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

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.

Bengali

Bengali (ethnonym: Bangla) is an Indo Aryan language spoken by about 170 million people across the world out of which about 100 million Bengali speaking population resides in Bangladesh. Populations of Bengali speakers are also found in India, Nepal and Singapore [1]. Bengali is the national language of Bangladesh and it is also the state language of the Indian state of West Bengal.

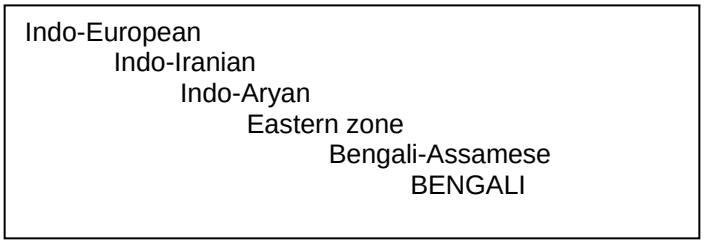


Figure 1: Language Family Tree for Bengali [1]

Bengali is written using Bengali script which is derived from Brahmi. Bengali script is also closely related to the Devnagari script [2].

Character Set and Encoding

Bengali character set encoding is included in Unicode 4.0. Unicode block 0980-09FF is the standard encoding used for Bengali script in computers. Bengali code block is also available in ISCII-91, for which the code page identifier for Bengali is 57003 [3, 4, 19]. This is shown in Figure 2.

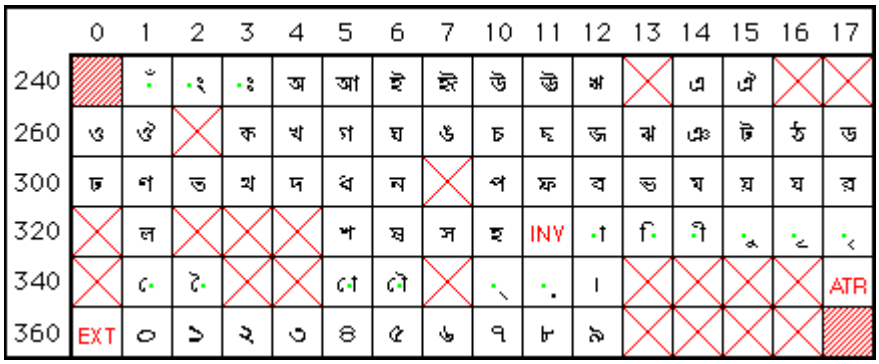


Figure 2: Language Family Tree for Bengali [19]

A Bangladesh national standard for Bengali text encoding, BSD 1520, has also been recommended by Bangladesh Standards and Testing Institute (BSTI) in 1995. However, this encoding is used infrequently. This standard was revised in 2000 and eventually Unicode was also formally adopted as national standard. A detailed overview is given in [22]. Most of the recent software development being done for Bengali is based on Unicode.

Fonts and Rendering

Many Bengali fonts are available free of cost and through private vendors (e.g. [5, 6, 7, 8]).

Microsoft Platform

Microsoft does not have in-built Bengali fonts. However many Bengali Unicode based fonts are available which can be appropriately rendered on Microsoft platform [5, 6, 7, 8]. Figure 3 presents results of using some of these fonts.

Figure 3: Unicode Based Fonts on Microsoft Platform [8]

Open Type font support for Bengali was first included in Service Pack 2 update for Microsoft Windows XP. This service pack installs an upgraded version of the Uniscribe engine as well as an on-screen keyboard based on Inscript layout for Bengali (details given later).

Linux Platform

Bengali fonts are shipped with all the Bengali Linux distributions. Ankur live CD includes Bengali fonts Aakash and Luxi. These are rendered adequately on all Linux platforms as shown in Figure 4 below.

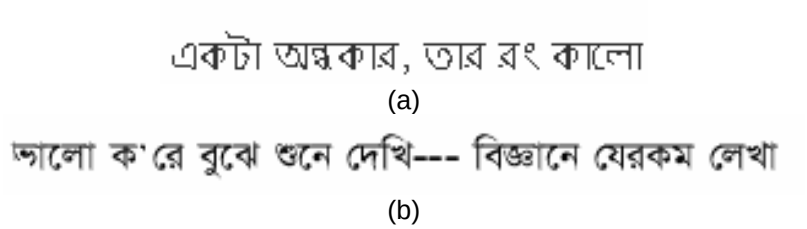


Figure 4: Bengali Fonts Rendered in (a) G-Edit (GNOME) and (b) Open Office 2.0

Keyboard

Bijoy keyboard layout [9] (see Figure 5) has been the de-facto standard since late 1980's, spreading beyond the border to the Indian state of West Bengal. While there have been recent layouts developed, all of these have been variations of Bijoy, e.g. UniBijoy. Another standard is based on Inscript keyboard layout, developed and standardized by Government of India.

