8.3 The Perso-Arabic Standard for Information Interchange

The standard proposed by C-DAC GIST is an extension to the standard 8-bit ASCII. It compliments the symbol set of Latin script by adding the symbol of the Perso-Arabic scripts. The standard supports storage for the Perso-Arabic languages like Urdu, Persian, Sindhi, Kashmiri, and Arabic.

Characteristics

i. Its a 8-bit standard

ii. Supports letters for Urdu, Arabic, Sindhi, Kashmiri

iii. Defines Perso-Arabic alphabets in the upper ASCII (This leaves the lower ASCII free. The lower ASCII can be used for English alphabets e.g. to give a bi-lingual font support).

iv. Defines numerals other than ASCII numbers (48 to 57) (This may help supporting both Arabic Numerals 0-9 and language specific numerals)

v. Maintains the order of alphabets for Perso-Arabic languages.

vi. Alphabets / letters are placed in their ascending order. Letters like “bhey” are not provided for URDU but kept for languages like Sindhi. Urdu may make use of the digraph “be” and “choTi-he” for that.

vii. Minimal erabs are provided. Tanveen, for example do-zabar, can be formed with the help of double zabar.

vii. Unicode compatability can be achieved by having PASCII to UNICODE & viceversa converter.

Superscripts

i. Place for superscripts like khaRa-alif is provided

ii. Place for superscripts for Arabic is provided

iii. Place for superscripts like “re-ze”, “ain”, etc. is provided.

iv. Numerals are placed after erabs and superscripts. (This is provided only to support display for language specific numerals and standard numerals i.e. the ASCII numerals are available).

Standardization of Perso-Arabic Fonts

Characteristics of Perso-Arabic languages :

Perso-Arabic languages are written in Naskh & Nastaliq scripts. Urdu & Kashmiri are traditionally written in the Nastaliq script ; while Sindhi is written in Naskh script. Although the script employs basic letters of the language, the rendering of these letters in a word is extremely complex. The reason for this complexity is that the text has traditionally been composed through calligraphy, a medium whose precepts are based on the aesthetic sense of the calligrapher rather than on any formula. So great is the variation in calligraphy that many times it is difficult to recognize the letters in a constituent word. This is because, in their calligraphed form, the individual letters partially or completely fused into each other thereby losing their identity. A degree of fusion is purposely introduced to make the resulting fused glyph visually appealing.

Another characteristic of the Perso-Arabic languages is the use of diacritics. Diacritics, although sparingly used, help in the proper pronunciation of the constituent word. The diacritics appear above or below a character to specify a vowel or emphasize a particular sound. These are essential for the removal of ambiguities, natural language processing and speech synthesis.

Standardization of Glyph Set

Following was taken into consideration while designing fonts for the Perso-Arabic languages. Considering the complexities of the script it was not possible to accommodate all the glyphs / ligatures in an 8 bit code space. Hence 16 – bit font code space was considered.

1. Alphabet

2. Numerals
3. Special characters
4. Diacritics
5. Religious and linguistic symbols
6. Control characters

The 16-bit *Nastaliq* font for Urdu & Kashmiri

Fonts developed by C-DAC for Urdu & Kashmiri are 16-bit. The Glyphs are defined in the User Area of the Unicode range. The ASCII range is not used and can be used for different purposes (it can be used to support English for example).

- Includes all the basic shapes
- Includes all the starting shapes and variations
- Includes all the middle shapes and variations
- Includes all the ending shapes and variations
- Includes levels for erabs (short vowels)
- Includes Complete ligatures
- Includes Beginning ligatures
- Includes Middle ligatures
- Includes Ending ligatures
- Includes dotted circle glyph

The 16-bit *Naskh* font for Sindhi, Urdu & Kashmiri.

Font developed by C-DAC for Sindhi, Urdu & Kashmiri are 16-bit. The Glyphs are defined in the User Area of the Unicode range. The ASCII range is not used and can be used for different purposes (it can be used to support English for example).

