

UZT 1.01 & UNICODE MAPPING FOR URDU



**CENTER FOR RESEARCH IN URDU LANGUAGE PROCESSING
NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES**

Center for Research in Urdu Language Processing (www.crup.org)

Introduction:

This document contains the mapping of UZT 1.01 (Urdu Zabita Takhti) Standard on the Unicode. UZT version 1.01 has been accepted by the Government of Pakistan as a standard code page for Urdu. It contains all the characters which are used in Urdu language. This character set includes the Control characters, Punctuations and arithmetic symbols, Digits, Urdu aerab, Urdu characters. Unicode is the two byte world wide standard character set. It contains almost complete character sets of all scripts of the world.

This document also deals with the duplication of Unicode of a single Urdu character and finalizes the single code. Some new codes for Urdu have been proposed in the future versions of Unicode which did not exist in the version 3.0. These codes are also part of the documents. These will be available in Unicode 4.0 or later versions.

Urdu Zabita Tahti (UZT)

UZT is the 256 bit code page. It has been divided into various logical sections.

- Control characters (0-31, 127)
- Punctuation and arithmetic symbols (32 – 47, 58-65)
- Digits (48-57)
- Urdu aerab/diacritics (66 – 79, 123 – 126)
- Urdu characters (80 – 122)
- Reserved control space (128 – 159, 255)
- Special symbols (160 – 176, 192 – 199)
- Reserved expansion space (177 – 191, 200 – 207, 240 – 253)
- Vendor area (208 – 239)
- Toggle character (254)

Unicode 3.0 / 3.2

Unicode standard is developed by the “The Unicode Consortium”. Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language.

Arabic, Urdu, and Persian scripts contain many similar characters and are placed under the general *Arabic* group (code range (hex): 0600-06FF).

Unicode standard version 3.2.0 contains three different Arabic code blocks. These code blocks with names and ranges are listed below

1. Arabic (U0600–U06FF)
2. Arabic Presentation Form-A (UFB50-UFDFF)
3. Arabic Presentation Form-B (UFE70-UFEFF)

First Code block of Arabic contains the standard Arabic characters and diacritics in accordance with ISO 8859-6 standard. It also contains characters listed below

1. Arabic-Indic digits (U0660-U0669)
2. Punctuation (% , decimal separator, * , and thousand separator)

Center for Research in Urdu Language Processing (www.crupl.org)

3. Archaic letters (dot less Beh, and Qaf)
4. Diacritics
5. Extended Arabic letters (U0671-U06D5)
6. Quranic annotation signs (U06D6-U06ED)
7. Eastern Arabic-Indic digits (U06F0-U06F9): *Digits for Persian, Sindhi, Urdu Etc.*
8. Signs for Sindhi (U06FD-U06FE)

Second code block of Arabic script contains not much needed codes. It contains codes for contextual shapes of letters and ligatures (ligatures are not to be used in normal contexts). This block contains

1. Glyphs for contextual forms of letters of Persian, Urdu, Sindhi etc (UFB50-UFBE9)
2. Two Elements Ligature (UFBEA-UFD3D)
3. Punctuation marks (UFD3E-UFD3F)
4. Three elements Ligatures (UFD50-UFD7C)
5. Word Ligatures (UFD7D-UFD7E)
6. Currency Sign of Riyal (UFD7F)

Third code block of Arabic script contains codes for Arabic letters contextual parts, some spacing forms of Arabic and special character (Zero-Width No-Break Space) (ligatures are not to be used in normal contexts). This code block contains

1. Elements for spacing forms of Arabic (like do-zabar) (UFE70-UFE7F)
2. Arabic language contextual forms (UFE80-UFEFC)
3. Special character (Zero-Width No-Break Space) (U200B)

It has complete character set of UZT except some of them. These codes are going to be added in latter version of Unicode. These codes are 0610, FEFB, 0612, 0613, 0614, 0611, 0658, 060F, 0603, 0659, FDFD, 0600, 0601, 0602, 060E, 0657.