Figure 5: Bijoy Bengali Keyboard Layout [9]

Bangladesh Computer Council published Bangladesh Jatiyo (national) keyboard standard in 2003, based on a phonetic layout. However it is reported to be used rarely [10].

Microsoft Platform

Bengali keyboard is available with Microsoft Windows XP. It is based on Inscript keyboard layout as shown in Figure 6.

(a)

(b)

Figure 6: Inscript Based Bengali On-Screen Keyboard, (a) Normal, and (b) Shift Version

Linux Platform

Linux platform supports two pre-defined Bengali keyboard layouts. These are extended Bengali and Probhat keyboard layouts. The phonetic based Probhat has been modified to deal with the latest Unicode 4.1 changes and has been incorporated into the Ankur live CD through Bangla Linux project [11]. Figure 7 shows Probhat phonetic keyboard layout in Linux.

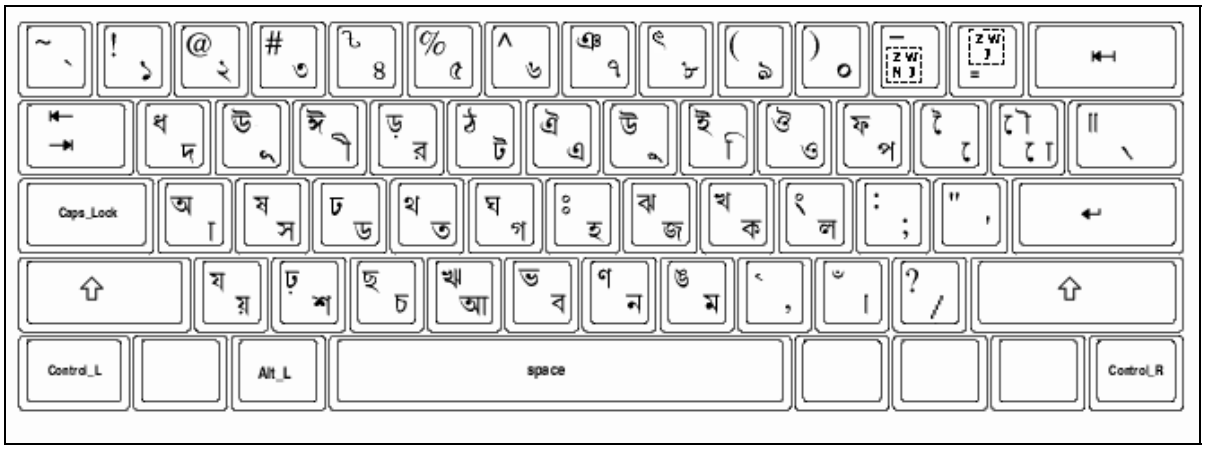


Figure 7: Probhat Phonetic Keyboard Layout [11]

Collation

Collation sequence for Bengali is yet to be standardized. However, Bangla Academy publishes a dictionary which defines the official collation sequence for Bangladesh. The Bangla Academy dictionary is also widely used in the Indian state of West Bengal.

Bengali sorting is not supported by Microsoft or Linux platforms.

Locale

Bengali Locale is defined in CLDR 1.3. There are two locales bn_BD and bn_IN for Bangladesh and India respectively.

Microsoft Platform

Microsoft provides partial support for Bengali locale in Windows XP. When the system locale is switched to Bengali, date, time and numbers, etc. change to Bengali. Figure 8 below shows Bengali settings for India for number and long date formats, and currency symbol in Windows XP.

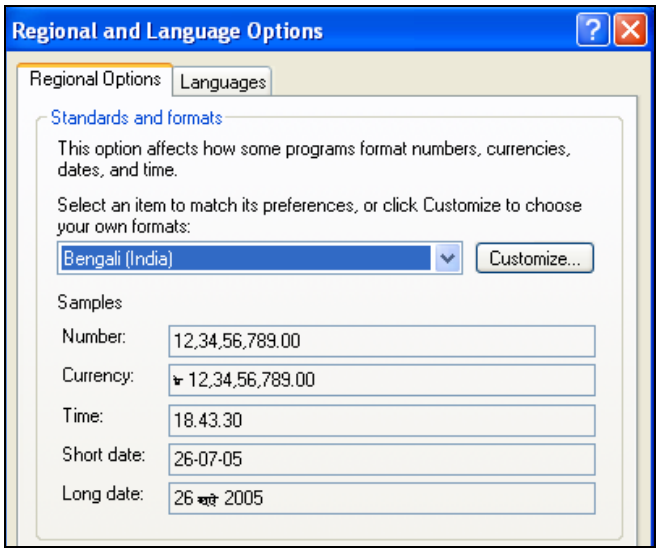


Figure 8: Bengali Locale on Windows Platform

Linux Platform

Bengali locale has also been defined on the Linux platform. The live Ankur CD developed through the Bangla Ankur Project includes a Bengali locale for Bangladesh [12]. Figure 9 shows the locale setting of Bengali Linux distribution.