- Includes all the basic shapes
- Includes all the starting shapes
- Includes all the middle shapes
- Includes all the ending shapes
- Includes levels for erabs (short vowels)
- Includes Complete ligatures
- Includes Beginning ligatures
- Includes Middle ligatures
- Includes Ending ligatures.
- Includes dotted circle glyph

India

India is a paradise in the foot of the great Himalayas in the northern end and lies cocooned by huge oceans on the other three sides. While the Arabian Sea borders the southwest side, the southeast is lulled by the Bay of Bengal, and the southern tip - Kanya Kumari (Cape Comorin) is washed by the Indian Ocean. Hence protected by such natural barriers like mountains and water, it is separated from the rest of Asia. For geographers, it lies to the north of the equator between 8.4 and 37.6 degrees north latitude and 68.7 and 97.25 degrees east longitude. India measures 3214 km from north to south and 2933 kms from east to west. it has a land frontier of 15,200 kms and a coastline of 7516.5 kms.

India shares its political borders with Pakistan and Afghanistan on the west; Bangladesh and Myanmar in the east; Nepal, China, Tibet and Bhutan in the north.The Capital of India is New Delhi.

Languages

India has 18 officially recognized languages among about 200 languages as enumerated in the census.

Names of Languages

Following languages are listed in the 8th schedule of the Constitution (given in Devanagri order):

- Assamese
- Urdu
- Oriya
- Kannada
- Kashmiri
- Konkani
- Gujarati
- Tamil

- Telugu
- Nepali
- Punjabi
- Bengali
- Manipuri
- Marathi
- Malayalam
- Sanskrit
- Sindhi
- Hindi

Urdu Design Guide : General Information

Introduction

This document provides general information about the Urdu language and some conventions of its usage in India.

The information presented in this document is intended to assist in understanding the nature and problems of Urdu implementation in the digital medium. It contains the generic description of Urdu.

Urdu is one of the official languages of India. It is the official language of Pakistan, and spoken in various countries around the world.

Language Description

Urdu belongs to the Indo-Aryan subgroup of the Indo-European family of languages. It has developed with the heavy influences of Arabic, Persian and Turkish languages. Urdu writing system is a super set of Arabic and Persian and contains 39 characters. Urdu is written from right side to left. Unlike English, the characters do not have upper and lower cases. Further, the shape assumed by a character in a word is context-sensitive i.e. the shape is different depending whether the position of the character is at the beginning, in the middle or at the end of the constituent word.

Urdu is traditionally written in Nastaliq, a script rich in calligraphic content. Owing to the complexities of rendering, a number of alternate shapes are possible for a single letter, considering its position in the word and the letter next to it. Due to this nature of Nastaliq, it increases the glyph set for the language.

The characters of Urdu also need diacritics to help in a proper pronunciation of the constituent word. There are a number of diacritics, the common ones being Zabar, Zer, and Pesh.

History of Urdu language

The word Urdu means 'Lashkar', derived from the Turkish language meaning 'armies'. In the south of India it flourished under the name of Dakhani and southwest as Gurjari while in Delhi its name changed from Hindi to Hindavi and Hindustani. Alternate names of Urdu are DAKHINI(DAKANI, DECCAN, DESIA, MIRGAN), PINJARI, REKHTA (REKHTI).

Population using the Urdu Language

48,062,000 in India (1997 IMA);

10,719,000 in Pakistan (1993), or 7.57% of the population;

600,000 in Bangladesh;

64,000 in Mauritius (1993 Johnstone).

170,000 in South Africa (1987).

18,500 in Bahrain (1979 WA);

17,800 in Oman (1980 WA);

15,400 in Qatar;

382,000 in Saudi Arabia;

3,562 in Fiji (1980 WA);

23,000 in Germany;

14,000 in Norway;

Totals :

60,290,000 or more in all countries

104,000,000 including second language users (1999 WA).

