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

www.idrc.ca

Published by

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

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.

Hindi

The word “Hindi” is derived from Sanskrit word ‘Hindva’ meaning 'language of Hind’. About 180 million people speak Hindi as their first language and many more across the globe use it as a second language. Hindi is the national language of India and is also widely spoken in Bangladesh, Fiji, Indonesia, Malaysia, Mauritius, Nepal, South Africa, Uganda and Yemen. It is the third most spoken language and comes after Chinese and English. Hindi belongs to the Indo-European language family and has influences from Persian and Arabic [1]. Its formal vocabulary is derived from Sanskrit and Prakrit.

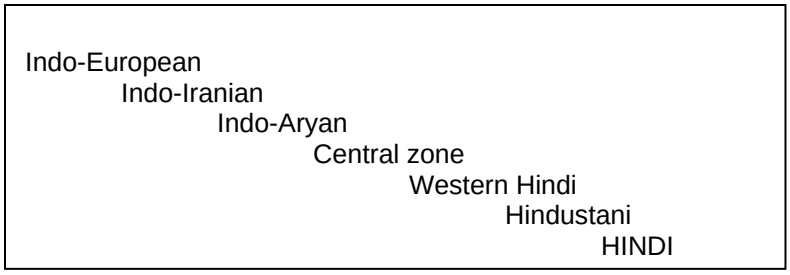


Figure 1: Language Family Tree for Hindi [1]

Hindi is written with Devanagari script, which was derived from Brahmi script [2].

Character Set and Encoding

ISCII (IS 13194:1991, earlier IS 13194:1988) is the national standard for Devanagari character set encoding, based on earlier standard IS 10402:1982 [3]. ISCII is a standard for Devanagari script and may be used for other languages. It is widely used in India. The standard contains ASCII in lower 128 slots and Devanagari alphabet superset in upper 128 slots and therefore it is a single byte standard. Though it is primarily an encoding standard (and sorting is usually not catered directly in such standards, e.g. see Collation section below), the standard was devised to do some implicit sorting directly on encoding. Variations of ISCII include PC-ISCII and language specific ISCII charts (see [4] for some details). ISCII standard is shown in Figure 2 below. Official standard publication is available at [12].

Unicode provides an international standard for Devanagari character set encoding based on IS 13194:1988 from 0900 till 097F (and therefore is not exactly equivalent to IS 13194:1991; also see [5, 14]). This may be used for Hindi and other Devanagari script based languages, including Marathi, Sanskrit, Prakrit, Sindhi, etc.

There are other encodings which have been used by vendors, in addition to those discussed. However, they are not as prevalent anymore. There are also encoding converters available which can convert among various encodings and platforms (e.g. [6, 7]).

Hex	Dec	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
Hex	Dec	0	16	32	48	64	80	96	112	128	144	160	176	192	208	224	240
0	0	NUL	DLE	SP	0	@	P	`	p				ओ	ढ	ऱ	े	EXT
1	1	SOH	DC1	!	1	A	Q	a	q			ॅ	औ	ण	ल	े	०
2	2	STX	DC2	"	2	B	R	b	r			ं	ऑ	त	ळ	ै	१
3	3	ETX	DC3	#	3	C	S	c	s			ः	क	थ	ढ	ॅ	२
4	4	EOT	DC4	$	4	D	T	d	t			अ	ख	द	व	ो	३
5	5	ENQ	NAK	%	5	E	U	e	u			आ	ग	ध	श	ो	४
6	6	ACK	SYN	&	6	F	V	f	v			इ	घ	न	ष	ौ	५
7	7	BEL	ETB	'	7	G	W	g	w			ई	ङ	ऩ	स	ॉ	६
8	8	BS	CAN	(	8	H	X	h	x			उ	च	प	ह	्	७
9	9	HT	EM	)	9	I	Y	i	y			ऊ	छ	फ	INV	़	८
A	10	LF	SUB	*	:	J	Z	j	z			ऋ	ज	ब	ा	।	९
B	11	VT	ESC	+	;	K	[	k	{			ऐ	झ	भ	ि		
C	12	FF	FS	,	<	L	\	l	|		ए	ञ	म	ी		
D	13	CR	GS	-	=	M	]	m	}			ऐ	ट	य	ु		
E	14	SO	RS	.	>	N	^	n	~			ऍ	ठ	य़	ू		
F	15	SI	US	/	?	O	_	o	DEL			ओ	ड	र	ृ	ATR	

