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

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.

To the languages which will be lost before they are saved

Preface

This report is an effort to document the state of localization in Asia. There are a lot of different initiatives undertaken to localize technology across Asia. However, no study surveys the extent of work completed. It is necessary to document the status to formulate effective and coordinated strategies for further development. Therefore, current work was undertaken to collect the available data to baseline local language computing in Asia. This work has been done through PAN Localization project.

There are about 2200 languages spoken in Asia. It is difficult to undertake the task of documenting the status of all these languages. Twenty languages are being surveyed to assess the level of language computing across Asia. The selected languages have official status in Asian countries of Middle East, South, South East and East Asia. The selection has been done to cover a variety of scripts and languages of Asia, but is eventually arbitrary.

Information in the survey has been collected from what is published on the internet in English, and through surveys conducted by PAN Localization for different countries from resources known to be working in localization. Internet has been used as a primary source because it reports the most recent status. However, the survey is bound to have missed some information, especially which is not published or reported. Accuracy of the information also depends on what is published and reported. The information has not been confirmed by actually testing the reported software, as that would require an extremely diverse team and significantly more effort. Finally, the information reported is what was available at the time of data collection (first half of 2005).

This survey covers computing standards and language technology. There are many other significant aspects of localization and language processing in addition to technology. For example, it is not easily possible to develop local language computing technology without local expertise. Similarly, as localization and language computing is a long term process, it requires sustained funding focus, only achievable through explicit policies. These aspects are not addressed in the current study. .

Initially data was collected through the partner countries in PAN Localization project and by those invited for trainings conducted through this project. We are thankful to Omar Mansoor Ansari, Amar Gurung, Chea Sok Huor, Mumit Khan, Phonpassit Phissamay, Sangay Wangchuk and Ruvan Weerasinghe (PAN Localization country project leaders) and PAN Localization country team members for contributing data on their respective languages and countries, and their participation in the review process. Furthermore, we are thankful to all our training participants who provided initial data for other languages. We would also acknowledge the contribution of Norbert Klien for collecting initial information for East and South East Asian languages (and people who volunteered information to him) and Guntupalli Karunakar for data on Hindi and Assamese (latter not included in this report). Moreover, we would like to acknowledge the contribution of Ahmed Muaz in collecting some of the data published in this report through the internet. We would finally like to show our gratitude to Maria Ng Lee Hoon for her feedback at various stages, from the inception of the project, and Phyllis Lim for her continued support.

This work has been possible through the support of National University of Computer and Emerging Sciences (NUCES), Pakistan, and Pan Asia Networking (PAN) program of International Development Research Center (IDRC), Canada.

Sarmad Hussain
Nadir Durrani
Sana Gul
(Lahore, Pakistan)

PAN Localization Project

Enabling local language computing is essential for access and generation of information, and also urgently required for development of Asian countries. PAN Localization project is a regional initiative to develop local language computing capacity in Asia. It is a partnership, sampling eight countries from South and South-East Asia, to research into the challenges and solutions for local language computing development. One of the basic principles of the project is to develop and enhance capacity of local institutions and resources to develop their own language solutions.

The PAN Localization Project has three broad objectives:

- To raise sustainable human resource capacity in the Asian region for R&D in local language computing

- To develop local language computing support for Asian languages

- To advance policy for local language content creation and access across Asia for development

Human resource development is being addressed through national and regional trainings and through a regional support network being established. The trainings are both short and long term to address the needs of relevant Asian community. In partner countries, resource and organizational development is also carried out by their involvement in development of local language computing solutions. This also caters to the second objective. The research being carried out by the partner countries is strategically located at different research entry points along the technology spectrum, with each country conducting research that is critical in terms of the applications that need to be delivered to the country's user market. Moreover, PAN Localization project is playing an active role in raising awareness of the potential of local language computing for the development of Asian population. This will help focus the required attention and urgency to this important aspect of ICTs, and create the appropriate policy framework for its sustainable growth across Asia.

The scope of the PAN Localization project encompasses language computing in a broader sense, including linguistic standardization, computing applications, development platforms, content publishing and access, effective marketing and dissemination strategies and intellectual property rights issues. As the PAN Localization project researches into problems and solutions for local language computing across Asia, it is designed to sample the cultural and linguistic diversity in the whole region. The project also builds an Asian network of researchers to share learning and knowledge and publishes research outputs, including a comprehensive review at the end of the project, documenting effective processes, results and recommendations.

Countries (and languages) directly involved in the project include Afghanistan (Pashto and Dari), Bangladesh (Bangla), Bhutan (Dzongkha), Cambodia (Khmer), Laos (Lao), Nepal (Nepali), Sri Lanka (Sinhala and Tamil) and Pakistan, which is the regional secretariat. The project started in January 2004 and will continue for three years, supporting a team of seventy five resources across these eight countries to research and develop local language computing solutions. Further details of the project, its partner organizations, activities and outputs are available from its website, www.PANL10n.net.

Table of Contents

Introduction

Asia is the largest and the most culturally and linguistically diverse continent. It covers 39 million square kilometers, about 60% of land area of the world [1], and has an estimated 3.8 billion population, which is approximately 60% of the world's population [2]. There are more than 50 countries and roughly 2200 languages spoken in Asia. Being the largest, most populous and most diverse, the challenge of development of Asian community is equally important, urgent and formidable. Utilization of Information and Communication Technology (ICT) to store, process and communicate information promises an effective and efficient remedy to socio-economic problems of poverty, health, education, gender parity, governance, etc. across this continent. This technology is increasingly being leveraged in the developed and developing countries across the world, and is bound to play significant role in Asia's future.

Most of Asia still lags in effectively gaining the promised benefits of ICT. As a measure, Asia has only 34.5% of total Internet users in the world. 90% of these are in seven Asian countries [3]. There are a variety of reasons why Asia is still behind in leveraging ICTs. One of the key factors has been the limited ICT infrastructure. However, significant investment has been made over past decade to improve this infrastructure in Asia. This has had significant impact. As infrastructure has improved and information has started to flow, it has increasingly been realized that the information is not usable unless it is generated or converted in languages that Asian populations can understand. About 10-15% of Asians can communicate in non-Asian languages, and only 11% of content on the Internet is available in Asian languages [4], most of which is in Chinese, Japanese and Korean. This indicates a significant barrier for Asians to access information, and therefore to synthesize this information for their development.

The solution is to empower Asian people to generate and access culturally relevant information content. But, before the problem of content can be addressed, it is an essential precursor to enable ICTs in Asian languages. Developing ICT "software framework," including standards, terminology, utilities and applications, to enable information processing in local language is called localization. Clearly, the foremost task is to develop this software framework for Asian languages. Once ICTs are enabled in local languages, they can be more effectively used towards generating and accessing the much needed local language content.

Unfortunately, large population in Asia is also deprived from information due to high illiteracy. However, with today's technology it is also possible to overcome this barrier by employing more innovative forms of ICT interface for accessing and generating information. This includes speech interface, visual interface using touch-screens, and usage of increasing pervasive mobile technology. After basic local language computing support has been achieved, the second step is to provide these higher-end user-centric tools which catalyze generation and access of content and overcome illiteracy and similar barriers. Advanced speech and language processing applications like Machine Translation, Text-to-Speech, Speech Recognition and Optical Character Recognition systems are some such tools.

