

## 10. Urdu

Urdu derives from Indo-Aryan family of languages and shares basic linguistic structure with Hindi, the two languages being mutually understandable. However, unlike Hindi, Urdu derives more of its vocabulary from Persian and Arabic. Urdu has 104 million speakers in Pakistan, Afghanistan, India, Bangladesh and many other countries [56]. Urdu is the national language of Pakistan and a state language of India. Urdu is written using Arabic script. Perso-Arabic Nastalique style is widely used for Urdu orthography [57, 58].

### 10.1. Writing System

#### 10.1.1. Character Set

Urdu uses characters from the extended Arabic character set used for Persian. It further extends this set to represent sounds which are present in Urdu but not in Arabic or Persian, including aspirated stop and alveolar consonants, and long vowels [59]. Altogether there are 58 letters in Urdu, given in Figure 10.1 ([60]; other sources may give slightly different set).

ا آ ب بھ پ پھ ت تھ ٹ ٹھ ث ج جھ چ چھ ح خ د دھ ڈ ڈھ ذ  
ر رھ ژ ژھ ز ژس ش ص ض ط ظ ع غ ف ق ک کھ گ گھ ل  
لھ م مھ ن نھ ل نھ و وھ ہ ہا ہا ی یھ اے

Figure 10.1. Urdu Character Set

Arabic script uses letters to represent consonants. Diacritics are used to specify the vowels. Urdu has both long and short vowels. Short vowels are indicated by placing diacritics with the consonant which precedes it in the syllable. Long vowels are indicated by a combination of the diacritic on a consonant followed by an additional letter (see [59] for a detailed discussion). These diacritics (also known as Aerab) are normally not written, though are implicitly present, and thus are optional in their usage. In addition to the Aerab which specify the vowels, diacritics are also used to add consonantal sounds, e.g. for germination (i.e. duplication of consonants). These Aerab are given in Figure 10.2 with the letter ب.

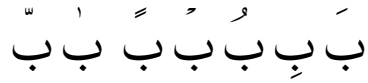


Figure 10.2. Urdu Diacritics

As is evident from the figure, different diacritics can occur above or below the consonant. A consonant may take two diacritics, one consonantal and the other vocalic. In case both diacritics are above the consonant, the consonantal diacritics stack under the vocalic one. These diacritics are always keyed in after the anchoring base letter.

Urdu also has honorific marks which are used to show respect, and are used with proper names. These honorifics are shown in Figure 10.3.

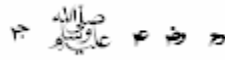


Figure 10.3. Honorific Marks in Urdu

Urdu has its own set of numerals based on numerals used in Arabic and Persian, but some numerals are unique in their shape. These numerals are listed in Figure 10.4.



Figure 10.4. Urdu Numerals

### 10.1.2. Bidirectionality

Urdu inherits the bidirectional property from Arabic script. Urdu words are written from right to left but numbers are written from right to left, as shown in Figure 10.5. However, bidirectionality is handled at rendering level and key press sequence for Urdu alphanumeric input is same as it would be for any other uni-directional language. Thus bidirectionality has no implication on collation.

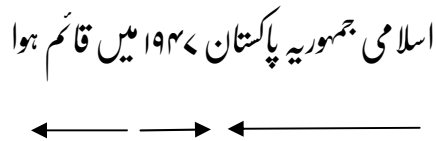
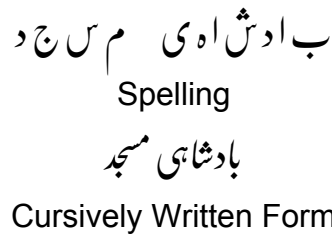


Figure 10.5. Bidirectional Urdu Text

(Arrows indicate reading direction)

### 10.1.3. Cursiveness, Ligation and Context Sensitive Glyph Shaping

Arabic script is cursive, that is, the letters in the script join together into units to form words. These connected units are called ligatures. There are two kinds of characters, joiners and non-joiners. While writing a word, all characters join together until a non-joiner is written. A new ligature starts after the non-joiner (thus, the name “non-joiner”). The process is repeated until the end of the word. In addition, depending on whether the character joins a ligature in the initial, medial or final position, or is unconnected, it takes a different shape. In Nastalique, the character may also change shape depending on the other characters around it. Thus, depending on the context, a single letter may take as many as 25 shapes. Cursiveness is shown in Figure 10.6.



