8.3.1 Urdu Alphabet

There are some peculiarities that can be seen in Perso-Arabic scripts:

- These scripts join letters with each other, and therefore letters have different forms as per their position in a ligature (the positions are beginning, middle, ending).
- Some shapes do not have middle shape i.e. they do not join at both the ends. For example alif, vao, daal, etc.
- Every letter has a standalone form

Following is a list of standalone shapes, which have been grouped by the similitude of their joins.

Characters	Group Shape	Connects (has middle shape)
ب پ ت ٹ ث	ب	Yes
ج چ ح خ	ج	Yes
د ڈ ذ	د	No
ر ڑ ز ژ	ر	No
س ش	س	Yes
ص ض	ص	Yes
ط ظ	ط	Yes
ع غ	ع	Yes
ف ق	ف	Yes
ک گ	ک	Yes

Vowels

The earliest Arabic script had no vowel markers, as the structure of the language and the context served to make the passage clear. Later on, however, a set of diacritics (zabar, zer, and pesh) was developed, written above or below the consonant symbol which they follow.

The Alphabet Chart (URDU)

Letter	Name	Connects	Homonyms
ا	Alif	No	ع
ب	Be	Yes	
پ	Pe	Yes	
ت	te	Yes	
ٹ	Te	Yes	
ث	Se	Yes	
ج	Jeem	Yes	
چ	Cheem	Yes	
ح	baRi-Hai	Yes	
خ	Khai	Yes	
د	Daal	No	
ڈ	Daal (retrofl.)	No	
ذ	Zaal	No	ظ ض ز
ر	Re	No	
ڑ	De	No	
ز	Ze	No	ذ ظ ض
ژ	Zhe	No	
س	Seen	Yes	ث ص
ش	Sheen	Yes	
ص	Suad	Yes	
ض	Zuad	Yes	
ط	Toe	Yes	
ظ	Zoe	Yes	ذ ض ز
ع	Ain	Yes	
غ	Ghain	Yes	گ
ف	Fe	Yes	
ق	Qaaf	Yes	ک
ک	Kaaf	Yes	
گ	Gaaf	Yes	
ل	Laam	Yes	
م	Meem	Yes	
ن	Noon	Yes	
ں	noon-ghunna	Yes	
و	Vao	No	
ہ	Hai	Yes	ھ
ھ	dochashmi-hai	Yes	
ی	Ye	Yes	
ے	baRi-ye	Yes	

Name	Final	Middle	Begin	Standalone	Name	Final	Middle	Begin	Standalone
Alif	ا			ا	Be	ب	بـ	بـ	ب
pe	پ	پـ	پـ	پ	te	ت	تـ	تـ	ت
Te	ٹ	ٹـ	ٹـ	ٹ	se	ث	ثـ	ثـ	ث
jeem	ج	جـ	جـ	ج	ce	چ	چـ	چـ	چ
baRi-Hai	ح	حـ	حـ	ح	khai	خ	خـ	خـ	خ
daal	د			د	daal	ڈ			ڈ
zaal	ذ			ذ	re	ر			ر
Re	ڑ			ڑ	ze	ز			ز
zhe	ژ			ژ	seen	س	سـ	سـ	س
sheen	ش	شـ	شـ	ش	suad	ص	صـ	صـ	ص
zuad	ض	ضـ	ضـ	ض	toe	ط	طـ	طـ	ط
zoe	ظ	ظـ	ظـ	ظ	ain	ع	عـ	ء	ع
ghain	غ	غـ	غـ	غ	fe	ف	فـ	فـ	ف
qaaf	ق	قـ	قـ	ق	kaaf	ک	کـ	کـ	ک
gaaf	گ	گـ	گـ	گ	laam	ل	لـ	لـ	ل
meem	م	مـ	مـ	م	noon	ن	نـ	نـ	ن
vao	و			و	hai	ہ	ہـ	ہـ	ہ
hai	ھ	ھـ	ھـ	ھ	hamza	ئ	ئـ	ئـ	ء
ye	ی	یـ	یـ	ی	baRi-ye	ے			ے

Technical characteristics

Urdu Alphabet Characteristics

Urdu alphabet utilizes consonant letters, vowels, diacritic marks, numerals, punctuations and a few superscripts signs.

The graphic representation of each alphabet has more than one form depending on its position and context in the word. In general each letter has four forms that is beginning, middle, final and standalone. The graphic representation (standalone form) of these alphabets can be found in the reference section, together with the naming convention according to the Unicode Standard Version 3.0.

Consonants or Basic Letters

There are total 39 alphabets. Each has, in general, four connection forms depending on its position in a word. There are consonants with similar phonetic sounds (homonyms). For example letters "THEH", "SEEN", "SAD" (U+062B, U+0633, U+0635 respectively) have a similar phonetic sound in Urdu. Similarly there are other consonants that have this characteristic.