This report surveys the availability of basic standards for localization, the extent of localization on different computing platforms, and also looks at the extent of work in speech and language processing applications for the languages covered. Advanced speech and language processing applications are included, in addition to basic localization requirements, because they present a very good indicator for the level of maturity of language technology for a language and are also significant in terms of providing end-user accessibility, as has been discussed. Mobile computing and platforms will also play a significant role in the future development of Asia. However, these are not considered in the current survey.

In the next chapter the scope and process of localization is introduced and some associated terminology and functionality is explained. This is followed by a chapter on each language surveyed. There are twenty languages covered in the survey, mostly the national languages of East, South East and South Asian countries, representing a variety of scripts and language families. The choice has been made to sample a reasonable diversity, but still may not be representative of the whole of Asia. A much larger task needs to be undertaken for that purpose. A comparative summary is given at the end of this report.

References

[1] http://www.nationsonline.org/oneworld/asia.htm
[2] PRB 2004 World Population Data Sheet, http://www.prb.org
[3] http://www.Internetworldstats.com/stats3.htm#asia
[4] Knowledge Indicators Measuring Information Societies, http://www.itu.int/ITU-D/ict/papers/2003/Knowledge%20indicators%20-%20measuring%20information%20societies%20in%20AP.pdf

What is Localization?

Localization is the process of enabling computing experience in local culture and language. This would require developing solutions to input, process and output information in local language. For oral cultures, which do not have written languages, this would also mean ability to input, process and output speech instead of text. Also, it is important that the input, processing and output are agreeable with culturally acceptable norms, e.g. writing direction (left to right, right to left, top to bottom, etc.), formatting (e.g. Arabic script does not have italics form of text), color (red color represents friendship in China but danger in North America), etc. This is not easily possible, as current computing has evolved out of western cultural traditions and languages and also because Asian languages and conventions are not always as well defined as required for computational modeling. This chapter explains the scope of localization for a language.

A greater part of localization is dependent on modeling linguistic details of languages. In order for proper computational modeling, very precise definitions are required for all the relevant linguistic phenomena. For many languages spoken in developing countries, these linguistic details are either not studied or at best partially and imprecisely defined. This poses a significant obstacle to localization. Therefore many times, a significant linguistic analysis is required before taking the localization process forward. Similar challenges also exist in cultural conventions, which are known but normally not documented. Thus, it becomes very important to involve native experts in the process.

As localization involves definition and standardization of linguistic phenomena for computers, the process requires technical experts and technical organizations (e.g. Ministries of Communication or IT) to work with linguists and related organizations (e.g. National Language Authorities and/or Cultural Ministries). This poses another challenge because in most of the developing countries there is little cooperation between these two disciplines and hardly any people who have cross-disciplinary expertise. In fact, many developing Asian countries have very limited number of formally trained and practicing computational linguists.

Listed below are some of the linguistic requirements and the corresponding modeling for localization. Though the linguistic side of modeling is briefly presented, it is not discussed in detail as that is beyond the scope of this survey. This report primarily focuses on the computational modeling or localization.

Character Set and Encoding

The most fundamental and foremost requirement of localization is the definition of the character set or alphabet of a language. This includes the basic characters, digits, punctuation marks, currency symbol, special symbols (e.g. honorifics, etc.), diacritical marks, and any other symbols conventionally used in dictionary making and publishing. Though the basic repository is normally known, it has been the experience of the authors that when more precise definition is required, especially for standardization, there are always a few ambiguities. Some common linguistic level challenges faced during standardization process are listed below, to illustrate the kind of decision standardization bodies may need to make.

- It is not always clear what is part of basic character set and what is to be included in auxiliary characters
- It is sometimes ambiguous if diacritics should be independently included or extra characters need to be defined which have the diacritics fused within them
- Though basic character set is known, larger set used for dictionary making and publishing is not known or well documented
- Some characters are not defined, e.g. currency symbols

Authors have also observed that there is always temptation to find solutions which are motivated by computational requirements and therefore careful analysis and screening is needed to try to capture the independent linguistic intuition. Until character set is not defined, through this collaborative process of relevant organizations, the localization process cannot proceed effectively. However, these decisions take considerable time. Also, as these decisions are made, national and international standards need to be appropriately formulated based on them.

After the character set is agreed at linguistic-level, the next step is to assign unique numbers or codes to each character so that it may be represented in a computer. This encoding scheme is critical to standardize so that all users use same codes, to enable them to exchange information in local language. Both national and international encoding standards need to be defined and aligned. This encoding may be done by assigning a unique number between 1 to 128 (7 bits) or 1 to 256 (8 bits)[1] to each character. This mapping of characters to numbers for a language is also referred to as a Code Page. 7-8 bits are used for code pages so that they are relatively less costly to store and transmit over networks. However, 256 numbers are barely enough to encode characters of a single language. Consequently, normally there is a separate code page for each language (e.g. IBM code pages [1], Windows code pages [2, 3], ISO 8859-x standards [4], etc.).

When multilingual computing initially started, where multiple languages needed to be written within the same document, code pages had to be switched frequently. However, the demand for multilingual computing has grown so radically (especially with the advent of the Internet) that code page switching has become cumbersome. Multilingual community therefore felt the need to develop a new code page large enough to store all characters of all the languages of world, to avoid toggling among smaller code pages. As is obvious, much more than 256 slots were needed to store these characters, therefore 16 bits (2^{16} = 65536 slots) were designated for it. This code page was initially developed by Unicode consortium and later also adopted by International Standards Organization (ISO), and is called Unicode or ISO IEC 10646 [5] standard. This is a script based standard, meaning that same code is used for a letter of a script for all languages which use it. For example, letter ‘a’ of Latin script has the same code whether it is used for English, French, Spanish, Turkish, Vietnamese or Malay. Unicode already supports many scripts and languages, and work is still actively being done to encode the characters which are missing. As most computing platforms now support Unicode, it is essential for all languages to get their character repository incorporated in this standard.

Language computing existed before international standardization. This has been possible because vendors had developed their own code pages. Therefore a number of ad hoc or national encoding schemes also exist, and there is significant amount of data already available based on these schemes. Thus encoding converters are normally required to port this previously generated data to the more recent encoding.

Finally, it should also be noted that even though the new international standard, Unicode, is getting popular in many countries, previous non-Unicode encodings are still used in many countries for various reasons, some of which are listed below.

- Unicode still does not completely support the language (or script)
- Unicode recently supported the language, but international vendors have not had time to incorporate the support in their software
- Unicode support is available in recent operating systems, e.g. Microsoft Windows XP, but users are still using earlier versions (e.g. Windows 95 or 98) which are not based on Unicode. The switching cost (e.g. buying new and more powerful computers to run the more recent operating systems, or training on newer software) is too high for the users

[1] As computers use binary numbering system (0 or 1, on or off), most numbers used in computers are multiples of 2, e.g. 2, 4, 8 , 16, 32, 64, 128, 256, …

- Users are locked-in with software which is based on non-Unicode encoding. This is mostly true for the publishing industry
- National standards do not agree with international Unicode standard and former standards are still prevalent

However, with the multilingual Internet becoming so widespread, Unicode is bound to become more prevalent. There is no other international encoding standard which supports such a diverse set of scripts and languages.