**Figure 10.6. Spelt-out and Cursive Version of Sample Text of Urdu in Nastalique Script**

Again, cursiveness, ligation and context sensitivity are rendering related issues and the though the output shapes of characters may vary with context, their internal encoding remains unchanged. For example, the letter ب may take many shapes but its internal encoding is always U+0628. Therefore, these properties have no implication on collation.

## 10.2. Collation

Urdu collation sequence has been standardized and published by National Language Authority for Pakistan [60]. The collation requires the characters to be sorted at three levels, letters, Aerab and honorifics. However, before the text can be sorted, it has to undergo text processing, as discussed in the next sub-section. Once the text is processed and collation elements are assigned, the regular sort-key generation and comparison process sorts the text.

### 10.2.1. Text Processing

#### 10.2.1.1. Inconsistent Use of Space

Nastalique style of writing does not have the concept of space to separate words. Similar to South-East Asian scripts like Lao, Thai and Khmer, Urdu readers are expected to parse the

ligatures into words as they read along the text. In typing, space is used to get the right character shapes. To achieve this end, it is sometimes used within a word to break the word into constituent ligatures, as in word *یونیورسٹی*. However, if the ligature form is achieved without the use of space, it is sometimes not even used in between two words, e.g. *اردو نط* is visually correct sequence of two words for the readers but has no space between them. This has implications on collation and thus proper word segmentation must be done before strings are collated. Currently there are no automatic word segmentation utilities available for Urdu and therefore the input for collation must be manually cleaned.

#### **10.2.1.2. Diacritics for Loan Words**

The diacritics used for Urdu are given in the figure above. However, there are additional diacritics which are sometimes used with loan words from Arabic. Though not part of Urdu, they have to be processed in case of loan words. Thus, they are also included in the collation element table.

#### **10.2.1.3. Normalization**

Two kinds of normalization are required for Urdu. First, a letter may be represented by multiple Unicode points, and thus the redundancy in encoding has to be cleaned in raw text before further processing. For example, letter *ی* may be represented by Unicode points U+0649, U+064A, and U+06CC in Urdu<sup>1</sup>. Second, a letter or a ligature is sometimes encoded in composed form as well as decomposed form. Thus, the two equivalent representations must also be reduced to same underlying form before further processing. This category includes two sub-categories. One category combines marks and base characters to form other characters. Other combines multiple base characters to form a ligature. Table 10.1 below gives some examples.

---

<sup>1</sup> These codes are normally used in Urdu corpus online to represent *ی* character. Additional codes in Arabic Presentations Forms are not listed here. Unicode does not recommend the use of this area, which was originally used for backward compatibility with legacy systems.

**Table 10.1. Composed and Decomposed Forms of an Urdu Character and a Ligature**

Glyph	Unicode	Individual letters/marks	Unicode Points
آ	0622	ا ~	0653 0627
لا	FEFB	ا ل	0627 06F1

There are many such characters and ligatures which can be represented in multiple ways. Many are not recommended by the Unicode standard, but users still use them due to the similarity of glyphs. An example is using Arabic digits for Urdu language (U+0660 – U+0669), where a separate similar looking set is also encoded (U+06F0 – U+06F9) for use in Arabic language.

#### 10.2.1.4. Contraction

In Urdu character ه (U+06BE or U+0647<sup>2</sup>) combines with most obstruents<sup>3</sup> to represent their aspirated version. Though the constituents are encoded separately, they combine to give a singular character with a single collation element. Thus, these combinations have to be contracted before collation elements are assigned. Some examples of these contractions are given in Figure 10.7.

$$بھ = ب + ه$$

$$چھ = چ + ه$$

$$دھ = د + ه$$

**Figure 10.7. Contraction of Letters with ه in Urdu**

There is no Unicode point available to directly encode the contracted form for aspirated obstruents.

