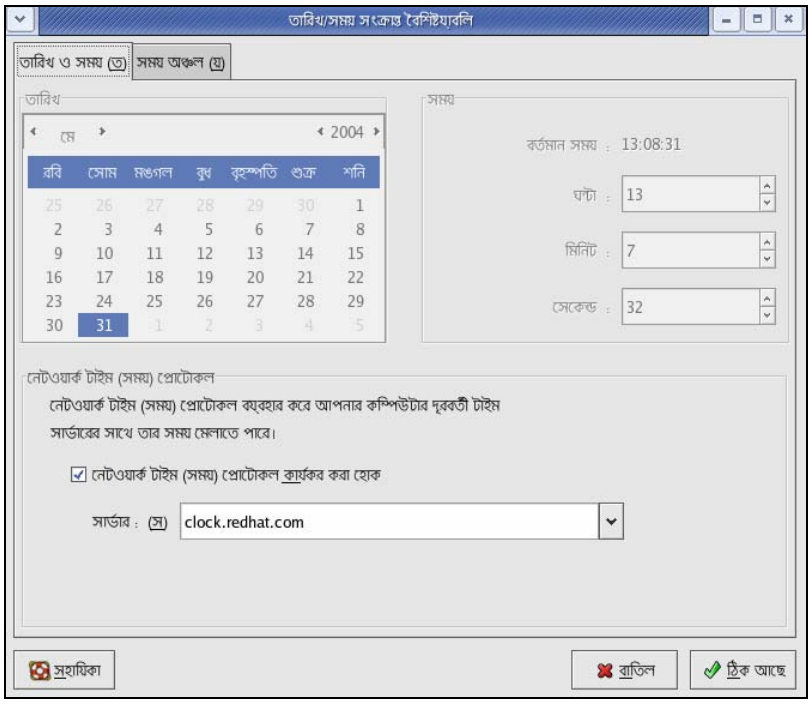


Figure 9: Bengali Locale on Ankur Live CD

Locale data has also been added for Bengali (bn_BD) in OpenOffice.org 2.0 Beta 2 [13] (see Figure 10 below).

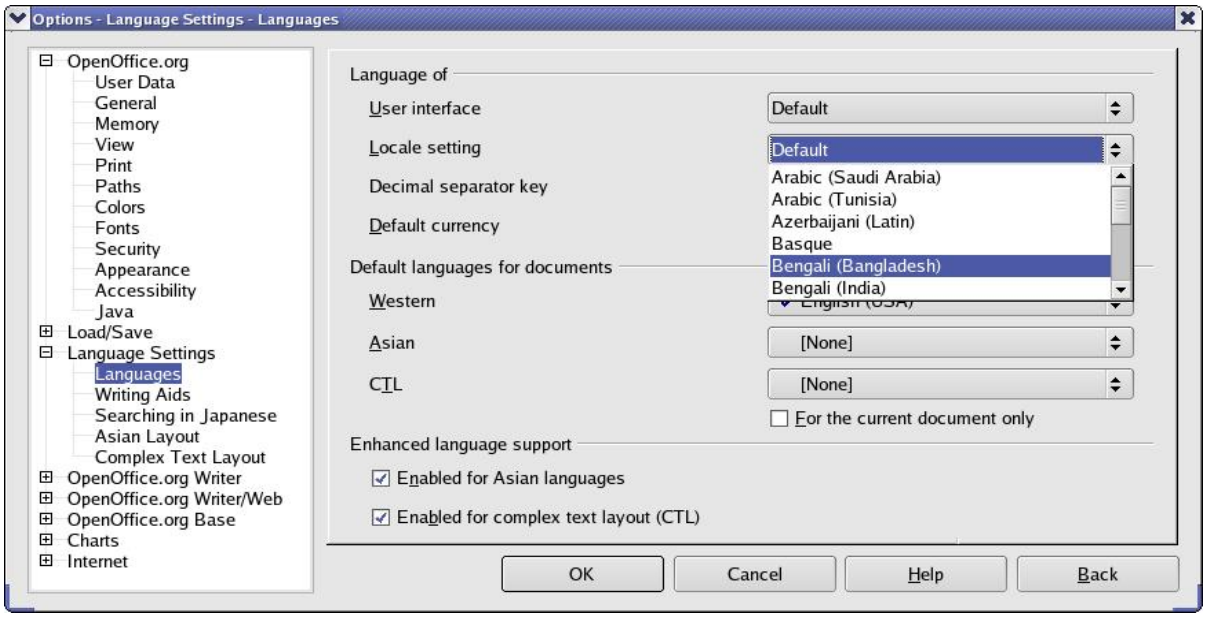


Figure 10: Open Office Language Settings

Interface Terminology Translation

The Bangla Linux developers have formulated a style guide for Bengali translations to be used for Linux. The complete list of translated strings in Bengali is available at Bangla Linux project website as well [14].