Figure 2: ISCII Code Chart IS 13194:1991 [3]

Fonts and Rendering

There are many fonts available to write Hindi on different platforms. Some are listed below.

Microsoft Platform

Windows provides Mangal font, which has been developed by CDAC, India, and other fonts for Hindi. Windows uses Uniscribe as the rendering engine, which supports rendering of Open Type fonts for Devanagari script. Results of Hindi fonts rendered on Microsoft are shown in Figure 3 below.

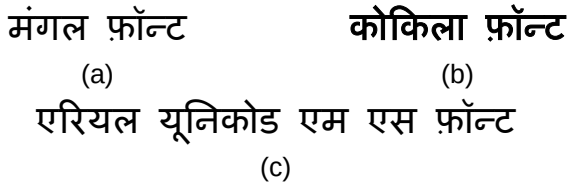
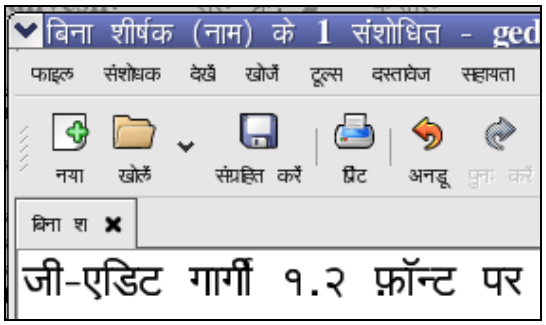


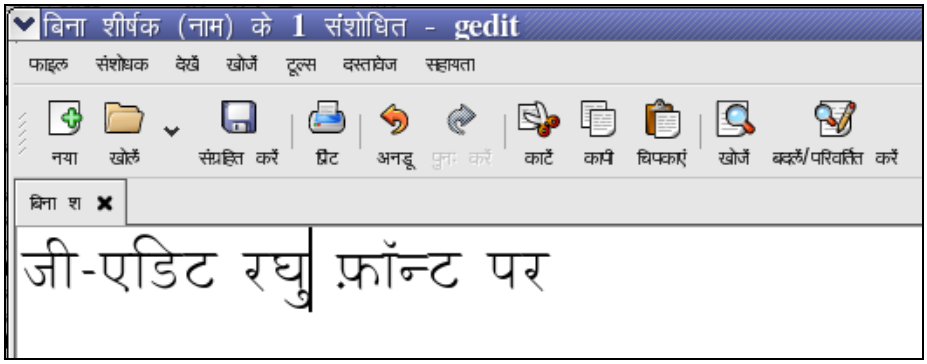
Figure 3: (a) Mangal, (b) Kokila, and (c) Arial Fonts on Microsoft Office 2003

Linux Platform

In Red Hat Fedora Core 3 Linux, system level support for Hindi fonts is now available. Red Hat has bundled Lohit series of Indic fonts with its distribution. Raghu and Gargi fonts have also been available for some time. Presented below are examples of a Hindi on G-Edit using Gargi and Raghu fonts.



(a)



(b)

Figure 4: (a) Gargi and (b) Raghu Fonts on G-Edit in GNOME

Pango rendering engine gives better results as it is more aware of substitution and positioning rules but these needs to be improved further. KDE does not provide any support for rendering Open Type fonts. Below is the demonstration of typing Hindi text using Gargi, on K-edit that runs Qt.