Fonts and Rendering

Defining an encoding is not sufficient for supporting a language in computers. The internal codes must be displayed on the screen in terms of textual characters for it to be put to any significant use. This is done through fonts and rendering. Fonts represent the shapes of characters (also called glyphs) corresponding to each code for the language and also rules to indicate how these characters may alter shape or position on the screen in context of other characters. Font files store this information. A software (called a rendering engine) is required to take the input from user and corresponding shapes and rules from a font file to generate the actual shape and position for display on the screen.

Initially fonts were “simple” as they were designed for Latin script in which character shapes or positions are not context dependent. For example, an ‘a’ always looks the same where ever it occurs and is always on the baseline. These fonts only stored the basic shape and position of each letter, e.g. True Type fonts (TTF). However, as more scripts were computerized, it was realized that they were context-sensitive, cursive and required multiple shapes and variable positioning for their characters. For example, in Arabic script, letters have different shape in isolation, and in word-initial, word-medial and word-final positions. So font formalisms were extended and improved to store multiple shapes for each character and positioning and contextual rules for them, e.g. Open Type fonts (OTF, open standard by Microsoft and Adobe) [6] and Apple Advanced Typography (AAT by Apple) [7].

As explained earlier, displaying output requires a rendering engine, which can read a font file and create appropriate output against the input. There are a few rendering engines being used. Microsoft has developed Uniscribe rendering engine (shipped as USP10.dll file), which allows Open Type fonts to be displayed on Windows platform [8]. Similarly, Apple has a rendering engine associated with its AAT fonts [7]. Graphite engine by Summer Institute of Linguistics (SIL) is available for both Microsoft and Linux platforms [9]. Pango is another engine available for GNOME (GTK+) platform on Linux [10]. A short comparison is given at [11]. These engines support Unicode but provide varying degree of support for different scripts and languages. Level of support by some of these engines is discussed for each language later in this report.

Keyboard Layout and Input Method Engines

After character set is finalized, the next step is to place the characters across the keyboard to allow users to key-in the text. In our experience, for keyboards lack of standards is normally not the problem; the problem is that there are multiple standards. These standards can be categorized in the following manner.

- Most of these standards are inherited from layout for typewriters, tele-printers and other such devices
- Due to easy to configure utilities, which enable users to define their own on-screen keyboard layouts for most languages, there are “phonetic” versions of keyboard layouts. These are defined by users who are used to English layout and map English letters to the similar sounding characters in their language
- Many vendors also offer their own keyboard layouts, based on their own encoding schemes. These may be arbitrarily different from others

The existing standards may be adopted and adapted for newer standards. The decision could be based on a variety of (not always scientific) reasons. Some of the problems associated with keyboards are listed below, which would need rectification.

- A keyboard layout may not include all characters in a language encoded by current computing standards, e.g. Unicode, because character set inventory has been expanded or altered from the earlier definition, e.g. tele-printers had different requirements from publishing industry so layouts for them may not have all the characters. Also, many countries are now introducing currency symbols, which did not exist earlier
- Due to mechanical limitations, earlier layout was not intuitive for writing system of a language; those mechanical limitations are not applicable to computing paradigm any more. For example, single vowels which surround a consonant from left and right in Thai, Lao, Khmer, etc. had to be broken into two parts, one typed before the consonant and other after the consonant due to mechanical limitations. This is not a limitation in computing paradigm
- Sometimes encoding has implications on keyboards. For example, Unicode has redundancies due to some design decisions, e.g. backward compatibility. It has been a compromise between practical and academic challenges. So it has to be decided which letter(s) within the encoding need to be placed on the keyboard.

Faced with these challenges, the countries need to reach a consensus on a formal layout which can serve their languages as comprehensively as possible and is intuitive for the users.

Some languages have many more “characters” (ideo- or pictograms) than can conveniently fit on a keyboard, e.g. Chinese, Japanese and Korean. Therefore, innovative ways have been defined to input these languages. They include typing strokes or typing Latin based “phonetic” sequences, which eventually create the required symbol(s). This requires intelligent programs working with the keyboards. These companion programs are normally referred to as Input Method Engines (IMEs). Languages may require them in the background, in addition to a keyboard. IMEs may also require algorithmic definition and standardization at national and international levels.

Collation

For applications which go beyond basic word processing, one of the most significant standards required for processing of any language is the definition of collation or sorting sequence, also sometimes called lexicographic sequence. Given different words in any language, collation determines the order in which they would be arranged, as is expected by the users. This is defined by their arrangement in the dictionaries. This standard is required for indexing in databases and any significant textual processing, e.g. making voter lists.

Encoding standards are normally implicitly based on character order, but often do not determine collation completely. This is especially true for Unicode standard, which defines an arbitrary collation order (based on default character collation weights given in DUCET) which does not sort languages properly. Unicode standard requires language specific collation weights specified and standardized independently by relevant organizations for each language. These weights can be used with Unicode Collation Algorithm (UCA, available at Unicode website [5]) for sorting. This algorithm orders words based on collation weights provided to it for a language (see [13] for further explanation).

Languages use a variety of mechanisms to collate strings. This may be based on stroke count or phonetically equivalent Latin strings (e.g. in Chinese, Japanese and Korean), letter sequence along with diacritics and/or capitalization (e.g. in Latin based scripts), consonantal root (e.g. in Arabic language), dictionary order (e.g. in Khmer on Choun Nat Dictionary) or syllabic content (e.g. in Lao), etc. (see [12] for examples). For many languages in developing countries, this sequence is not very precisely defined. In authors' own experience, analyses have shown that different dictionaries in at least some languages do not agree in collation especially in finer details. However, for the computer, these orders must be defined to last detail. First step, again, is to involve language and cultural authorities and other relevant organizations to finalize the linguistic level standards for collation very precisely for all the characters encoded. This has to be done at the level of each language, for at least a country or a region. Second step would then involve developing effective algorithms or collation weights to realize that order. Many times lexicographic order for existing words may be determined based on dictionaries. However, in these cases mechanisms still have to be devised for the introduction of new words and proper names not present in the dictionaries.

Locale

Locale is used to define some basic language and cultural conventions for the user interface of computers and other ICT devices. It includes definition of date, time, number and other formats preferred by different countries. For example, fractional part in a number is separated by a dot in US and UK but by a comma in some European countries. It also specifies day, month and other common strings, currency symbols and calendars used by different cultures. Locales need to be defined in standard repositories so that same information can be used by everybody for consistency. One such repository, recently established to eliminate any variations, is Common Locale Data Repository (CLDR), available through Unicode website [5]. IBM ICU also has locale definitions. Locales are also maintained by other vendors. Locales are defined for every language for every country. Therefore, a combined language and country identification is used, e.g. ur_PK indicates Urdu as spoken in Pakistan and ur_IN indicates Urdu as spoken in India. These language and country codes are standardized through ISO 639.2 [14] and ISO 3166 [15] standards respectively.

