



**PAN
Localization**

**Survey of Language Computing in Asia
2005**

Sarmad Hussain
Nadir Durrani
Sana Gul

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences



www.nu.edu.pk

IDRC  CRDI

Canada

www.idrc.ca

Published by

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
Lahore, Pakistan

Copyrights © International Development Research Center, Canada

Printed by Walayatsons, Pakistan

ISBN: 969-8961-00-3

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Ottawa, Canada, administered through the Centre for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences (NUCES), Pakistan.

Urdu

Urdu belongs to the Indo-European language family, has influences from Persian and Arabic and is closely related to Hindi. About 104 million people speak Urdu as its first or second language across the globe. Urdu is the national language of Pakistan and is also widely spoken in Afghanistan, Bahrain, Bangladesh, Botswana, Fiji, Germany, Guyana, India, Malawi, Mauritius, Nepal, Norway, Oman, Qatar, Saudi Arabia, South Africa, Thailand, UAE, United Kingdom and Zambia [1, 2]. Figure 1 shows the language tree for Urdu.

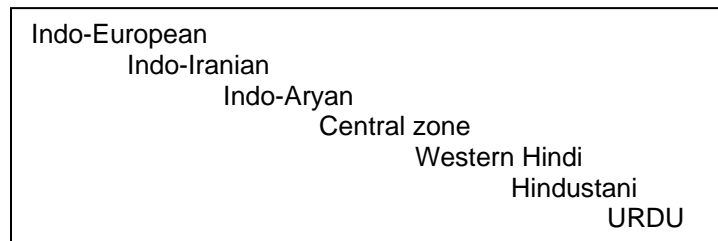


Figure 1: Language Family Tree for Urdu [1]

Perso-Arabic script written in Nastalique style is widely used for Urdu orthography [3].

Character set and Encoding

Vendors had developed multiple encoding schemes for Urdu language in 1980's. These encodings did not allow users to exchange data, and therefore, efforts were taken to standardize Urdu encoding in 1998. These efforts resulted in the formation of Urdu Zabta Takhti (UZT 1.01), the national standard for Pakistan in 2000 (see [4] for details). This standard is given in Figure 2 below. Though the standard is not widely utilized, it has been used to update the Unicode support for Urdu.

Unicode provides an international standard for Urdu character set encoding. Arabic script block from 0600 to 06FF and ligatures FDFx are used. This standard was updated in Unicode 4.0 after a gap analysis with UZT [5].

Though Unicode is increasingly being used, especially for web development, currently, most of the publishing and other word processing is still done using the ad hoc encoding based on Inpage Urdu word processing software. Other ad hoc encodings are not widely used anymore.

	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0			Sp	.	@		ژ	م			ل	o	[
1			!	\	HS	آ	ز	ن			ع		\			
2			"	۲	بھرتہ اضافہ	آ	ژ	ن			ب]			
3			#	۳	کڑھ اضافہ	ب	و	س			و		U			
4			C	۴	۱	پ	ش	ذ			۴		{			
5			۵	۵	۱	ن	ص	و			۶					
6			&	۶	۴	ث	ض	ق			۶		}			
7			'	<	۵	ن	ط	و			ع		D			
8			(۸	۱	ج	ظ	ی			۸		a			
9)	۹	۱	ع	ع	ع			۹					
a			*	:	۱	ح	غ	۶			۶					
b			+		۱	خ	ف	۶			۶					
c			,	<	ط	د	ق	۱			۶					
d			.	=	۸	ڈ	ک	۱			۸					
e			D	>	۷	ذ	ک	۱			۷					→
f			D	?	۱	ر	ل				۵					

Figure 1: Urdu Zabta Takhti (Urdu Code Plate) ver 1.01

Abbreviations

Sp: space, Cr: currency, Dc: decimal, Dv: division,
HS: hard space, Us: underscore, Ds: dash,
→: code plate switching

Legend




	Control area (not to be used)
	Reserved area (for future use by the standard)
	Vendor area

Figure 2: Urdu Zabta Takhti (UZT 1.01) [4]

Fonts and Rendering

Microsoft Platform

Microsoft Windows XP provides rendering support for Urdu. It also ship two fonts, Tahoma and Sans Serif, which provide Urdu character support in Naskh style but not the preferred Nastalique style, latter being very complex [6]. Nafees Nastalique [7] is developed in Urdu style by Centre for Research in Urdu Language Processing (CRULP) using Open Type font format. Figure 3 shows the results for Tahoma and Nafees Nastalique as typed on Windows XP.

