characters do not have a

representation in Unicode. These have been listed in Table 2. Each character is given a proposed description and a symbol, where applicable. If these “missing characters” are given a place in Unicode standard, it would make Urdu compatible with Unicode and ISO/IEC 10646.

It should be noted that Unicode does not specify the collating sequence. In case of Urdu too, the collating sequence is defined through software. Unicode can serve as a source table for all the character and ligatures of Urdu, as it does for other languages of the world.

3. CONCLUSION

ISO/IEC 10646 /Unicode is fast assuming a standard for representing national character codes. After analysis of Urdu character codes with Unicode standard, a table of missing Urdu characters is drawn up. It is proposed that these characters be included in the Unicode standard.

4. REFERENCES

1. ZIA, Khaver (1999),“Standard Code Table for Urdu”. *4th Symposium on Multilingual Information Processing (MLIT-4)*. Yangon. Myanmar. Organized by CICC Japan. October.

2. ZIA, Khaver (1999), “A Survey of Standardization in Urdu.” *4th Symposium on Multilingual Information Processing (MLIT-4)*. Yangon. Myanmar. Organized by CICC Japan. October.

3. LUA Kim Teng (1989), “Standardization for Multilingual Computing”. Keynote Address. *Proc. of 3rd AFSIT Symposium held at Singapore*. Organized by CICC. Japan. December.

4. SHIBANO Koji (1993), “ISO/IEC 10646-1 in Japan”. Technical Report. *Proc. of 7th AFSIT held in Tokyo. Japan.* Organized by CICC Japan. October.

5. ACKNOWLEDGEMENTS

The author thanks the management of Beaconhouse Informatics Pakistan, for its support in the preparation of this paper. The author gratefully acknowledges the provision of scanned bit-images of Urdu characters and ligatures by Mr. Humayun Qureshi, formerly of IBM, Pakistan.

TABLE 1

Standard Urdu Codes mapped to ISO/IEC 10646 /Unicode

Serial No.	Code Point (hex)	Symbol	Unicode	Unicode Description (where applicable) or Proposed Description
1-32	00-1F			CONTROL AREA (Lower Block)
33	20		0020	SPACE
34	21	!	0021	EXCLAMATION MARK
35	22	"	0022	QUOTATION MARK
36	23	#	0023	NUMBER SIGN
37	24	Cr	00A4	CURRENCY SIGN
38	25	%	0025	PERCENTAGE SIGN
39	26	&	0026	AMPERSAND
40	27	ٔ		ARABIC-URDU INVERTED PESH SIGN ***Urdu***
41	28	(	0028	LEFT PARENTHESIS
42	29	)	0029	RIGHT PARENTHESIS
43	2A	*	002A	ASTERISK
44	2B	+	002B	PLUS SIGN
45	2C	،	060C	ARABIC COMMA
46	2D	-	002D	HYPHEN-MINUS
47	2E	٫		ARABIC-URDU DECIMAL SIGN ***Urdu***
48	2F	÷	00F7	DIVISION SIGN

Serial No.	Code Point (hex)	Symbol	Unicode	Unicode Description (where applicable) or Proposed Description
49	30	۰	06F0	EASTERN ARABIC-INDIC DIGIT ZERO
50	31	۱	06F1	EASTERN ARABIC-INDIC DIGIT ONE
51	32	۲	06F2	EASTERN ARABIC-INDIC DIGIT TWO
52	33	۳	06F3	EASTERN ARABIC-INDIC DIGIT THREE
53	34	۴	06F4	EASTERN ARABIC-INDIC DIGIT FOUR
54	35	۵	06F5	EASTERN ARABIC-INDIC DIGIT FIVE
55	36	۶	06F6	EASTERN ARABIC-INDIC DIGIT SIX
56	37	۷	06F7	EASTERN ARABIC-INDIC DIGIT SEVEN
57	38	۸	06F8	EASTERN ARABIC-INDIC DIGIT EIGHT
58	39	۹	06F9	EASTERN ARABIC-INDIC DIGIT NINE
59	3A	؛		ARABIC-URDU COLON SIGN ***Urdu***
60	3B	؛	061B	ARABIC SEMI-COLON
61	3C	<	003C	LESS-THAN SIGN
62	3D	=	003D	EQUALS SIGN
63	3E	>	003E	GREATER-THAN SIGN
64	3F	؟	061F	ARABIC QUESTION MARK
65	40	@	0040	COMMERCIAL AT
66	41			ARABIC-URDU HARD SPACE ***Urdu***
67	42	ـٔ		ARABIC-URDU HAMZA E IZAFAT ***Urdu***
68	43	ـٖ		ARABIC-URDU KASRA E IZAFAT ***Urdu***