Vowels

The letters "ALIF" (U+0627), "WAW" (U+0648), "HAMZA" (U+0621), "YEH" (U+06CC) serve as vowels in Urdu. These letters are part of the basic letters that form the character set of Urdu.

Diacritic Marks

In addition to above vowels, there are diacritic marks that appear above or below a character to specify a vowel or emphasize a particular sound. These diacritics have been taken from Arabic script.

The common diacritical marks used are "FATHA" (zabar U+064E), "KASAR" (zer U+0650), "DAMMA" (pesh U+064F), "SHADDA" (tashdeed U+0651), etc. The other diacritics developed in the course of time are آ (as in star), او (as in go), اؤ (as in how), اُو (as in cool), etc. However, the vowel diacritics are generally not written in materials intended for native speakers of the language.

Erabs

zabar ـَ	zer ـِ	pesh ـُ	tashdid ـّ
jazm ـْ	alif-short ـٰ		Mad ـٓ

Other diacritics: wasl, bat, etc.

Punctuations

question mark	؟	comma	،
semi-colon	؛	colon	:
full stop		dash	ـ
opening bracket	(	closing bracket	)
exclamation mark	!	forward slash	/
open quote		close quote	

Numerals

Zero	۰	Five	۵
One	۱	Six	۶
Two	۲	Seven	۷
Three	۳	Eight	۸
Four	۴	Nine	۹

Consonant Conjuncts

As the script joins letters that have different forms, there are some combinational forms (ligatures) used to write certain combinations of letters. For example "LAAM" (U+0644) and "ALIF" (U+0627) are not written like taking starting shape of LAAM and final shape of ALIF, but in a different way. Similarly, many ligatures are possible, for example : laam-meem, laam-jeem, laam-meem-jeem, etc. These ligatures are part of the script meant to beautify the text.

Other Superscripts & Diacritical Signs

There are certain superscripts commonly used in Urdu text. These are abbreviations for words. Examples of some superscripts are "re-ze" superscript (short form for "Raziyallahu-Anhu"), "suad" superscript (short for "Sallallahu-Alaihe-Wassallam"), etc.

Apart from these superscripts, there are certain signs such as "batt", used over the nickname of a poet. Example of other sign is "wasl".

Kashida

The letter with kashida is written with a stretch in its width and shape. Not all but most of the letters can be written in kashida.

Numerals

The Urdu script has its own numerals. The international numerals are also used in place of Urdu numerals. Numerals in Urdu are written left to right. The decimal separator in Urdu numerals is called "ASHARYA" (U+066B) and is similar to "HAMZA" in shape. A dot may also be used in place of "ASHARYA".

Punctuation Marks

Urdu borrows some punctuation marks from English, some of which have been modified to suit the right to left script behavior. For example a question mark is written like the English question mark flipped horizontally. Examples of other punctuation marks used in Urdu are Division sign, Exclamation, Sentence Dash, etc.

Character Set Considerations

Urdu Characteristics

In addition to 39 basic characters, there are a few diacritics (five in number), numerals, a few superscripts, and punctuation markers.

Fonts

Considering the Nastaliq script, as mentioned earlier that it being rich in calligraphic shapes, multiple alternate shapes are possible for a single letter. The shape is decided by the context i.e. position of the character in a word and/or character next to it.

Considering the Naskh script, there can be at the most four shapes of a character; the shapes being starting, medial, final and standalone.

Character Cell Size

The characters in Urdu do not have the same height and width. The character cell can be given the height of the tallest character. But considering Nastaliq i.e. there is a vertical shift of characters involved, one will have to put a limit on the levels it can have. In general, meaningful words will not be extended in height. Only a non-meaningful word can have a flexible height. This will have to be controlled by the rendering mechanism, and break a word whenever required or possible.

Glyphs to be supported in Urdu Fonts

All the basic shapes plus alternate shapes required for a character have to be provided. A single character thus would have at least four or more glyphs for it. The diacritic marks will have to be provided, and these may have multiple forms in levels, as characters in Urdu do not have the same height. Further, the common religious and linguistic symbols are used as diacritics or superscripts. These symbols may also have levels defined. The Urdu numerals, and punctuations will have to be povided in the font. Additionally, a number of ligatures will have to be provided to justify the calligraphic need of the script. Ligatures that are mandatory, for example *laam-alef* will have to be provided.

Keyboards

There is no standard keyboard available for Urdu language. There are vendor specific keyboards available which all differ in their layouts.

There are a lot of keyboards designed by different companies who provide Perso-Arabic support in their applications.

Unfortunately, every company has its own keyboard layout that entirely differs from the other's keyboard Layouts. For example Microsoft has its own keyboard layouts for Perso-Arabic languages.

i. Drawbacks of available keyboards.

ii. No standard keyboard available.

iii. All vendor specific keyboards differ in their layout.

iv. Most of the keyboards have been designed for Arabic

Although Arabic and Urdu have similar script, they are both different languages, and hence the keyboard designed for Arabic may not be that useful for Urdu language. C-DAC, GIST has designed a keyboard for Urdu, and has come up with an optimum solution as given below.