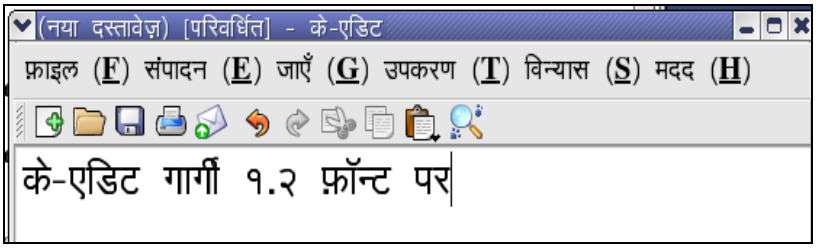


Figure 5: Gargi Font on K-Edit in KDE

Rendering engine of Open Office extracts and consequently displays only the True Type features from the font file. Rendering through Open Office 1.1.1 is shown in Figure 6.

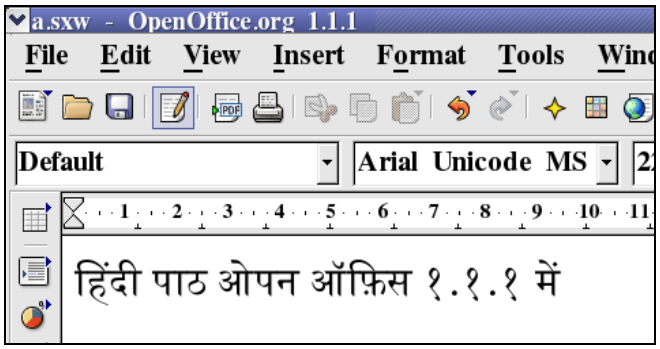


Figure 6: Hindi Rendering in Open Office Version 1.1.1

A Pango enabled Mozilla renders Hindi reasonably well. Where required, India Linux project has updated the rendering engine to work for Devanagari script [8].

Keyboard

As for Hindi character set encoding formats, different software vendors have implemented many different keyboard layouts e.g. Godrej, Ramington, Phonetic, Shusha, and Traditional keyboard layout. Inscript is the standard Hindi keyboard layout and is the most commonly used [9, 11]. It is shown in Figure 7 below.

Figure 7: Inscript Keyboard for Hindi [11]

Microsoft Platform

Microsoft provides support for a built-in Hindi keyboard in Traditional layout, shown in Figure 8. Once this keyboard is enabled it supports Hindi typing on all Microsoft applications. Further details are provided in [10].