Many developing countries are still not decided on standard conventions and therefore it becomes difficult to define these locales. For example, in Urdu in Pakistan both Latin and Arabic script digits are used and people disagree on which conventions should be used in the future. Once the conventions are defined by a country for a language they are submitted for standardization.

Interface Terminology Translation

With input and output of text possible, and basic locale defined, another task is to translate the remaining software interface in local languages. This means translating all the words of an application which appear on the screen into local language. This includes words or phrases in menu items, help files, error messages, etc. Basic interface words and phrases per application are normally less than 5000, and need to be translated to have core interface available. However, for complete translation, including help files, the number of words and phrases goes into hundreds of thousands. This is a formidable task, especially because consistency and quality of translation requires careful analysis. Translation memory softwares are normally used to assist in this task.

Many times another challenge is that no words in local language are not available to provide the sense required. Thus, considerable invention is required within language for the translation. Other possibility is to transliterate and adopt the foreign language words (e.g. "Computer" in many languages is the same word transliterated; an author's experience has shown "cursor" as another difficult word to translate). The decision between coining a new word or sense of a word versus transliteration is often difficult. Furthermore, in many developing countries linguists are not very proficient in computer usage and computer scientists are not very linguistically competent. This makes the translation task even harder. Many platforms (e.g. Microsoft and Linux) provide glossaries with English examples to assist in translation, for example, the glossaries for Linux are provided in *.po files. To allow basic use of computers, it would be essential to localize at least a word processor, an email client and a web browser.

Advanced Applications

Basic localization paves the way to access and generate information. However, if ability of user to access and generate information is the criteria, basic localization is not sufficient. More advanced language processing applications are required to enable end-users to completely leverage the ability of the ICTs. Language based applications facilitate and enhance the information access and generation process. Simpler user-end applications include spelling checkers, find/replace utilities, etc. which assist word processing. However, advanced language applications give the end-user significantly more accessibility. Speech applications would provide illiterate and handicapped population access to technology. Users will be able to hear their email, even if they cannot read it, using text-to-speech systems. They will be able to generate and send email using speech recognition systems. Automatic machine translation systems make existing information from other languages available in local languages. Optical character recognition systems quickly convert culturally relevant typewritten content in textual (and therefore searchable and concise) format. These applications are now mostly developed based on statistical modeling for higher performance and accuracy. However, statistical modeling is not possible without lexica and corpora (parallel corpora for applications like machine translation, in which same text is available and aligned in multiple languages). Development of these applications requires significant linguistic and computational skills and numerous linguistic resources.

Development Platforms

Localization and language processing can be done on proprietary or open platforms. Currently, Microsoft Windows platform is widely used in developing Asia. In addition, Linux platform, being open and free, is slowly gaining strength as a viable alternative. There are additional platforms, including Apple, Unix, etc. However, the survey is limited to the first two because of their presence or potential for end-user deployment in developing Asia. Within Linux there are many environments, e.g. Xwindows, KDE, GNOME etc. Many of these have also been discussed. In addition, localization of Open Office suite, which can work with both Microsoft and Linux platforms, is also evaluated. Enabling applications on these platforms requires following through all the steps discussed earlier in this section.

References

[1] http://www-03.ibm.com/servers/eserver/iseries/software/globalization/codepages.html
[2] http://www.microsoft.com/globaldev/reference/wincp.mspx
[3] http://www.microsoft.com/globaldev/reference/oem.mspx
[4] http://en.wikipedia.org/wiki/ISO_8859
[5] www.unicode.org
[6] http://www.adobe.co.uk/type/opentype/main.html
[7] http://developer.apple.com/fonts/TTRefMan/RM06/Chap6AATIntro.html
[8] http://www.microsoft.com/typography/developers/uniscribe/default.htm
[9] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&cat_id=RenderingGraphite
[10] http://www.pango.org/
[11] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=GraphiteFAQ#e81fde23
[12] Wissink, C. and Kaplan, M., “Sorting it all out: An Introduction to Collation”, in *Proceedings of 23rd International Unicode Conference,* Prague, Czech Republic, March 2003.
[13] Gillman, R., *Unicode Demystified: A Practical Programmer’s Guide to the Encoding Standard,* Addison-Wesley, Boston, USA, 2003.
[14] http://www.loc.gov/standards/iso639-2/
[15] http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1.html

Arabic

Arabic is a Semitic language spoken by about 206 million people across the world, especially in Middle East and North Africa, where it is also the national language of many countries [1]. There are many dialectical variations of Arabic across this region. It is widely used as a medium of communication in schools, government institutions and media in most of these Arabic speaking countries. Figure 1 below shows the linguistic lineage of standard Arabic [1].

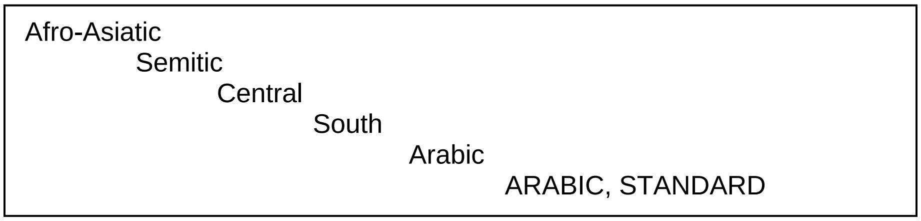


Figure 1: Language Family Tree for Arabic [1]