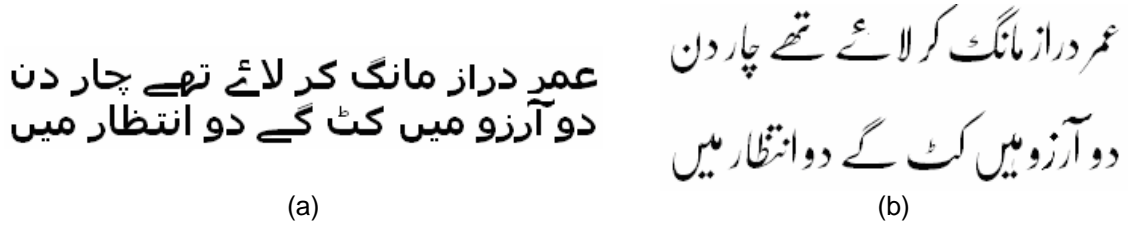


Figure 3: (a) Tahoma and (b) Nafees Nastalique Fonts on Microsoft Platform

One of the most unique fonts for Urdu Nastalique, and still the most widely used, is a ligature based font, which stores about 32,000 ligatures. It is not based on any standard encoding and can only be used by its parent application, Inpage word processor.

Linux Platform

Urdu is not rendered properly on Linux, as Linux cannot render complex font formats like Open Type font. However, simpler four-shaped Naskh style fonts can be rendered as they do not use positioning and substitution tables of Open Type fonts. Complex rule-based fonts like Nafees Nastaleeq and Nafees Naskh are not rendered properly. Figure 4 displays results of rendering three fonts in G-Edit, which uses Pango rendering engine.

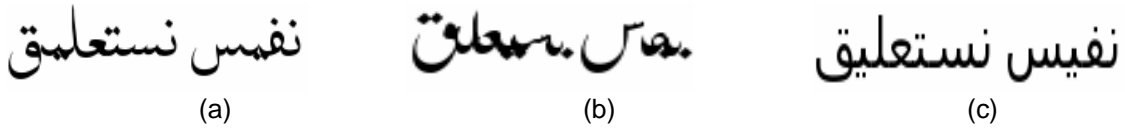


Figure 4 (a) Nafees Naskh, (b) Nafees Nastalique and (c) Simple Four-Shaped Fonts Rendered on G-Edit (GNOME platform)

Pango rendering engine gives better results than KDE. Latter displays boxes when an Open Type font is used on K-Edit, as shown in Figure 5.

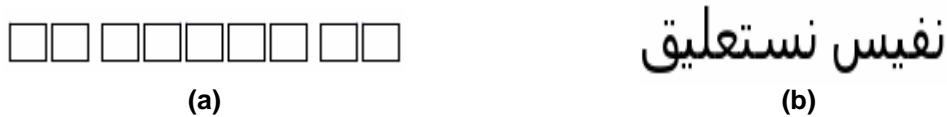


Figure 5: (a) Nafees Nastaleeq and (b) Simple Four-Shaped Fonts on KDE

Open Office renders Urdu text using its own rendering engine. It extracts and consequently displays only the True Type features from Open Type font file and does not use other rules in the font to connect letters, as shown for Open Office 1.1 in Figure 6 below.

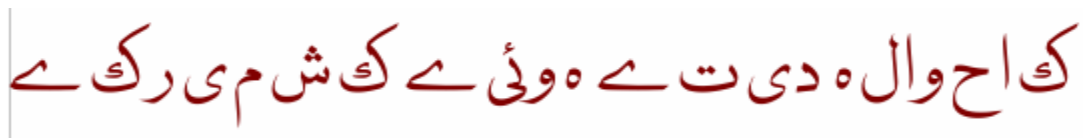


Figure 6: Urdu Text Written in Nafees Naskh in Open Office Version 1.1.0

Results of rendering Urdu Open Type fonts on Mozilla are similar to those obtained through KDE's rendering engine. However, Mozilla, instead of displaying boxes when rendering Open Type fonts (as in KDE), displays the same text in a simple four-shaped Naskh font.

Keyboard Layout

Work is underway to standardize Urdu keyboard layout in Pakistan. Similar to Urdu character set encoding formats, different software vendors have implemented different layouts for sequencing Urdu characters on the keyboard e.g. Katib, Rakim, Inpage, IBM, NLA and Microsoft keyboard layout [8, 9]. The keyboard layouts provided by Inpage are the most commonly used, and include Inpage and Phonetic layouts.