Serial No.	Code Point (hex)	Symbol	Unicode	Unicode Description (where applicable) or Proposed Description
69	44		0670	ARABIC ALEF ABOVE
70	45			ARABIC-URDU ALEF BELOW ***Urdu***
71	46			ARABIC-URDU PESH ABOVE ***Urdu***
72	47			ARABIC-URDU SPECIAL INVERTED PESH ***Urdu***
73	48			ARABIC-URDU ZARE BELOW ***Urdu***
74	49		064B	ARABIC SPACING FATHATAN
75	4A		064D	ARABIC SPACING KASRATAN
76	4B		064C	ARABIC SPACING DAMMATAN
77	4C			ARABIC-URDU SMALL TAH ***Urdu***
78	4D			ARABIC-URDU SAKOON ***Urdu***
79	4E			ARABIC-URDU REVERSE SAKOON ***Urdu***
80	4F		0651	ARABIC SHADDAH
81	50	ا	0627	ARABIC LETTER ALEF
82	51	أ	0623	ARABIC LETTER HAMZAH ON ALEF
83	52	آ	0622	ARABIC LETTER MADDAH ON ALEF
84	53	ب	0628	ARABIC LETTER BAA
85	54	پ	067E	ARABIC LETTER TAA WITH THREE DOTS BELOW = peh
86	55	ت	062A	ARABIC LETTER TAA
87	56	ٹ	0679	ARABIC LETTER TAA WITH SMALL TAH
88	57	ث	062B	ARABIC LETTER THAA

Serial No.	Code Point (hex)	Symbol	Unicode	Unicode Description (where applicable) or Proposed Description
89	58	ج	062C	ARABIC LETTER JEEM
90	59	چ	0686	ARABIC LETTER HAA WITH MIDDLE THREE DOTS DOWNWARD = tcheh
91	5A	ح	062D	ARABIC LETTER HAA
92	5B	خ	062E	ARABIC LETTER KHAA
93	5C	د	062F	ARABIC LETTER DAL
94	5D	ڈ	0688	ARABIC LETTER DAL WITH SMALL TAH
95	5E	ذ	0630	ARABIC LETTER THAL
96	5F	ر	0631	ARABIC LETTER RA
97	60	ڑ	0691	ARABIC LETTER RA WITH SMALL TAH
98	61	ز	0632	ARABIC LETTER ZAIN
99	62	ژ	0698	ARABIC LETTER RA WITH THREE DOTS ABOVE = jeh
100	63	س	0633	ARABIC LETTER SEEN
101	64	ش	0634	ARABIC LETTER SHEEN
102	65	ص	0635	ARABIC LETTER SAD
103	66	ض	0636	ARABIC LETTER DAD
104	67	ط	0637	ARABIC LETTER TAH
105	68	ظ	0638	ARABIC LETTER DHAH
106	69	ع	0639	ARABIC LETTER AIN
107	6A	غ	063A	ARABIC LETTER GHAIN
108	6B	ف	0641	ARABIC LETTER FA

Serial No.	Code Point (hex)	Symbol	Unicode	Unicode Description (where applicable) or Proposed Description
109	6C	ق	0642	ARABIC LETTER QAF
110	6D	ک	06A9	ARABIC LETTER OPEN CAF
111	6E	گ	06AF	ARABIC LETTER GAF
112	6F	ل	0644	ARABIC LETTER LAM
113	70	م	0645	ARABIC LETTER MEEM
114	71	ں	06BA	ARABIC LETTER DOTLESS NOON
115	72	ن	0646	ARABIC LETTER NOON
116	73	و	0648	ARABIC LETTER WAW
117	74	ؤ	0624	ARABIC LETTER HAMZAH ON WAW
118	75	ہ	0647	ARABIC LETTER HA
119	76	ة	0629	ARABIC LETTER TAA MARBUTAH
120	77	ء	0621	ARABIC LETTER HAMZAH
121	78	ی	0649	ARABIC LETTER ALEF MAQSURAH
122	79	ے	06D2	ARABIC LETTER YA BARREE
123	7A	ھ	06BE	ARABIC LETTER KNOTTED HA
124	7B			ARABIC-URDU NO-DICRITIC SIGN ***Urdu***
125	7C	ـَ	064E	ARABIC FATHAH
126	7D	ـِ	0650	ARABIC KASRAH
127	7E	ـُ	064F	ARABIC DAMMAH
128	7F			NOT USED