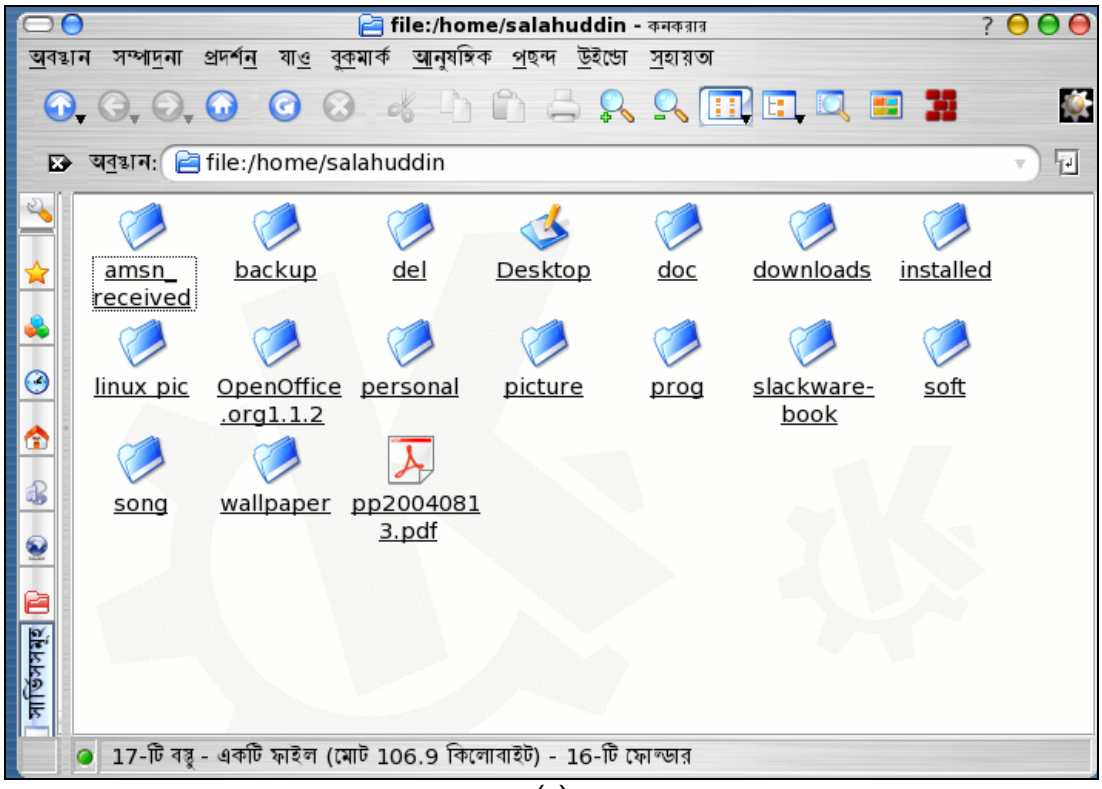
Microsoft Platform

Recently Microsoft has signed a Memorandum of Understanding with the Government of West Bengal in India to develop localized computing applications by providing LIP in Bengali for Windows and Office 2003. Bengali LIP for Microsoft Windows and Office 2003 editions is now under development. This support will offer a complete Bengali user experience.

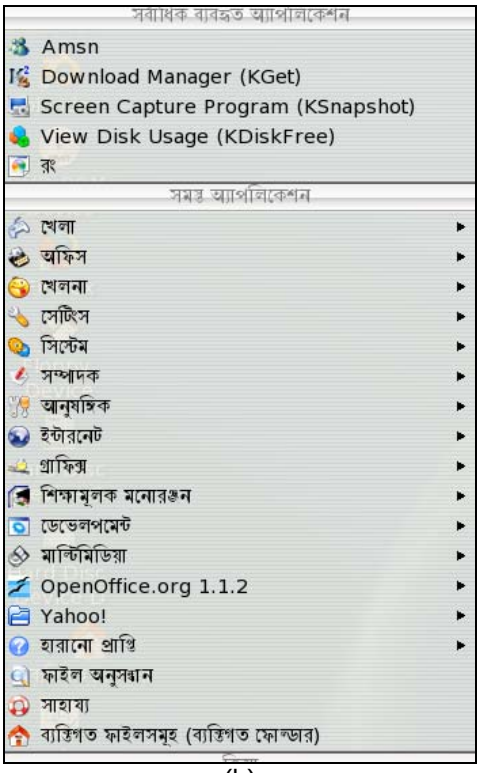
Linux Platform

Red Hat has launched a beta version of Bengali Linux distribution. Red Hat Enterprise Linux Bengali version includes applications such as Office suite with a word processor, spread sheet, presentation tool, as well as a web browser and an e-mail client [15].