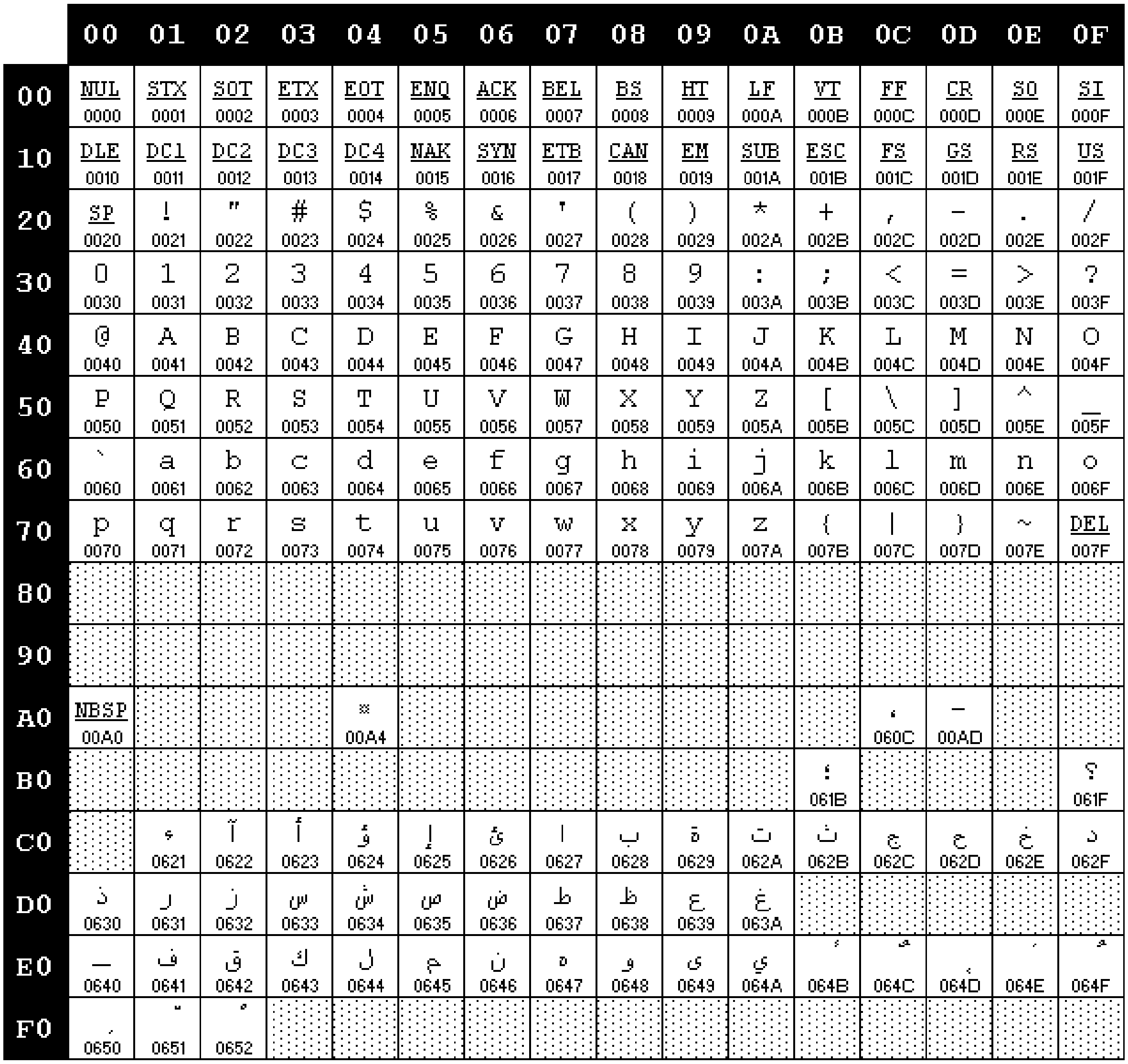
Arabic script has evolved from the ancient Aramaic script, and has been in use since the 4th century AD. Earliest known Arabic inscriptions date back to 512 AD [2].

Character Set and Encoding

Unicode Arabic script block ranging from 0600-06FF is the standard character set encoding used for Arabic language. There is an additional set of code tables given for Arabic. Arabic Supplement (0750-0764) contains additional Arabic script characters used in African languages and Arabic Presentation Forms A and B (FB50-FDFF and FE70-FEFF respectively). Presentation forms have been introduced for backward compatibility and except for ligatures (FDFx) other codes are not recommended for current use.

ISO 8859-6 is also widely used, and is shown in Figure 2. This standard contains Arabic in addition to basic Latin characters and is an 8-bit standard.

These standards have been derived from earlier standards, e.g. ASMO 449, CODAR-U and ISO 9036. For a more comprehensive overview and reference list see [4]. Microsoft also used Arabic code page 1256 based on these earlier standards [5].



	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	DEL 007F
80																
90																
A0	NBSP 00A0				¤ 00A4								، 060C	- 00AD		
B0												؛ 061B				؟ 061F
C0		ء 0621	آ 0622	أ 0623	ؤ 0624	إ 0625	ئ 0626	ا 0627	ب 0628	ة 0629	ت 062A	ث 062B	ج 062C	ح 062D	خ 062E	د 062F
D0	ذ 0630	ر 0631	ز 0632	س 0633	ش 0634	ص 0635	ض 0636	ط 0637	ظ 0638	ع 0639	غ 063A					
E0	ـ 0640	ف 0641	ق 0642	ك 0643	ل 0644	م 0645	ن 0646	ه 0647	و 0648	ى 0649	ي 064A	ً 064B	ٌ 064C	ٍ 064D	َ 064E	ُ 064F
F0	ِ 0650	ّ 0651	ْ 0652													

Figure 2: ISO 8859-6 Code Page for Arabic [3]

Fonts and Rendering

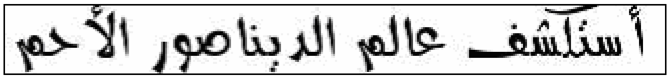
Arabic fonts are widely available. In addition, these fonts are well supported on multiple platforms.

Microsoft Platform

Microsoft ships an exclusive version of Windows and Office products in Arabic language. Microsoft Arabic Windows includes a rich inventory of fonts for Arabic, many of which are not available in the English version of Microsoft Windows.

Linux Platform

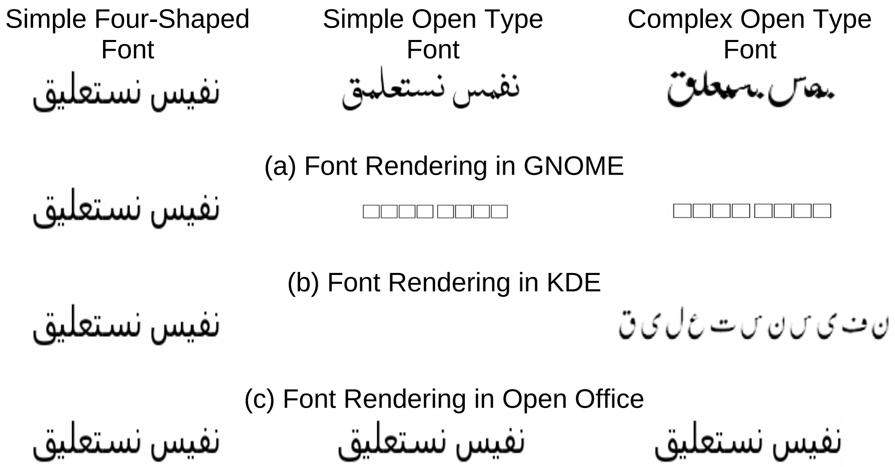
Arabic script is fully supported on all Linux based applications. However, Open Type fonts do not exhibit satisfactory results. Arabic distributions provide support for rendering only basic four shaped fonts. Arabic text in Figure 3 is written in Open Office version 1.1.0.



أستكشف عالم الديناصور الأحمر

Figure 3: Ae_Dimnah Font in Open Office 1.1.0

Results of rendering Open Type fonts in Linux distributions vary with applications as different applications in Linux use rendering support from different rendering engines. Comparative font rendering results in GNOME, KDE, Open Office and Mozilla are presented below.