PASCII (Perso-Arabic Standard for Information Interchange) Version 1.0

	128	144	160	176	192	208	224	240
	8	9	A	B	C	D	E	F
0		ج	ژ	ل	ُ	ض	۴	-
1	Kasheeda	چ	س	م	ٔ	ح	۵	/
2	ا	چ	ش	ن	ٔ	ع	۶	؛
3	آ	چ	ص	ں	ٔ	ق	۷	:
4	ب	ح	ض	ط	ٔ	ک	۸	؟
5	ب	خ	ط	و	ٔ	ل	۹	=
6	پ	د	ظ	ۆ	ّ	ج	!	۔
7	پ	ذ	ع	ه	ٔ	م	”	○
8	ق	د/ڈ	غ	ھ/ھ	ٔ	ـــنـــ	“	●
9	ت	ڈ	ف	ء	ٔ	%	’	،
A	ة	د	ق	ي/ى	ا	/	‘	Reserved
B	ت	ذ	ک	ي	ٔ	ء	(	Reserved
C	ت/ٹ	ر	ک	ے	...	۰	)	◌
D	ت	ڑ/ڑ	گ	ے	ٔ	ا	*	۔
E	ث	ڑھ	گ	َ	ٔ	۲	+	Reserved
F	ج	ز	گ	َ	ص	۳	ATR	Reserved

Code Chart Details of Pascii Storage Standard

Code Point	Character	Description
129	*Kasheeda*	Kasheeda Indicator (used to stretch character)
130	ا	LETTER ALIF Urdu, Sindhi, Kashmiri
131	آ	LETTER ALIF WITH MADD Urdu, Sindhi, Kashmiri
132	ب	LETTER BE Urdu, Sindhi, Kashmiri
133	ٻ	LETTER BBE Sindhi
134	ڀ	LETTER BHE Sindhi
135	پ	LETTER PE Urdu, Sindhi, Kashmiri
136	ڦ	LETTER PHE Sindhi
137	ت	LETTER TE Urdu, Sindhi, Kashmiri.
138	ة	LETTER TE MARBUTA Urdu.
139	ٽ	LETTER THE Sindhi
140	ټ/ٹ	LETTER TEY Urdu/Sindhi
141	ٺ	LETTER TTE Sindhi
142	ث	LETTER SE Urdu, Sindhi
143	ج	LETTER JEEM Urdu, Sindhi, Kashmiri
144	ڃ	LETTER JJE Sindhi
145	ڄ	LETTER JNE Sindhi
146	چ	LETTER CHE Urdu, Sindhi, Kashmiri
147	ڇ	LETTER CHHE Sindhi
148	ح	LETTER HAY Urdu, Sindhi, Kashmiri
149	خ	LETTER KHAY Urdu, Sindhi, Kashmiri
150	د	LETTER DAAL Urdu, Sindhi, Kashmiri.
151	ڈ	LETTER DHAAL Sindhi
152	ډ/ڈ	LETTER DAAL (retroflex) Urdu, Kashmiri/Sindhi
153	ڈ	LETTER DDAAL (implosive) Sindhi
154	ډ	LETTER DHAAL (retroflex) Sindhi
155	ذ	LETTER ZAAL Urdu, Sindhi, Kashmiri
156	ر	LETTER RE Urdu, Sindhi, Kashmiri
157	ڑ/ڑ	LETTER REY Urdu, Kashmiri/Sindhi
158	ڑھ	LETTER RHEY Sindhi
159	ز	LETTER ZE Urdu, Sindhi, Kashmiri
160	ژ	LETTER ZHAY Urdu, Kashmiri
161	س	LETTER SEEN Urdu, Sindhi, Kashmiri
162	ش	LETTER SHEEN Urdu, Sindhi, Kashmiri