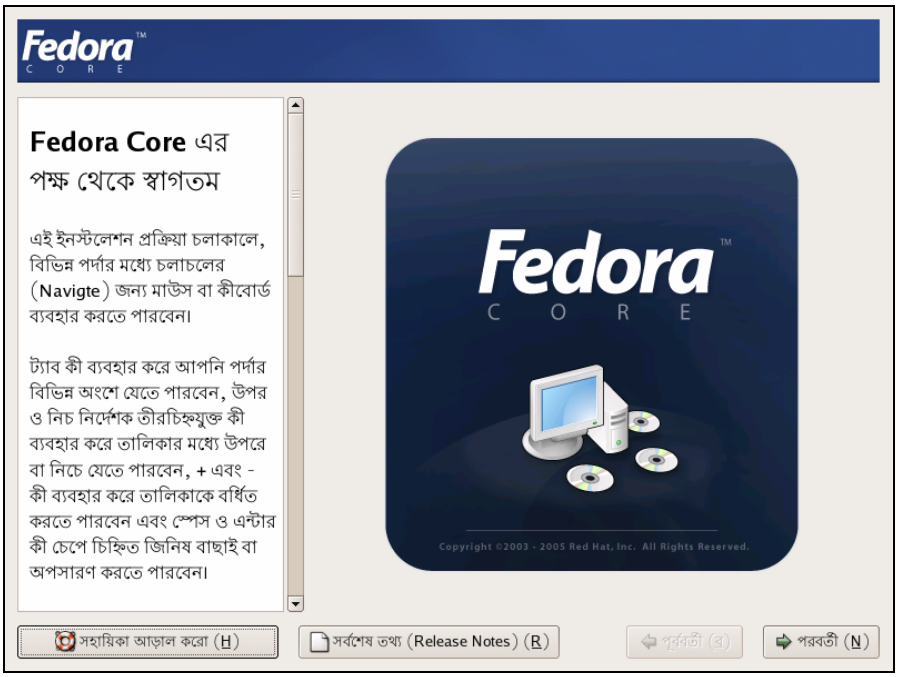
Ankur Bangla project is working on translations for KDE, GNOME, Fedora Core 4, SUSE and Mandrake [16]. Translations for GNOME 2.6 have been completed while 35% of the KDE translations have also been done. Ankur project is also working on developing Bengali translation and localization of Open Office and Mozilla. Figure 11 presents the localized Konqueror interface, partially localized KDE start up menu and localized Fedora Core 4.


file:/home/salahuddin - কনকরার
অবস্থান সম্পাদনা প্রদর্শন যাও বুকমার্ক আনুষঙ্গিক পছন্দ উইন্ডো সহায়তা
অবস্থান: file:/home/salahuddin
amsn_ received
backup
del
Desktop
doc
downloads
installed
linux pic
OpenOffice .org1.1.2
personal
picture
prog
slackware- book
soft
song
wallpaper
pp2004081 3.pdf
সার্ভিসসমূহ
17-টি বস্তু - একটি ফাইল (মোট 106.9 কিলোবাইট) - 16-টি ফোল্ডার