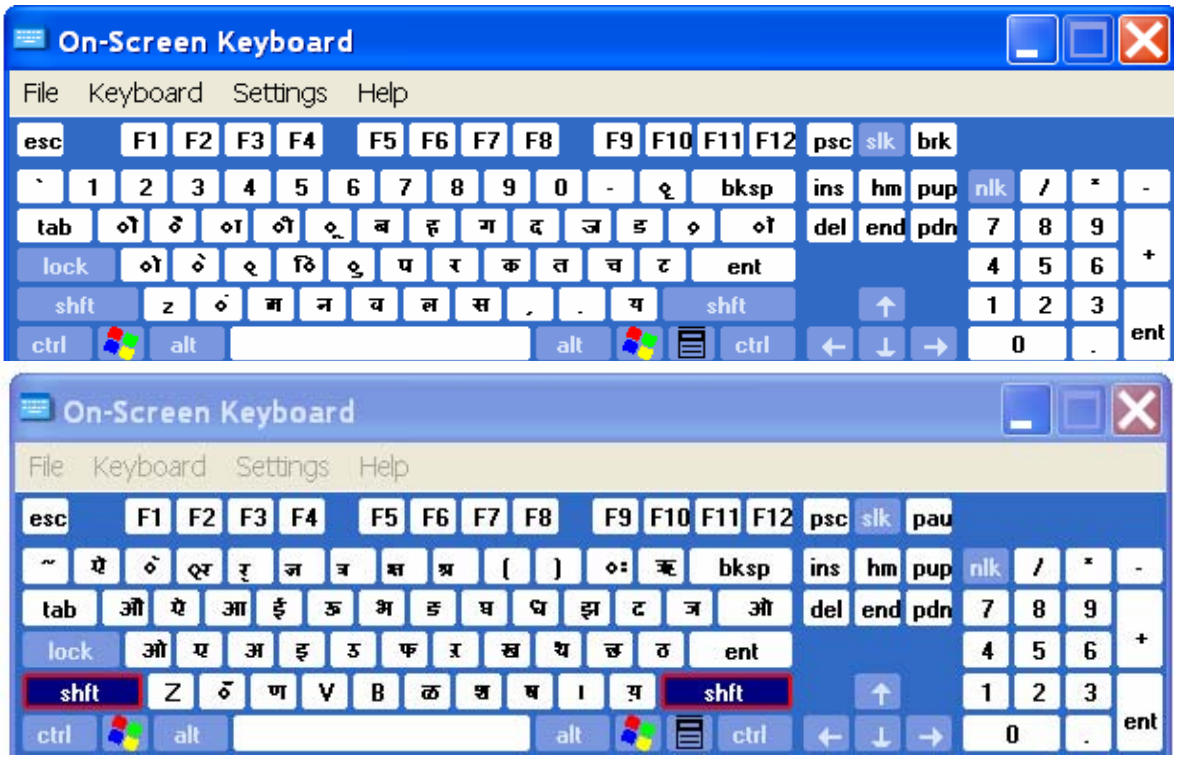


Figure 8: Traditional Keyboard Layout for Microsoft Platform

Linux Platform

There is built in support for Hindi keyboard in Red Hat Fedora Core 3 Linux. There are many Hindi keyboards available that include Inscript, Devrom and Bolnagri, which provide support for all Unicode applications running on Linux, including Open Office, Mozilla, etc. A live Linux Morphix-based CD has been developed by Indlinux.org. This is a complete Linux distribution in Hindi, which also has many Hindi keyboards such as Phonetic, Bolnagri and Inscript packaged within it [8].

Collation

Work had been in progress to finalize a single collation sequence standard for Government of India. However, ambiguities in the linguistic sorting order of the Hindi character set has hampered this standardization [8]. The work is still in progress.

Microsoft Platform

A collation sequence for Hindi is supported on Microsoft platform [13]. This order gives satisfactory results for Hindi users.

Linux Platform

Collation on Linux platform is done using LC_COLLATE in the locale definition. The current hi_IN locale file does not have collation data included, so default sort order is used. As such this suffices for basic Hindi sorting, because Devanagari range in Unicode is based on ISCII-88 document, which does implicit sorting for Hindi.

Locale

No significant effort at national or regional level has been undertaken to standardize Hindi locale

(hi_IN), though it is included in CLDR 1.3. Some work has been done by CDAC in defining a locale for Microsoft to enable Hindi in Windows [15].

Microsoft Platform

Microsoft Windows XP does provide support for Hindi locale. If system locale is switched to Hindi, appropriate changes are observed in all application of Microsoft. An example of setting User locale to Hindi is shown for Microsoft Windows XP in Figure 9.

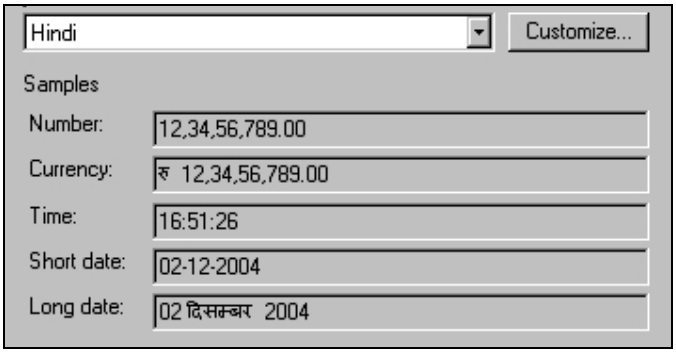


Figure 9: Hindi Locale Definition on Microsoft Platform

Linux Platform

Locale for Hindi is defined in Red Hat Linux version 9 and above. Though not complete, it still provides significant Hindi related information. Most Linux distributions support Hindi locale.