(a) Font Rendering in GNOME

(b) Font Rendering in KDE

(c) Font Rendering in Open Office

(d) Font Rendering in Mozilla

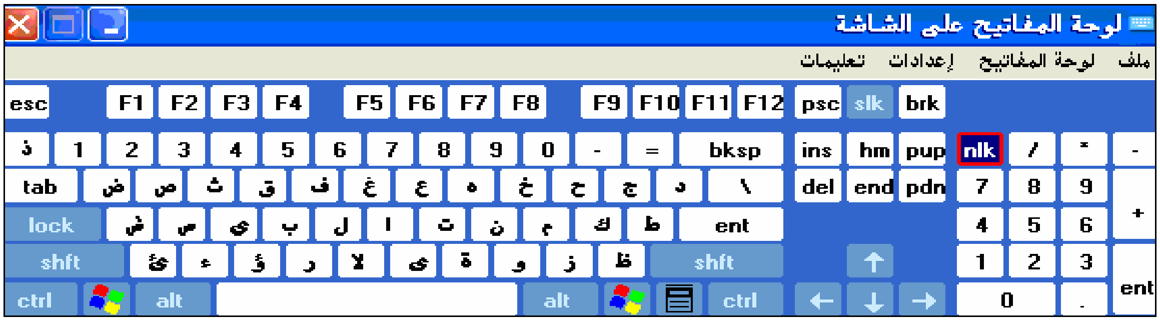
Figure 4: Font Rendering on Linux Platform

Figure 4 shows that GNOME displays the best results, but still does not work properly for Open Type fonts. KDE does not render Open Type fonts, Open Office only displays True Type fonts, and Mozilla defaults to a simple four shaped True Type font fonts because it fails to display Open Type fonts properly.

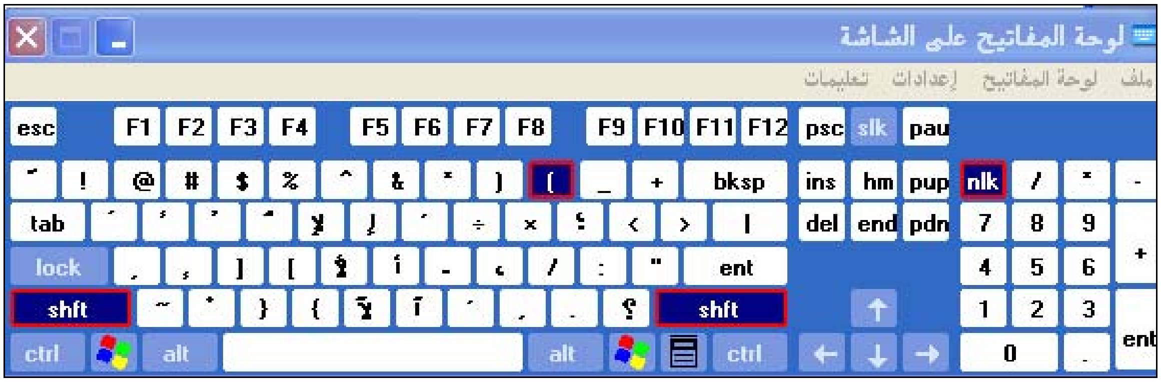
Keyboard

Microsoft Platform

Microsoft Windows XP provides an Arabic keyboard layout both in Arabic and English versions. Arabic keyboard layout is enabled by default in Arabic version. This layout is shown in Figure 5 below. Also see [6].


لوحة المفاتيح على الشاشة
ملف
لوحة المفاتيح
إعدادات
تعليمات
esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 psc slk brk
ذ 1 2 3 4 5 6 7 8 9 0 - = bksp ins hm pup nlk / * -
tab ض ص ث ق ف غ ع ه خ ح ج د \ del end pdn 7 8 9 +
lock ش س ي ب ل ا ت ن م ك ط ent 4 5 6
shft ئ ء ؤ ر لا ى ة و ز ظ shft 1 2 3
ctrl alt alt ctrl 0 . ent

(a)


لوحة المفاتيح على الشاشة
ملف لوحة المفاتيح إعدادات تعليمات
esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 psc slk pau
! @ # $ % ^ & *) (_ + bksp ins hm pup nlk / * -
tab ÷ × ؟ < > | del end pdn 7 8 9 +
lock] [: " ent 4 5 6
shft ~ } { ؟ shft 1 2 3
ctrl alt alt ctrl 0 . ent

(b)

Figure 5: (a) Normal and (b) Shift Version of Keyboard on MS Windows

Linux Platform

Red Hat 9 provides an Arabic keyboard layout which can be used both with KDE and GNOME. Once enabled the Arabic keyboard can be used on all Unicode compliant applications.

Collation

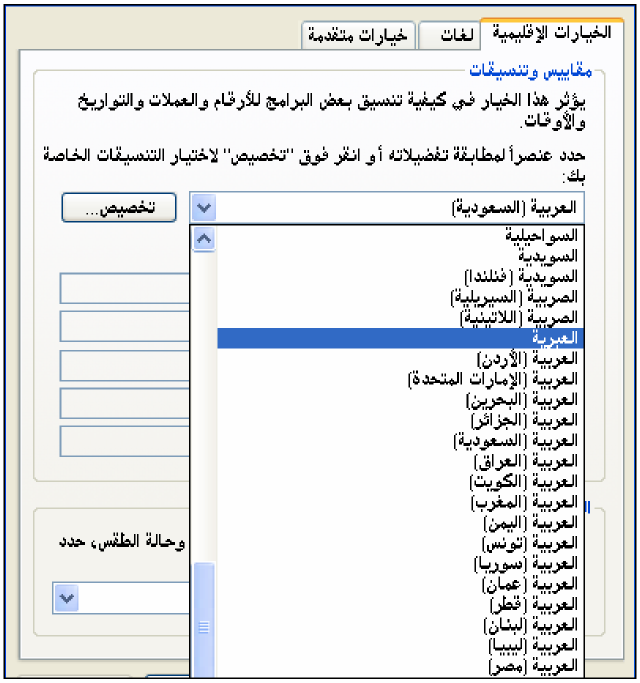
Arabic collation is supported in Arabic Windows. LC_COLLATE for Arabic language has not been defined yet. Default sequence for sorting is used, which sorts data similar to original collation sequence for Arabic language.

Locale

Arabic (ar) locales are defined in IBM ICU library and CLDR 1.3 for different countries. They support Arabic date, time and number formats, currency symbol and collation. The locale is available for many countries where Arabic is spoken as a national language (see [7] for a complete list).

Microsoft Platform

Both versions of Microsoft Windows (Arabic and English version of Microsoft XP) provide locale settings for date, time and number formats. They also support currency symbol for some countries. Figure 6 displays the text box to set the country specific Arabic locale in Arabic Windows.



مقاييس وتنسيقات

يؤثر هذا الخيار في كيفية تنسيق بعض البرامج للأرقام والعملات والتواريخ والأوقات.

حدد عنصراً لمطابقة تفضيلاته أو انقر فوق "تخصيص" لاختيار التنسيقات الخاصة بك:

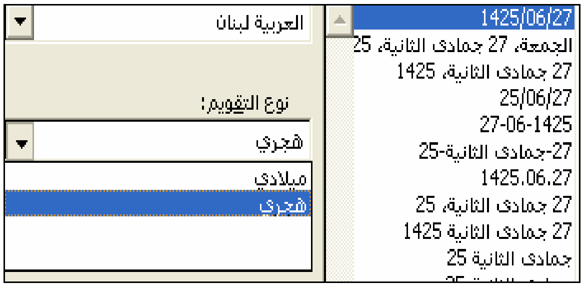
العربية (السعودية)

السواحيلية
السويدية
السويدية (فنلندا)
الصربية (السيريلية)
الصربية (اللاتينية)
العربية
العربية (الأردن)
العربية (الإمارات المتحدة)
العربية (البحرين)
العربية (الجزائر)
العربية (السعودية)
العربية (العراق)
العربية (الكويت)
العربية (المغرب)
العربية (اليمن)
العربية (تونس)
العربية (سوريا)
العربية (عمان)
العربية (قطر)
العربية (لبنان)
العربية (ليبيا)
العربية (مصر)

وحالة الطقس، حدد

Figure 6: Available Locales of Different Countries in Arabic Windows

In addition, different calendars such as Mailadi and Hijri can also be enabled, as shown for Lebanese Arabic in Figure 7.