C-DAC Urdu Keyboard

Shifted

Unshifted

Keyboard Characteristics

Erabs:

i. Erab (diacritic marks for short vowels) have been isolated from consonants, so that they are available in shifted state.

ii. The erabs are also grouped e.g. zabar, zer. hamza-above, hamza-below are placed in key orders.

Superscripts:

i. Superscripts like re-zuad, suad, bat etc, have been grouped and positioned on the top line.

ii. This isolates them from erab and also gives a block view to the keyboard.

Consonants:

i. Consonants have been placed according to their phonetic values wherever suitable e.g. kaaf is placed on k, jeem on j, etc.

ii. Homonyms (similar sounding letters) like zuad, zaal, zhe, zoe are placed nearby and the ones that are rarely used like zhe are placed in the shifted state.

iii. Care has been taken to provide maximum consonants in the un-shifted state.

Combinations

i. Combinations like laam-alif, vao-hamza, etc have been provided and placed in the shifted state.

ii. Also combinations like alif-hamza-above, alif-hamza-below are provided.

Numerals & Punctuation

i. Numerals and punctuations have been grouped and placed near by.

Urdu Composing

Unlike English script, the characters in Urdu have different shapes depending on their context in the constituent word. Further, there can be alternate forms of each of these shapes, again depending on character next to it. In general, there can be four shapes of a character as per its position in the word i.e. starting, medial, and final. The fourth being standalone or isolated. Following is a sample text composed in Naskh and Nastaliq scripts.

نستعلیق — Script Naskh

نستعلیق — Script Nastaliq

A Sample Text in Naskh & Nastaliq

Formats of Units & Localization

Calendar

The commonly used calendars in the Urdu language are the Hijri and Gregorian. The Hijri Calendar is usually abbreviated AH in Western languages from the Latinized Anno Hejirae “In the year of the Hijra”. 1st of Muharram-ul-Haraam, AH 1 corresponds to Friday July 16th, 622 CE in the Julian calendar. There are 12 months in Hijri calendar (from Muharram-ul-Haraam to Dzul-Hajj), however no abbreviations are used for writing these month names in Urdu.

Week

Urdu has specific weekday names: Peer (Monday), Mangal (Tuesday), Buddh (Wednesday), Jume’raat

(Thursday), Jum'a(Friday), Saneechar(Saturday), and Itwaar(Sunday). No abbreviations are used for writing weekday names.

Date

Dates in Urdu are written using date number, followed by an optional forward slash-like shape less in height (this shape is generally used to separate numbers and month name, or any other unit name used), then month name (month numbers are generally not written), and then the year.

Hamza is written at the end to indicate Gregorian date, and a *Dochashmi-he* to indicate Hijri date. An optional word *San* with extended *Seen* is sometimes written below the year numbers (that is year numbers are written above the extended *Seen* of the word *San*). Following is an example of sample dates written in Urdu.

۱۲ / ذالحجہ ۱۴۲۲ء or مارچ ۲۰۰۲ء or مارچ ء ۲۰۰۲ سنہ

Sample Dates In Urdu

Time

There is no abbreviation (like AM or PM in English) used for time in Urdu. The time whether morning or evening is written in text. Following is example of time in Urdu.

صبح ۱۰ بجکر ۵۴ منٹ　　شام کے ۵ بجکر ۱۰ منٹ

10:54AM　　05:10 PM

Sample Time in Urdu

Number

Numerals in Urdu are written left to right just like they are written in English. Urdu.However has its own shapes for the numerals, and the decimal point is written like a thin Hamza shape. Following are examples of numerals in Urdu.

۱۰,۰۰۰ء۵۵　　۲۰۰　　۱۴ء۵۰

10,000.55　　200　　14.50

Sample numbers in Urdu

Currency

There are no currency symbols defined. The name of the currency is written after the currency value.

۱۰۰ روپے　　۲۵ڈالرس

100 Rupees　　25 Dollars

Sample Currency in Urdu.

Collating Sequence and Sorting

The collating sequence can be seen in the PASCII standard proposed by C-DAC, GIST. The sorting sequence has a few issues. The sorting sequence for the basic characters (excluding Erabs or diacritics) is fixed from *Alif* to *Ye*. The diacritics are skipped or ignored when sorting out the correct order. Short vowels may optionally have an order that is *Zabar*, *Zer*, and then *Pesh*, followed by *Tashdeed*.

Character, Line, Word and Sentence Break Rules

Letters *Alif*, *Daal*, *Zaal*, *Waov*, *Choti-ye*, etc. always break a word as the characters following them do not join with them. Space character can be used to put explicit break for the words.

Line Heights however will have to be decided by the user, as there can be vertical shift in the joins, and certain ligatures may vary in height.