163	ص	LETTER SUAD Urdu, Sindhi, Kashmiri
164	ض	LETTER ZUAD Urdu, Sindhi, Kashmiri
165	ط	LETTER TOE Urdu, Sindhi, Kashmiri
166	ظ	LETTER ZOE Urdu, Sindhi, Kashmiri
167	ع	LETTER AIN Urdu, Sindhi, Kashmiri
168	غ	LETTER GHAIN Urdu, Sindhi, Kashmiri
169	ف	LETTER FE Urdu, Sindhi, Kashmiri
170	ق	LETTER QAAF Urdu, Sindhi, Kashmiri
171	ک/ك	LETTER KAAF Urdu, Kashmiri/Sindhi
172	ک	LETTER KHE Sindhi
173	گ	LETTER GAAF Urdu, Sindhi, Kashmiri
174	ڳ	LETTER GGE Sindhi
175	ڱ	LETTE NGE Sindhi
176	ل	LETTER LAAM Urdu, Sindhi, Kashmiri
177	م	LETTER MEEM Urdu, Sindhi, Kashmiri
178	ن	LETTER NOON Urdu, Sindhi, Kashmiri
179	ں	LETTER NOON GHUNNA Urdu
180	ڻ	LETTER NNOONN (retroflex) Sindhi
181	و	LETTER VAO Urdu, Sindhi, Kashmiri
182	ۄ	LETTER VAO with Ring Kashmiri
183	ہ	LETTER HE Urdu, Sindhi, Kashmiri
184	ھ	LETTER DOCHASHMI HE Urdu, Sindhi, Kashmiri
185	ء	LETTER HAMZA Urdu, Sindhi
186	ي/ى	LETTER YE Urdu/Sindhi
187	ي	LETTER YE (CIRCLE BELOW) Kashmiri
188	ۓ	LETTER BARI YE(HALF CIRCLE ABOVE Kashmiri
189	ے	LETTER BARI YE Urdu, Kashmiri
190	◌َ	Diacritic Mark (Zabar)
191	◌ِ	Diacritic Mark (Zer)
192	◌ُ	Diacritic Mark (Pesh)
193	◌ٗ	Diacritic Mark (Ulta Pesh)
194	◌ٔ	Diacritic Mark (Hamza Above)
195	◌ٕ	Diacritic Mark (Hamza Below)
196	◌ؕ	Diacritic Mark (Hamza Above) Kashmiri
197	◌ؔ	Diacritic Mark (Hamza Below) Kashmiri
198	◌ّ	Diacritic Mark (Tashdeed)
199	◌ٓ	Diacritic Mark (Madd)
200	◌ْ	Diacritic Mark (Jazm)
201	◌ٰ	Diacritic Mark (KhaRa Alif Above)

202	ٕ	Diacritic Mark (KhaRa Alif Below)
203	ؕ	Diacritic Mark (Wasl)
204	...	Punctuation (Tafsiliya)
205	ؔ	Punctuation (Batt)
206	،	Punctuation (Comma)
207	ؐ	Superscript (Suad)
208	ؑ	Superscript (Re-zuad)
209	ؒ	Superscript (Re-he)
210	ؑ	Superscript (Ain)
211	ق	Superscript (Qaf) ARABIC
212	ک	Superscript(Kaaf) ARABIC
213	ﻻ	Superscript (Laam-alif) ARABIC
214	ج	Superscript (Jeem) ARABIC
215	م	Superscript (Meem) ARABIC
216	ســـنه	Year symbol -Gregorian
217	%	Punctuation (Percentage symbol)
218	/	Number & Text separator symbol
219	٫	Decimal point (Asharya)
220	٠	DIGIT 0
221	١	DIGIT 1
222	٢	DIGIT 2
223	٣	DIGIT 3
224	۴	DIGIT 4
225	۵	DIGIT 5
226	٦	DIGIT 6
227	۷	DIGIT 7
228	٨	DIGIT 8
229	٩	DIGIT 9

230	!	Exclamation symbol
231	”	Open double quote
232	“	Close double quote
233	’	Open single quote
234	‘	Close single quote
235	(	Open bracket
236	)	Close bracket
237	*	Start symbol
238	+	Plus sign
239	ATR	Attribute.
240	-	Minus sign
241	/	Forward slash
242	؛	Semi colon
243	:	Colon
244	؟	Question mark
245	=	Equal sign
246	ـ	Sentence dash
247	○	Ayat end (Arabic)
248	•	Filled circle.
249	,	Thousand separator
250	Reserved	(Control char for DM)
251	Reserved	(Control char for LB)
252	◌	Empty Circle
253	.	Dot symbol
254	Reserved	
255	Reserved	