<sup>2</sup> Not recommended for use in Urdu but is used in the online Urdu corpus.

<sup>3</sup> Phonological term for all sounds which cause a constriction in the oral tract during articulation.

## 10.2.2. Unicode Collation Elements

Collation Elements for Urdu character set are given in Table 10.2 below.

**Table 10.2. Urdu Collation Elements**

<b>Glyph</b>	<b>Uni-code</b>	<b>Collation Elements</b>	<b>Unicode Name</b>
ZWNJ	200C	0000 0010 0002	ZERO WIDTH NON-JOINER
<b>← Diacrtics →</b>			
	0652	0000 00C4 0002	ARABIC SUKUN
	064E	0000 00C9 0002	ARABIC FATHA
	0650	0000 00CA 0002	ARABIC KASRA
	064F	0000 00CB 0002	ARABIC DAMMA
	0670	0000 00CD 0002	ARABIC LETTER SUPERSCRIPT ALEF
	0656	0000 00D5 0002	ARABIC SUBSCRIPT ALEF
	0657	0000 00D8 0002	ARABIC INVERTED DAMMA
	064B	0000 00DB 0002	ARABIC FATHATAN
	064D	0000 00DE 0002	ARABIC KASRATAN
	064C	0000 00E2 0002	ARABIC DAMMATAN
	0654	0000 00E5 0002	ARABIC HAMZA ABOVE
	0651	0000 00E8 0002	ARABIC SHADDA
	0658	0000 00EA 0002	ARABIC MARK NOON GHUNNA
	0653	0000 00F1 0002	ARABIC MADDAH ABOVE
<b>← Honorifics and Special Signs →</b>			
	0610	0000 0000 000A	ARABIC SIGN SALLALLAHOU ALAYHWASSALLAM
	0611	0000 0000 001A	ARABIC SIGN ALAYHE ASSALLAM

۞	0613	0000 0000 002A	ARABIC SIGN RADI ALLAHOU ANHU
۞	0612	0000 0000 003A	ARABIC SIGN RAHMATULLAH ALAYHE
۞	0614	0000 0000 004A	ARABIC SIGN TAKHALLUS
<b>← Punctuation Marks →</b>			
۞	0600	0000 0000 0000	ARABIC NUMBER SIGN
۞	0601	0000 0000 0000	ARABIC SIGN SANAH
۞	0602	0000 0000 0000	ARABIC FOOTNOTE MARKER
۞	0603	0000 0000 0000	ARABIC SIGN SAFHA
۞	0615	0000 0000 0000	ARABIC SMALL HIGH TAH
،	060C	0000 0000 0000	ARABIC COMMA
،	060D	0000 0000 0000	ARABIC DATE SEPARATOR
،	066B	0000 0000 0000	ARABIC DECIMAL SEPARATOR
،	066C	0000 0000 0000	ARABIC THOUSANDS SEPARATOR
؟	061F	0000 0000 0000	ARABIC QUESTION MARK
؛	061B	0000 0000 0000	ARABIC SEMICOLON
-	06D4	0000 0000 0000	ARABIC FULL STOP
%	066A	0000 0000 0000	ARABIC PERCENT SIGN
ء	060E	0000 0000 0000	ARABIC POETIC VERSE SIGN
ء	060F	0000 0000 0000	ARABIC SIGN MISRA
لا	FEFB	[13AB 0020 0002],[ 1350 0020 0002]	ARABIC LIGATURE LAAM WITH ALEF ISOLATED FORM
الله	FDF2	[13AB 0020 0002], [13AB 0020 0002], [13AB 0020 0002],[ 13D3 0020 0002]	ARABIC LIGATURE ALLAH