العربية لبنان

نوع التقويم:

هجري

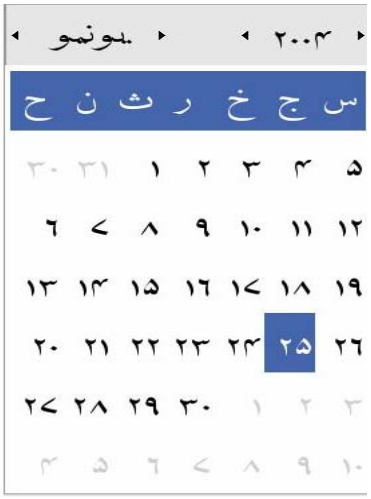
ميلادي
هجري

1425/06/27
الجمعة، 27 جمادى الثانية، 25
27 جمادى الثانية، 1425
25/06/27
27-06-1425
27-جمادى الثانية-25
1425.06.27
27 جمادى الثانية، 25
27 جمادى الثانية 1425
جمادى الثانية 25

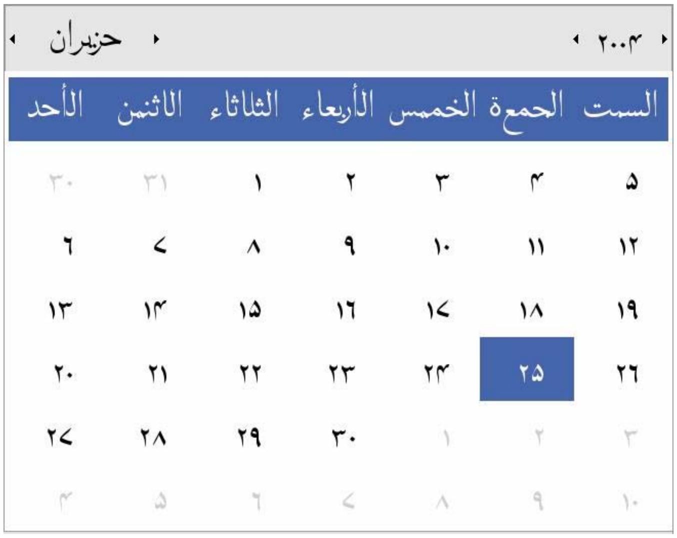
Figure 7: Calendar and Date Settings

Linux Platform

Eighteen locales have been defined in Glibc locale file for different Arabic speaking countries. Each locale includes localized settings for date, time, number formats, and collation specific to the country. Two Arabic locales, Egypt and Lebanon, are also available in Red Hat 9. Figure 8 shows date settings for Egypt and Lebanon from this distribution.



Arabic Egypt



Arabic Lebanon

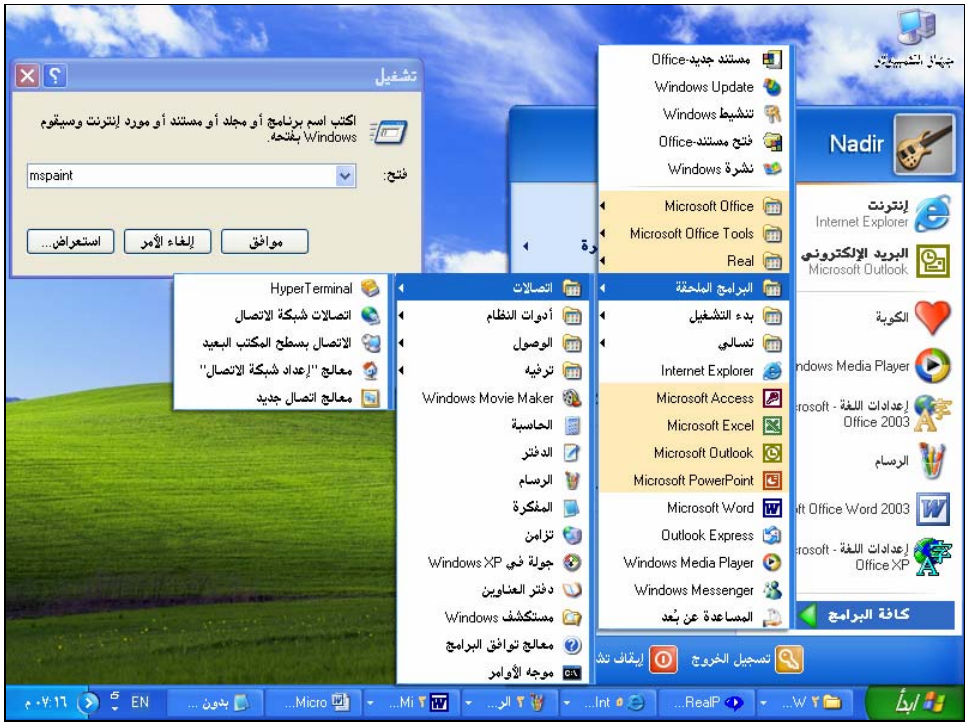
Figure 8: Calendar Settings, in GNOME for Egypt and Lebanon in Red Hat 9

Both Mozilla and Open Office (within Linux distribution) also provide locale support for Arabic.

Interface Terminology Translation

Microsoft Platform

As discussed before, Microsoft ships a complete Arabic version for Windows and Office products. This is different from other language versions, which are based on the English systems and then localized. Arabic version of Microsoft Windows is completely localized in Arabic language. All dialogue boxes, menu, messages and help files have been translated. This version has been available for a long time. Figure 9 below shows the localized version of Arabic Windows desktop and Internet Explorer.