Urdu Character	UNICODE (Hex)	Unicode Name	UZT (Hex)
ا	0627	Arabic Letter Alef	50
آ	0622 0627 + 0653 (Optional)	Arabic Letter Alef with Madda above	52
أ	0623 0627 + 0654 (Optional)	Arabic Letter Alef with Hamza above	51
ب	0628	Arabic Letter Beh	53
پ	067E	Arabic Letter Peh	54
ت	062A	Arabic Letter Teh	55

ت	0679	Arabic Letter Tteh	56
ث	062B	Arabic Letter Theh	57
ج	062C	Arabic Letter Jeem	58
چ	0686	Arabic Letter Tcheh	59
ح	062D	Arabic Letter Hah	5A
خ	062E	Arabic Letter Khah	5B
د	062F	Arabic Letter Dal	5C
ڈ	0688	Arabic Letter Ddal	5D
ذ	0630	Arabic Letter Thal	5E
ر	0631	Arabic Letter Reh	5F
ڑ	0691	Arabic Letter Rreh	60
ز	0632	Arabic Letter Zain	61
ژ	0698	Arabic Letter Jeh	62
س	0633	Arabic Letter Seen	63
ش	0634	Arabic Letter Sheen	64
ص	0635	Arabic Letter Sad	65
ض	0636	Arabic Letter Dad	66
ط	0637	Arabic Letter Tah	67
ظ	0638	Arabic Letter Zah	68
ع	0639	Arabic Letter Ain	69
غ	063A	Arabic Letter Ghain	6A

فا	0641	Arabic Letter Feh	6B
قا	0642	Arabic Letter Qaf	6C
كا	06A9	Arabic Letter Kaf	6D
گا	06AF	Arabic Letter Gaf	6E
لا	0644	Arabic Letter Lam	6F
ما	0645	Arabic Letter Meem	70
نا	06BA	Arabic Letter Noon Ghunna	71
ن	0646	Arabic Letter Noon	72
وا	0648	Arabic Letter Waw	73
ھا	06BE, 0647(not recommended)	Arabic Letter Heh Doachashmee	7A
ہ	06C1	Arabic Letter Heh Goal	75
ة	0629	Arabic Letter The Marbuta	
ء	0621, 0654(Optional), 0674(not recommended)	Arabic Letter Hamza	77
ی	06CC, 0649 (Optional)	Arabic Letter Farsi Yeh	78
ے	06D2	Arabic Letter Yeh Barree	79
ؤ	0624 0648 + 0647 (Optional)	Arabic Letter Waw with Hamza above	74
ئ	0626 064A + 0674 (Optional)	Arabic Letter Yeh with Hamza above	
ئے	06D3 06D2 + 0654 (Optional)	Arabic Letter Yeh Barree with Hamza above	
ء	0654	Arabic Hamza below	42 (بمزه اضافت)
-	0650	Arabic Kasra	43 (کسرہ اضافت)

ه	06C2 06C1 + 0654 (Optional)	Arabic Letter Heh Goal with Hamza above	
=	064B	Arabic Fathatan	49
ھ	064C	Arabic Dammatan	4B
=	064D	Arabic Kasratan	4A
-	064E	Arabic Fathah	7C
ـ	064F	Arabic Damma	7E
-	0650	Arabic Kasra	7D
ا	0670	Arabic Letter Superscript Alef	44
ا	0656 (Proposed)	Arabic Subscript Alef	45
”	0652	Arabic Sukun (Jazm; inclusion of separate Jazm code is still under consideration of Unicode)	4D
(Let aPesh)			47
(Let i zer)			48
\	2018, 0027	Left Single Quotation, Apostrophe	27
!	0021	Exclamation Mark	21
؛	061B	Arabic Semicolon	3B
:	003A	Colon	3A
”	0022	Quotation Mark	22
،	060C	Arabic Comma	2C
۔	06D4	Arabic Full Stop	2D
؟	061F	Arabic Question Mark	3F
(0028	Left Parenthesis	38