Open Office and Mozilla suites pick locale definitions from their underlying system locale definitions. Therefore as Hindi locale has been defined in Linux, it is also available in these applications. Figure 10 below shows calendar in Hindi.

Figure 10: Hindi Calendar on Linux

Interface Terminology Translation

Hindi Interface is available on both Microsoft and Linux platforms. Microsoft provides a completely localized version of Windows and Office [16], as illustrated in Figure 11.

Interface terminology translation has been performed in the Hindi Linux distribution developed by IndLinux. In this distribution, KDE base files and desktop files have been fully translated, and GNOME files have been partially translated. Localized interface is illustrated in Figure 12.

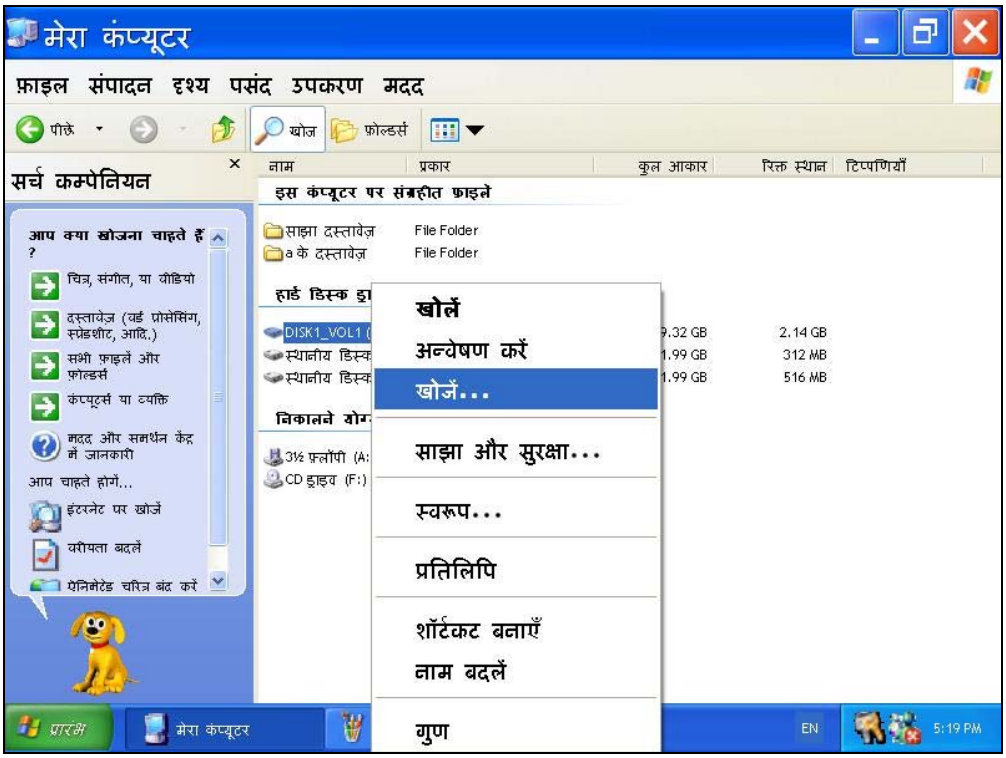


Figure 11: Localized Microsoft Windows XP

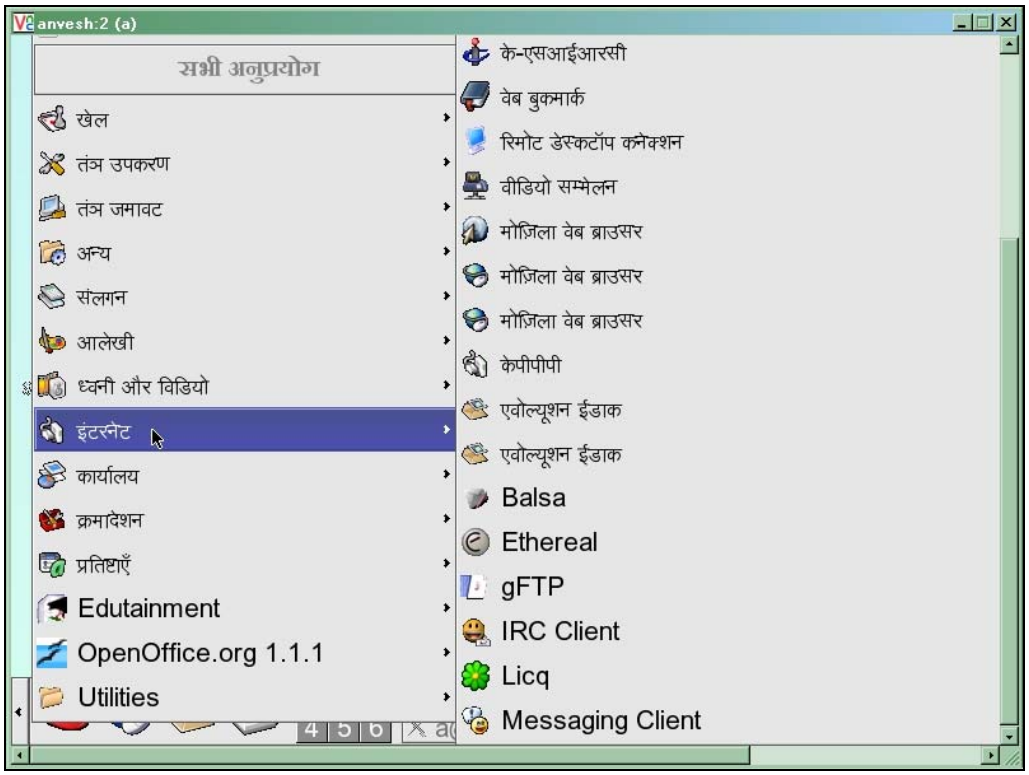


Figure 12: Localized Start-Up Menus for Hindi Linux Distribution

Status of Advanced Applications

Significant work is being done on Hindi language processing in Indian language technology centers at universities, CDAC and TDIL resource centers across India. Applications being developed include Hindi spell checkers, Hindi mono- and multi-lingual lexicons, Hindi text-to-speech systems and Hindi speech recognition systems. Work is also being done on Hindi machine