Microsoft Platform

Microsoft Windows XP provides support for a built-in Urdu keyboard, shown in Figure 7. This is based on the keyboard layout given by National Language Authority [9]. However, this layout is not widely used.



(a)



(b)

Figure 7: (a) Normal and (b) Shift Versions of Microsoft Urdu Keyboard

A phonetic keyboard extended from Inpage keyboard to incorporate Unicode character set for Urdu is distributed by Center for Research in Urdu Language Processing (CRULP) for Microsoft platform [10], and is shown in Figure 8.

Phonetic Keyboard Layout

Base version

(64D) ں	(6F1) ا	(6F2) ا	(6F3) ا	(6F4) ا	(6F5) ا	(6F6) ا	(6F7) ا	(6F8) ا	(6F9) ا	(6F0) .	(623) ا	(624) و	(602) ے
(642) ق	(648) و	(639) ع	(631) ر	(62A) ت	(602) ے	(621) ع	(6CC) ی	(6C1) ہ	(67E) پ	(05D) [(05B)]		
(627) ا	(633) س	(62F) ج	(641) ف	(6AF) گ	(6BE) ہ	(62C) ج	(6A9) ک	(644) ل	(61B) ہ	(027) '			
(632) ز	(634) ش	(686) ی	(637) ط	(628) ب	(646) ل	(645) م	(60C) ے	(6D4) ے		(615) ط			

Shift version

(64B) ں	(021) !	(600) م	(02F) /	(626) ع	(60F) ط	(610) ں	(654) م	(64C) ں	(029) ((028))	(651) ں	(622) آ	(614) ں
(6E1) ں	(PDF) ہ	(611) ں	(691) ر	(679) ت	(601) ع	(657) ں	(670) ں	(6C3) ج	(64F) ں	(603) ں	(60E) ں		
(653) ں	(635) ص	(688) ڈ	(656) ں	(63A) غ	(62D) ح	(636) ض	(62E) خ	(612) ں	(03A) :	(022) "			
(630) ڈ	(698) ث	(628) ث	(638) ط	(613) ں	(6BA) ل	(658) ل	(650) ں	(64E) ں	(61F) ؟				

Figure 8: Urdu Phonetic Keyboard Layout by CRULP Extended From InPage Phonetic Keyboard [10]

Linux Platform

There is no in-built keyboard for Urdu in Red Hat or any other Linux distributions. However, Urdu keyboard has been added in the Urdu Distribution developed by CRULP. It is based on phonetic layout shown in Figure 8 [10]. Another Urdu keyboard available for Linux platform has been developed by Sindhi Computing group [11].

Collation Sequence

Urdu collation sequence has recently been standardized and published by National Language Authority of Pakistan [12]. The collation sequence for characters and diacritics is shown in Figure 9. This is not yet supported on any platform.

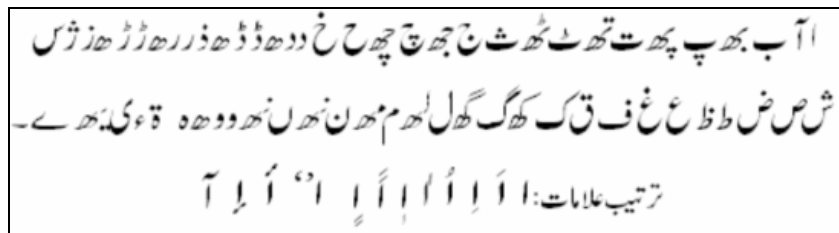


Figure 9: Standard Collation Sequence for Urdu [12]

Locale

Urdu locale has not been standardized nationally in Pakistan or India, but there has been some work towards its definition internationally. Urdu locale (ur_PK) is partially defined in CLDR 1.3.

Microsoft Platform

Microsoft Windows XP provides support for Urdu locale. If system locale is switched to Urdu, changes may be monitored in the date, time, and currency symbol for all applications of Microsoft. This is shown in Figure 10.