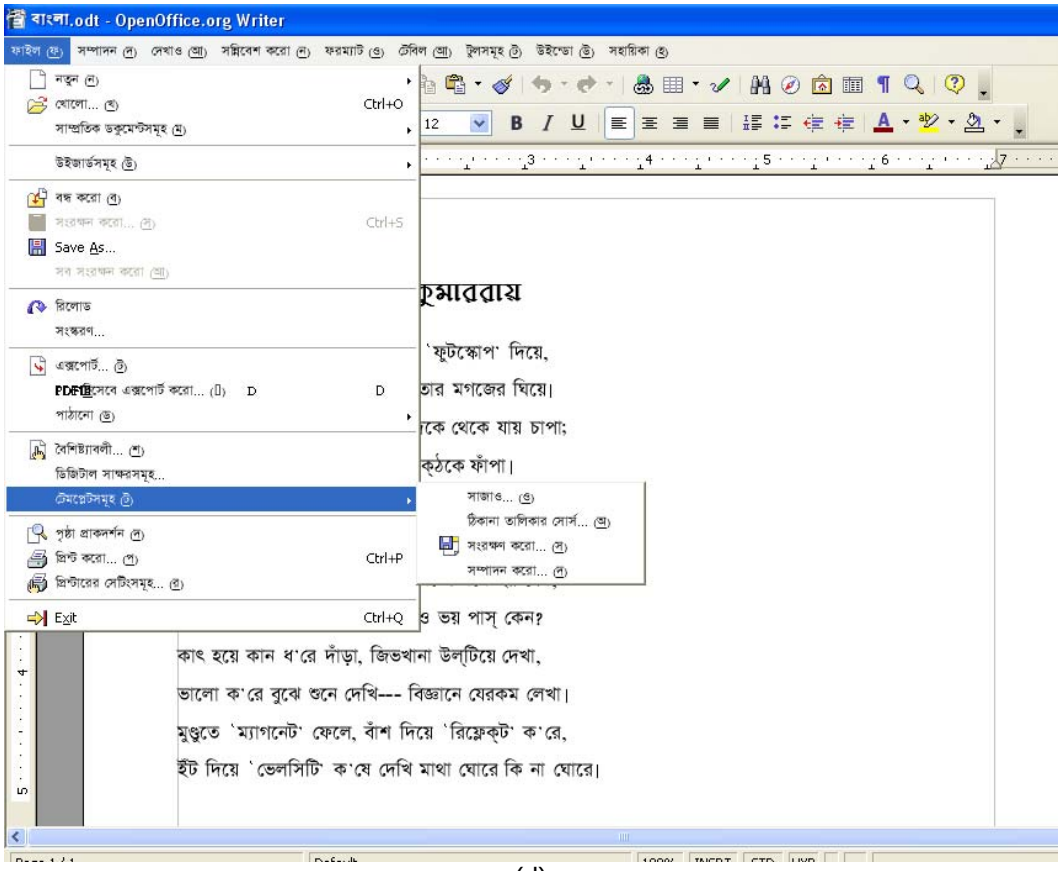
(a)


সর্বাধিক ব্যবহৃত অ্যাপ্লিকেশন
Amsn
Download Manager (KGet)
Screen Capture Program (KSnapshot)
View Disk Usage (KDiskFree)
রং
সমস্ত অ্যাপ্লিকেশন
খেলা
অফিস
খেলনা
সেটিংস
সিস্টেম
সম্পাদক
আনুষঙ্গিক
ইন্টারনেট
গ্রাফিক্স
শিক্ষামূলক মনোরঞ্জন
ডেভেলপমেন্ট
মাল্টিমিডিয়া
OpenOffice.org 1.1.2
Yahoo!
হারানো প্রাপ্তি
ফাইল অনুসন্ধান
সাহায্য
ব্যক্তিগত ফাইলসমূহ (ব্যক্তিগত ফোল্ডার)

(b)



(c)



(d)

Figure 11: (a) Localized Konqueror (KDE) Interface, (b) Localized KDE Start-Up Menu, (c) Localized Fedora Core 4, and (d) Localized Open Office

Status of Advanced Applications

A Bengali spell checker [17, 19] has been developed by the Bangla Resource Center at Indian Statistical Institute (ISI), Kolkata, which works in online and offline mode.

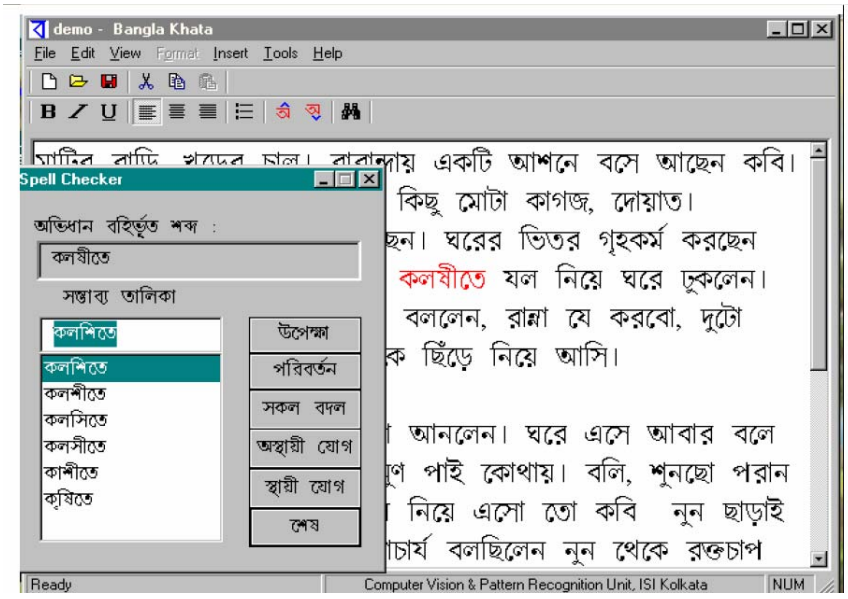


Figure 12: Bengali Spell Checker [19]

Bspeller, another spell checker shown in Figure 13, has been developed by Bangla Linux project [15], which builds on Bangla dictionary project, and uses its word list [18].