translations systems with English and Indic languages, Hindi OCR systems and other related applications and resources including corpora, POS taggers, morphological analyzers, etc. Information about this work is available at [15, 17, 18, 19, 20, 21, 22]. Hindi WordNet has also been recently released [19]. Work is also in progress on online handwriting recognition.

References

[1] www.ethnologue.com
[2] www.omniglot.com
[3] http://tdil.mit.gov.in/standards.htm
[4] http://homepages.cwi.nl/~dik/english/codes/indic.html
[5] http://acharya.iitm.ac.in/multi_sys/exist_codes.html
[6] http://www.iiit.net/ltrc/FC-1.0/fc.html
[7] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=EncCnvtrs
[8] http://indlinux.org/
[9] http://tdil.mit.gov.in/keyoverlay.htm
[10] http://www.bhashaindia.com/Developers/IndianLang/TypingDnagari/dnpages.htm?lang=en
[11] http://en.wikipedia.org/wiki/Devanagari
[12] http://varamozhi.sourceforge.net/iscii91.pdf
[13] http://bhashaindia.com/ForumV2/shwmessage.aspx?ForumID=5&MessageID=821
[14] http://tdil.mit.gov.in/pchangeuni.htm
[15] http://www.cdacindia.com/
[16] http://www.bhashaindia.com/downloadsV2/Category.aspx?ID=2
[17] http://www.cse.iitk.ac.in/users/rmk/proj/proj.html
[18] http://ltrc.iiit.net/
[19] http://www.cfilt.iitb.ac.in/
[20] http://www.ciil.org
[21] http://speech.cs.iitm.ernet.in/
[22] http://www.isical.ac.in/~rc_bangla/activities.html