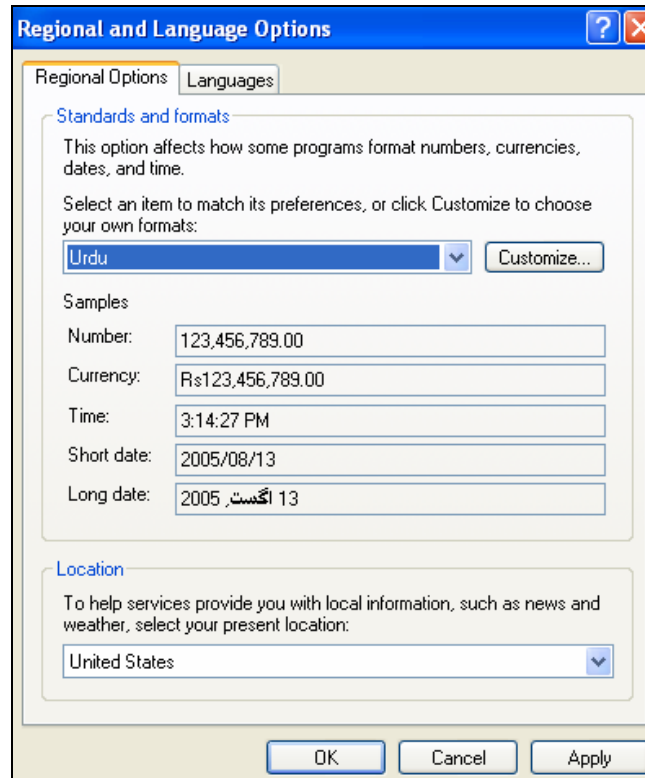


Figure 10: Urdu Locale on Microsoft Platform

Microsoft also displays localized information for Urdu language by enabling a thread locale for Urdu. If the user locale is set to Urdu, it automatically retrieves Urdu version of a multilingual resource file even if no explicit settings are made. For example, accessing Google with Urdu locale will automatically retrieve Urdu version of Google as shown in Figure 11.



Figure 11: Urdu Version of a Multilingual Website www.Google.com with Urdu Locale

Linux Platform

Locale for Urdu is defined in Red Hat Linux version 9 and above. Though incomplete, locale definitions for Urdu time, month names and days of week etc. are displayed, as shown in Figure 12.



Figure 12: Urdu Calendar in Urdu Distribution for Linux

Interface Terminology Translation

Standard translation for interface has recently been published by National Language Authority of Pakistan, after the translation work for Microsoft [12]. This terminology has been realized on Microsoft platform. Urdu LIP for Microsoft is due to be released in 2006. Partial interface terminology translation has been performed in the Urdu Linux distribution by CRULP. In this distribution, KDE base files, desktop files and K-Office suite have been partially translated as shown in Figure 13.



Figure 13: Localized Start-Up Menu for KDE in Urdu Distribution by CRULP

Status of Advanced Applications

CRULP has been working on developing advanced solutions for Urdu. To date, it has developed the following applications [10].

- Urdu spell checker
- Prototype Hindi to Urdu transliteration engine
- Prototype Urdu Naskh and Nastalique optical character recognition system
- Prototype Urdu speech recognition system
- Urdu morphological parser
- Urdu Corpus
- Urdu Lexicon
- English to Urdu machine translation system
- Urdu text-to-speech system
- Website and email reader (based on Urdu TTS)
- Website and email translator (based on English to Urdu MT)

The Urdu lexicon, Urdu text-to-speech system and English to Urdu machine translation system are being developed through Urdu Localization project, an initiative of E-Government Directorate of Ministry of IT, Government of Pakistan.

Corpus has also been developed by EMILLE project [13].

References

- [1] http://www.ethnologue.com/show_language.asp?code=urd
- [2] http://en.wikipedia.org/wiki/Urdu_language
- [3] <http://www.omniglot.com/writing/urdu.htm>
- [4] <http://www.crup.org/Publication/n2413-2.pdf>
- [5] <http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2413-1.pdf>
- [6] Hussain, S. "Complexity of Asian Writing Script: A case study of Nafees Nastaleeq." Proceedings of SCALLA, Kathmandu, Nepal. 2003.
- [7] <http://www.crup.org/nafeesNastaleeq.html>
- [8] Aziz, T. "Urdu type Machine kay kaleedi Takhtay." Muqtadra Qaumi Zaban, Islamabad, Pakistan, 1987.
- [9] http://www.nla.gov.pk/keyboard_files/slide0001.htm
- [10] www.crup.org
- [11] <http://groups.msn.com/SindhiComputing>
- [12] www.nla.gov.pk
- [13] <http://www.emille.lancs.ac.uk/home.htm>