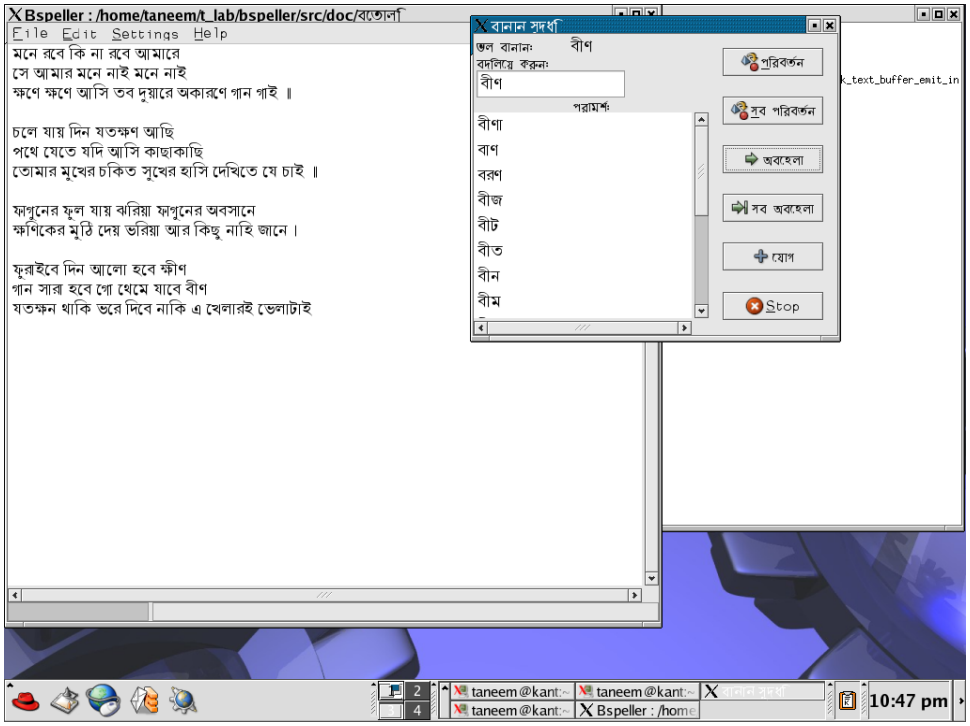


Figure 13: Bspeller Interface [18]

Significant amount of work has been done on Bengali dictionaries. These include the dictionaries by Indian Statistical Institute [19], which contains 65,000 words with parts-of-speech and meaning, and Bangla Linux project dictionary, which only contains a word list [18].

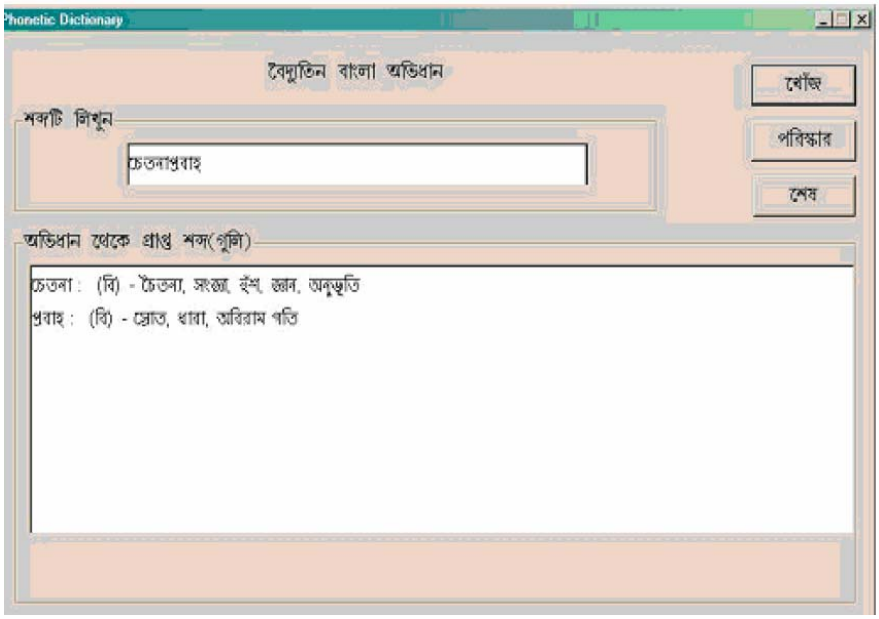


Figure 14: Bengali Dictionary by ISI [19]

ISI has also developed a Bengali OCR. It deals with single column text image. The application can handle small amount of skew, in the range of –5 to + 5 degree. The reported average character level accuracy is about 96% and word level accuracy is about 86% [19]. A Bengali text and image corpus has also been developed, which can also be used for testing any Bengali OCR system [19]. The output of this system is shown in Figure 15.