Serial No.	Code Point (hex)	Symbol	Unicode	Unicode Description (where applicable) or Proposed Description
129-160	80-9F			CONTROL AREA (Upper Block)
161	A0	ﷲ	FDF2	ARABIC LIGATURE ALLAH ISOLATED FORM
162	A1	ﷻ	FDFB	ARABIC LIGATURE JALLA JALALOUHOU
163	A2			ARABIC-URDU LIGATURE BISMILLAH ***Urdu***
164	A3		FDFA	ARABIC LIGATURE SALLALLAHOU ALAYHE WASALLAM
165	A4	ﷺ	FDF9	ARABIC LIGATURE SALLA ISOLATED FORM
166	A5	ؑ		ARABIC-URDU LIGATURE ALAYHE AS SALAM ***Urdu***
167	A6	ؓ		ARABIC-URDU LIGATURE RADIALLAH ***Urdu***
168	A7	ؒ		ARABIC-URDU LIGATURE REHMATULLAH ***Urdu***
169	A8	ـــــ		ARABIC-URDU TAKHALLUS SIGN (Poetry) ***Urdu***
170	A9	؏		ARABIC-URDU MISRA SIGN (Poetry) ***Urdu***
171	AA	ـــــ		ARABIC-URDU FOOTNOTE SIGN ***Urdu***
172	AB	ـــــ		ARABIC-URDU SAFAH SIGN ***Urdu***
173	AC	ـــــ		ARABIC-URDU NUMBER SIGN ***Urdu***
174	AD	ـــــ		ARABIC-URDU SANAH SIGN ***Urdu***
175	AE	ـــــ		ARABIC-URDU LONG MADD ***Urdu***
176	AF	ﻻ	FEFB	ARABIC LAAM ALEF ISOLATED
177	B0	۝		ARABIC-URDU END OF SECTION SIGN ***Urdu***
178-192	B1-BF			RESERVED AREA

Serial No.	Code Point (hex)	Symbol	Unicode	Unicode Description (where applicable) or Proposed Description
193	C0	[	005B	LEFT SQUARE BRACKET
194	C1	\	005C	REVERSE SOLIDUS (BACKSLASH)
195	C2	]	005D	RIGHT SQUARE BRACKET
196	C3	_	005F	LOW LINE (UNDERSCORE)
197	C4	{	007B	LEFT CURLY BRACKET
198	C5	:	003A	COLON
199	C6	}	007D	RIGHT CURLY BRACKET
200	C7	ـ	06D4	ARABIC PERIOD (DASH)
201-208	C8-CF			RESERVED AREA
209-254	D0- FD			VENDOR AREA
255	FE			LANGUAGE TOGGLE
256	FF			NOT USED

TABLE 2

Characters and Ligatures from Standard Urdu Code Page proposed for inclusion in ISO/IEC 10646 / Unicode

Serial No.	Code Point (hex)	Symbol	Unicode	Proposed Description
1	2E	۔		ARABIC-URDU DECIMAL SIGN ***Urdu***
2	3A	؛		ARABIC-URDU COLON SIGN ***Urdu***
3	41			ARABIC-URDU HARD SPACE ***Urdu***
4	42	ـٔـ		ARABIC-URDU HAMZA E IZAFAT ***Urdu***
5	43	ـٖـ		ARABIC-URDU KASRA E IZAFAT ***Urdu***
6	45	ـٙـ		ARABIC-URDU ALEF BELOW ***Urdu***
I7	46	ـٗـ		ARABIC-URDU PESH ABOVE ***Urdu***
8	47	ـٚـ		ARABIC-URDU SPECIAL INVERTED PESH ***Urdu***
9	48	ـٛـ		ARABIC-URDU ZARE BELOW ***Urdu***
10	4C	ـٜـ		ARABIC-URDU SMALL TAH ***Urdu***
11	4D	ـٝـ		ARABIC-URDU SAKOON ***Urdu***
12	4E	ـٞـ		ARABIC-URDU REVERSE SAKOON ***Urdu***
13	7B			ARABIC-URDU NO-DICRITIC SIGN ***Urdu***
14	A2	بسم الله الرحمن الرحيم		ARABIC-URDU LIGATURE BISMILLAH ***Urdu***
15	A5	ؑ		ARABIC-URDU LIGATURE ALAYHE AS SALAM ***Urdu***
16	A6	ؓ		ARABIC-URDU LIGATURE RADIALLAH ***Urdu***

Serial No.	Code Point (hex)	Symbol	Unicode	Proposed Description
17	A7			ARABIC-URDU LIGATURE REHMATULLAH ***Urdu***
18	A8			ARABIC-URDU TAKHALLUS SIGN (Poetry) ***Urdu***
19	A9			ARABIC-URDU MISRA SIGN (Poetry) ***Urdu***
20	AA			ARABIC-URDU FOOTNOTE SIGN ***Urdu***
21	AB			ARABIC-URDU SAFAH SIGN ***Urdu***
22	AC			ARABIC-URDU NUMBER SIGN ***Urdu***
23	AD			ARABIC-URDU SANAH SIGN ***Urdu***
24	AE			ARABIC-URDU LONG MADD ***Urdu***
25	B0	০		ARABIC-URDU END OF SECTION SIGN ***Urdu***