و	0624	[13BD 0020 0002],[0000 00E5 0002]	ARABIC LETTER WAW WITH HAMZA ABOVE
ي	0626	[13C9 0020 0002],[0000 00E5 0002]	ARABIC LETTER CHOTI YEH WITH HAMZA ABOVE
أ	0623	[1350 0020 0002],[0000 00E5 0002]	ARABIC LETTER ALEF WITH HAMZA ABOVE
<b>← Numerals →</b>			
٠	06F0	0E29 0020 0002	ARABIC-INDIC DIGIT ZERO
١	06F1	0E2A 0020 0002	ARABIC-INDIC DIGIT ONE
٢	06F2	0E2B 0020 0002	ARABIC-INDIC DIGIT TWO
٣	06F3	0E2C 0020 0002	ARABIC-INDIC DIGIT THREE
٤	06F4	0E2D 0020 0002	ARABIC-INDIC DIGIT FOUR
٥	06F5	0E2E 0020 0002	ARABIC-INDIC DIGIT FIVE
٦	06F6	0E2F 0020 0002	ARABIC-INDIC DIGIT SIX
٧	06F7	0E30 0020 0002	ARABIC-INDIC DIGIT SEVEN
٨	06F8	0E31 0020 0002	ARABIC-INDIC DIGIT EIGHT
٩	06F9	0E32 0020 0002	ARABIC-INDIC DIGIT NINE
ا	0627	1350 0020 0002	ARABIC LETTER ALEF
آ	0627 0653	1351 0020 0002	ARABIC LETTER ALEF WITH MADDA ABOVE
أ	0622	1351 0020 0002	ARABIC LETTER ALEF WITH MADDA ABOVE
ب	0628	1352 0020 0002	ARABIC LETTER BEH
بھ	0628 06BE	1353 0020 0002	ARABIC LETTER BEH + ARABIC LETTER HEH DOCHASHMEE

پ	067E	1354 0020 0002	ARABIC LETTER PEH
پھ	067E 06BE	1355 0020 0002	ARABIC LETTER PEH + ARABIC LETTER HEH DOCHASHMEE
ت	062A	1357 0020 0002	ARABIC LETTER TEH
تھ	062A 06BE	1358 0020 0002	ARABIC LETTER THE + ARABIC LETTER HEH DOCHASHMEE
ط	0679	135A 0020 0002	ARABIC LETTER TTEH
طھ	0679 06BE	135B 0020 0002	ARABIC LETTER TTEH + ARABIC LETTER HEH DOCHASHMEE
ث	062B	135D 0020 0002	ARABIC LETTER THEH
ج	062C	135E 0020 0002	ARABIC LETTER JEEM
جھ	062C 06BE	135F 0020 0002	ARABIC LETTER JEEM + ARABIC LETTER HEH DOCHASHMEE
چ	0686	1361 0020 0002	ARABIC LETTER TCHEH
چھ	0686 06BE	1362 0020 0002	ARABIC LETTER TCHEH + ARABIC LETTER HEH DOCHASHMEE
ح	062D	1364 0020 0002	ARABIC LETTER HAH
خ	062E	1365 0020 0002	ARABIC LETTER KHAH
د	062F	1369 0020 0002	ARABIC LETTER DAL
دھ	062F 06BE	136A 0020 0002	ARABIC LETTER DAL + ARABIC LETTER HEH DOCHASHMEE
ڈ	0688	136B 0020 0002	ARABIC LETTER DDAL
ڈھ	0688 06BE	136C 0020 0002	ARABIC LETTER DDAL + ARABIC LETTER HEH DOCHASHMEE
ذ	0630	1370 0020 0002	ARABIC LETTER THAL
ر	0631	1375 0020 0002	ARABIC LETTER REH