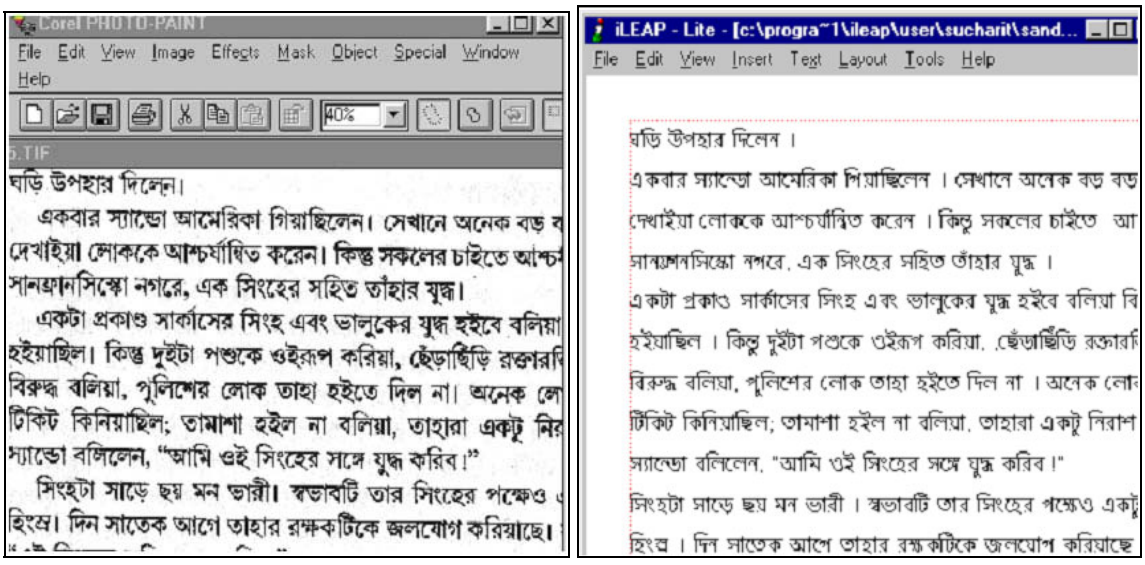


Figure 15: Bengali Image Input and Bengali Text Output of OCR [19]

Work is also being done on Bengali handwriting system [19], English-Bengali machine translation system [20], encoding converters, Bengali POS tagger and morpho-phonological analyzer at IIT Kharagpur. Additionally, Named Entity Tagger is being developed by Jadavpur Univeristy in India [20]. Limited Bengali corpus is also available through Central Institute of Indian Languages [20]. Work is also underway to develop Bengali spell checker for Open Office, Bengali lexicon, morphological analyzer and OCR through PAN Localization project [21].

References

[1] http://www.ethnologue.com
[2] http://www.omniglot.com/writing/bengali.htm
[3] http://www.cicc.or.jp/english/hyoujyunka/mlit4/7-3India/India.htm
[4] http://msdn2.microsoft.com/library/78bsyefa(en-us,vs.80).aspx
[5] http://www.angelfire.com/tx/rezaul/font.htm
[6] http://www.nongnu.org/freebangfont/downloads.html
[7] http://www.sil.org/computing/fonts/Lang/bengali.html
[8] http://salrc.uchicago.edu/resources/fonts/bengalifonts.html
[9] http://www.angelfire.com/tx/rezaul/bijoy.htm
[10] http://www.bcc.net.bd
[11] http://www.bengalinux.org/images/probhat_layout.png
[12] http://www.bengalinux.org/downloads/bn_BD
[13] http://bn.openoffice.org/
[14] http://www.bengalinux.org/devel_guide/ch03s04.html
[15] http://www.in.redhat.com/news/article/45.html
[16] http://www.bengalinux.org/
[17] http://tdil.mit.gov.in/TDIL-OCT-2003/spell%20checkers%20in%20indian%20langauges.pdf
[18] http://www.bengalinux.org/projects/dictionary/bspeller.php
[19] http://www.isical.ac.in/~rc_bangla/Brochure.pdf
[20] *Asia Pacific Association for Machine Translation Journal, Specila Issue,* Thailand 2005
[21] http://www.PANL10n.net
[22] Haque, S. "Bangladesh Experience on Unicode Standardization," in Proceedings of Seminar on Enhancement of the International Standardization Activities in Asia Pacific Region on Information Technology (SEISA-AP/IT) 2003, Mongolia.