)	0029	Right Parenthesis	39
[005B	Left Square Bracket	C0
]	005D	Right Square Bracket	C2
{	007B	Left Curly Bracket	C4
}	007D	Right Curly Bracket	C6
٠	066B	Arabic Decimal Separator	2E (Urdu Decimal Point)
.	06F0	Extended Arabic-Indic Digit Zero	30
١	06F1	Extended Arabic-Indic Digit One	31
٢	06F2	Extended Arabic-Indic Digit Two	32
٣	06F3	Extended Arabic-Indic Digit Three	33
٤	06F4	Extended Arabic-Indic Digit Four (different shape for Urdu)	34
٥	06F5	Extended Arabic-Indic Digit Five	35
٦	06F6	Extended Arabic-Indic Digit Six (different shape for Urdu)	36
٧	06F7	Extended Arabic-Indic Digit Seven (different shape for Urdu)	37
٨	06F8	Extended Arabic-Indic Digit Eight	38
٩	06F9	Extended Arabic-Indic Digit Nine	39
-	002D	Hyphen-Minus	2D
+	002B	Plus Sign	2B
=	003D	Equals Sign	3D
/	2215	Division Slash	2F

*	002A 066D (Optional)	Asterisk	2A
%	066A	Arabic Percent Sign	25
@	0040	Commercial At	40
#	0023	Number sign	23
&	0026	Ampersand	26
—	005F	Low Line	C3
~	0653	Arabic Maddah Above	AE
	007C	Vertical Line	C7
\	005C	Backslash	C1
<	003C	Less-than sign	3C
>	003E	Greater-than sign	3E
الله	FDF2 0627 + 0644 + 0644 + 0647 (Optional)	Arabic Ligature Allah Isolated Form	A0
اكبر	FDF3 0627 + 0643 + 0628 + 0631 (Optional)	Arabic Ligature Akbar Isolated Form	
محمد	FDF4 0645 + 062D + 0645 + 062F (Optional)	Arabic Ligature Mohammad Isolated Form	
صلعم	FDF5 0635 + 0644 + 0639 + 0645 (Optional)	Arabic Ligature Salam Isolated Form	
رسول	FDF6 0631 + 0633 + 0648 + 0644 (Optional)	Arabic Ligature Rasoul Isolated Form	
عليه	FDF7 0639 + 0644 + 064A + 0647 (Optional)	Arabic Ligature Alayhe Isolated Form	
وسلم	FDF8 0648 + 0633 + 0644 + 0645 (Optional)	Arabic Ligature Wasallam Isolated Form	
صلى	FDF9 0635 + 0644 + 0649 (Optional)	Arabic Ligature Salla Isolated Form	
صلى الله عليه وسلم	FDF9 0635 + 0644 + 0649 + 0020 + 0627 + 0644 +	Arabic Ligature Sallahallahou Alayhe Wasallam	A3

	0644 + 0647 + 0020 + 0639 + 0644 + 064A + 0647 + 0020 + 0648 + 0633 + 0644 + 0645 (Optional)		
جل جلاله	FDFB 062C + 0644 + 0020 + 062C + 0644 + 0627 + 0644 + 0647 (Optional)	Arabic Ligature Jallajalalouhou	A1
صله	FDF0 0635 + 0644 + 06D2 (Optional)	Arabic Ligature Salla Used as Koranic Sign Isolated Form	
.	0651	Arabic Shadda	4F
	200C	Zero Width Non-joiner	41 (Hs)
م	0610 (Proposed)	Arabic Sign Sallaallahou Alayhe Wasallam	A4
ع	0611 (Proposed)	Arabic Sign Alayhe Wasallam	A5
رح	0612 (Proposed)	Arabic Sign Rahmatullah Alayhe	A7
رح	0613 (Proposed)	Arabic Sign Radi Allahou anhu	A6
ا	0614 (Proposed)	Arabic Sign Takhallus	A8
	0657 (Proposed)	Arabic Inverted Damma	46
ص	0658 (Proposed)	Arabic Noon-Ghunna	4E
ع	060F (Proposed)	Arabic Sign Misra	A9
ط	0603 (Proposed)	Arabic Sign Safha	AB
ط	0659 (Proposed)	Arabic Small high Tah	4C
بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ	FDFD (Proposed)	Arabic Ligature Bismillah Arahman Araheem	A2
ا	0600 (Proposed)	Arabic Number Sign	AC
خ	0601 (Proposed)	Arabic Year Sign	AD
ع	0602 (Proposed)	Arabic Footnote Marker	AA

	060E (Proposed)	Arabic Date Separator	
ﻻ	FEFB	Arabic Ligature Lam with Alef Isolated Form	AF