رھ	0631 06BE	1376 0020 0002	ARABIC LETTER REH + ARABIC LETTER HEH DOCHASHMEE
ڑ	0691	1377 0020 0002	ARABIC LETTER RREH
ڑھ	0691 06BE	1378 0020 0002	ARABIC LETTER RREH + ARABIC LETTER HEH DOCHASHMEE
ز	0632	137C 0020 0002	ARABIC LETTER ZAIN
ژ	0698	137E 0020 0002	ARABIC LETTER JEH
س	0633	1381 0020 0002	ARABIC LETTER SEEN
ش	0634	1382 0020 0002	ARABIC LETTER SHEEN
ص	0635	1387 0020 0002	ARABIC LETTER SAD
ض	0636	1388 0020 0002	ARABIC LETTER DAD
ط	0637	138C 0020 0002	ARABIC LETTER TAH
ظ	0638	138D 0020 0002	ARABIC LETTER ZAH
ع	0639	138F 0020 0002	ARABIC LETTER AIN
غ	063A	1390 0020 0002	ARABIC LETTER GHAIN
ف	0641	1393 0020 0002	ARABIC LETTER FEH
ق	0642	139B 0020 0002	ARABIC LETTER QAF
ک	06A9	139F 0020 0002	ARABIC LETTER KEHEH
کھ	06A9 06BE	13A2 0020 0002	ARABIC LETTER KEHEH + ARABIC LETTER HEH DOCHASHMEE
گ	06AF	13A5 0020 0002	ARABIC LETTER GAF
گھ	06AF 06BE	13A6 0020 0002	ARABIC LETTER GAF + ARABIC LETTER HEH DOCHASHMEE

ل	0644	13AB 0020 0002	ARABIC LETTER LAM
لھ	0644 06BE	13AC 0020 0002	ARABIC LETTER LAM + ARABIC LETTER HEH DOCHASHMEE
م	0645	13B0 0020 0002	ARABIC LETTER MEEM
مھ	0645 06BE	13B1 0020 0002	ARABIC LETTER MEEM + ARABIC LETTER HEH DOCHASHMEE
ن	0646	13B4 0020 0002	ARABIC LETTER NOON
نھ	0646 06BE	13B5 0020 0002	ARABIC LETTER NOON + ARABIC LETTER HEH DOCHASHMEE
ں	06BA	13B9 0020 0002	ARABIC LETTER NOON GHUNNA
ںھ	06BA 06BE	13BA 0020 0002	ARABIC LETTER NOON GHUNNA + ARABIC LETTER HEH DOCHASHMEE
و	0648	13BD 0020 0002	ARABIC LETTER WAW
وھ	0648 06BE	13BE 0020 0002	ARABIC LETTER WAW + ARABIC LETTER HEH DOCHASHMEE
ہ	06C1	13C2 0020 0002	ARABIC LETTER HEH GOAL
ھ	06BE	13C4 0020 0002	ARABIC LETTER HEH DOCHASHMEE
ہٴ	06C3	13C6 0020 0002	ARABIC LETTER TEH MARBUTA GOAL
ع	0621	13C7 0020 0002	ARABIC LETTER HAMZA
ی	06CC	13C9 0020 0002	ARABIC LETTER FARSI YEH
یھ	06CC06BE	13CB 0020 0002	ARABIC LETTER FARSI YEH + ARABIC LETTER HEH DOCHASHMEE
ے	06D2	13CE 0020 0002	ARABIC LETTER YEH BARREE

### 10.3. Results

The sorting performed using the collation elements given results in the following sequence.

**Table 10.3. Input and Corresponding Sorted Output for Urdu**

Sample Output		Sample Input	
دائرة	اب	بہن	بہنگی
دائرة المعروف	ابھی	بی بی	اگنا
زکوت	اگنا	عمر	بیٹی
زکوہ	ایمان	دائرة المعروف	دائرہ
زکوۃ	آب	آبن	گنا
زکوۃ	آبن	عمر	عمر
عمر	بن	مان	گنا
عمر	بن	بی	آب
عمر	بن	گنا	ابھی
عمر	بہن	زکوۃ	ایمان
گنا	بی	بے	ٹیلیفون
گنا	بیٹی	زکوۃ	عمر
گنا	بے	زکوت	مان
گنا	بہنگی	زکوہ	ٹیلی فون
مان	ٹیلیفون	بن	گنا
مان	ٹیلیفون	دائرة	اب
	دائرہ	بن	گنا
			بن

### 10.4. Conclusion

Sorting in Urdu is carried out at three different levels. Letters are sorted at primary level, diacritics are handled at secondary level, and honorifics are handled at tertiary level. Normalization and contraction are also required for Urdu collation. However, regular sorting algorithm is applicable after appropriate text processing is done and collation elements are assigned.