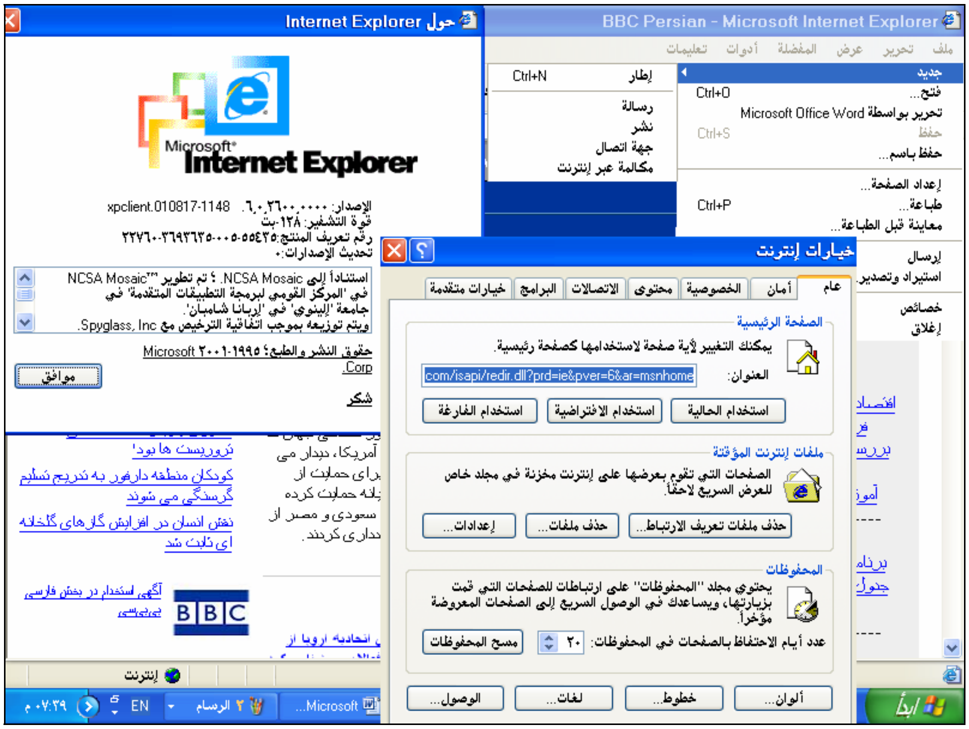


تشغيل

اكتب اسم برنامج أو مجلد أو مستند أو مورد إنترنت وسيقوم Windows بفتحه.

فتح: mspaint

(a)



حول Internet Explorer

Microsoft® Internet Explorer

الإصدار: ٦٫٠٫٢٦٠٠٫٠٠٠٠. xpclient.010817-1148
قوة التشفير: ١٢٨-بت
رقم تعريف المنتج: ٥٥٤٣٥-٠٠٠٥-٣٦٩٣٦٣٥-٢٢٧٦٠
تحديث الإصدارات: ٠

استناداً إلى NCSA Mosaic. ؛ تم تطوير NCSA Mosaic™ في 'المركز القومي لبرمجة التطبيقات المتقدمة' في جامعة 'إلينوي' في 'إربانا شامبان'.
ويتم توزيعه بموجب اتفاقية الترخيص مع Spyglass, Inc.

حقوق النشر والطبع؛ ١٩٩٥-٢٠٠١ Microsoft Corp.

شكر

BBC Persian - Microsoft Internet Explorer

خيارات إنترنت

الصفحة الرئيسية

يمكنك التغيير لأية صفحة لاستخدامها كصفحة رئيسية.

العنوان: com/isapi/redir.dll?prd=ie&pver=6&ar=msnhome

ملفات إنترنت المؤقتة

الصفحات التي تقوم بعرضها على إنترنت مخزنة في مجلد خاص للعرض السريع لاحقاً.

المحفوظات

يحتوي مجلد "المحفوظات" على ارتباطات للصفحات التي قمت بزيارتها، ويساعدك في الوصول السريع إلى الصفحات المعروضة مؤخراً.

عدد أيام الاحتفاظ بالصفحات في المحفوظات: ٢٠

(b)

Figure 9: (a) Localized Start-Up Menus in Arabic Windows, and (b) Arabic Internet Explorer

Linux Platform

Arrabix, a Debian based Linux distribution, has been localized for Arabic. Arrabix provides complete translation of the graphical user interface in Arabic. It has completely localized versions of all desktop applications including Open Office, Mozilla suite, spell checking utility (Duali), Arabic fonts, localized version of development tools like C/C++ environment, Python Development Kit and also the multimedia software and graphics tools. The following figures show desktop terminology translation in Arrabix CD.

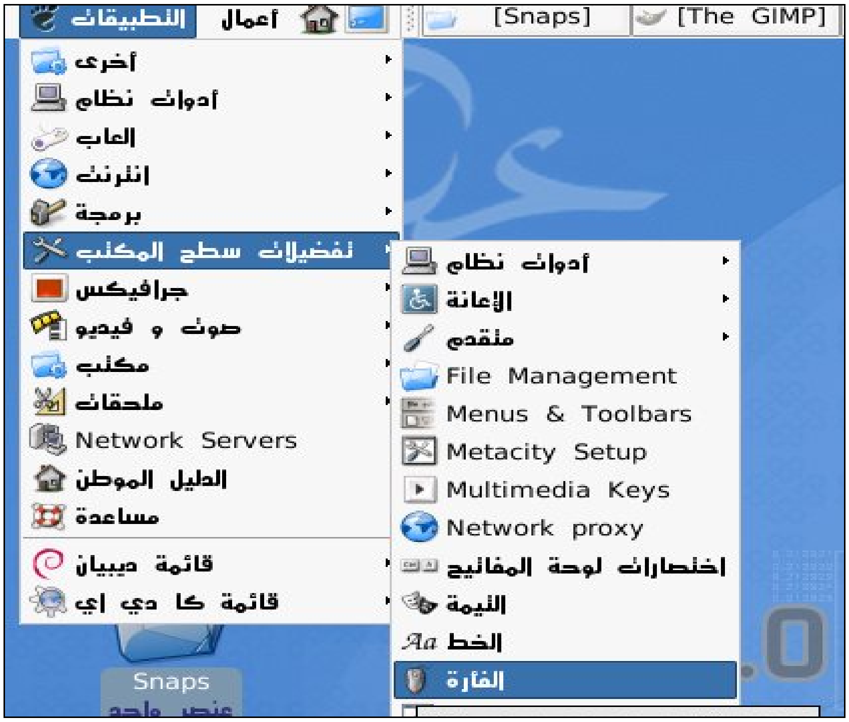
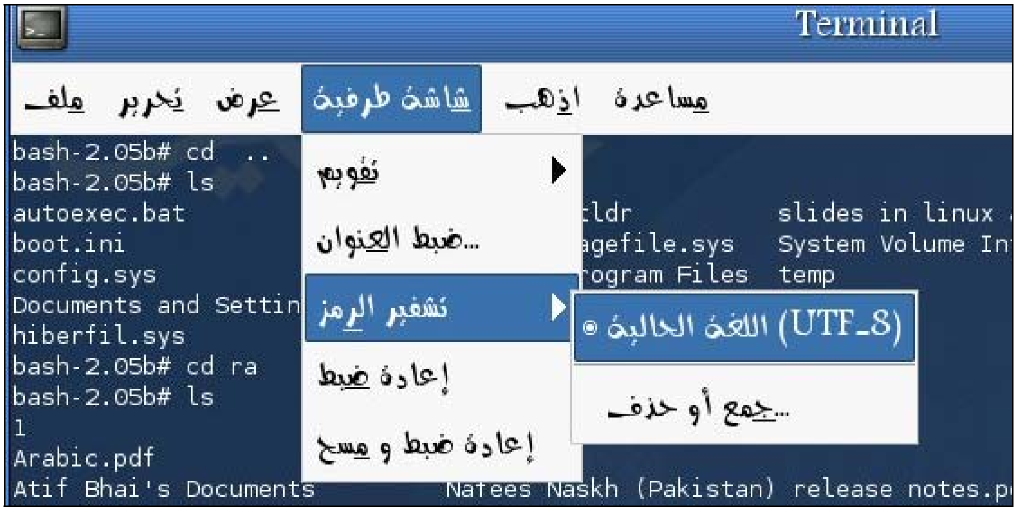


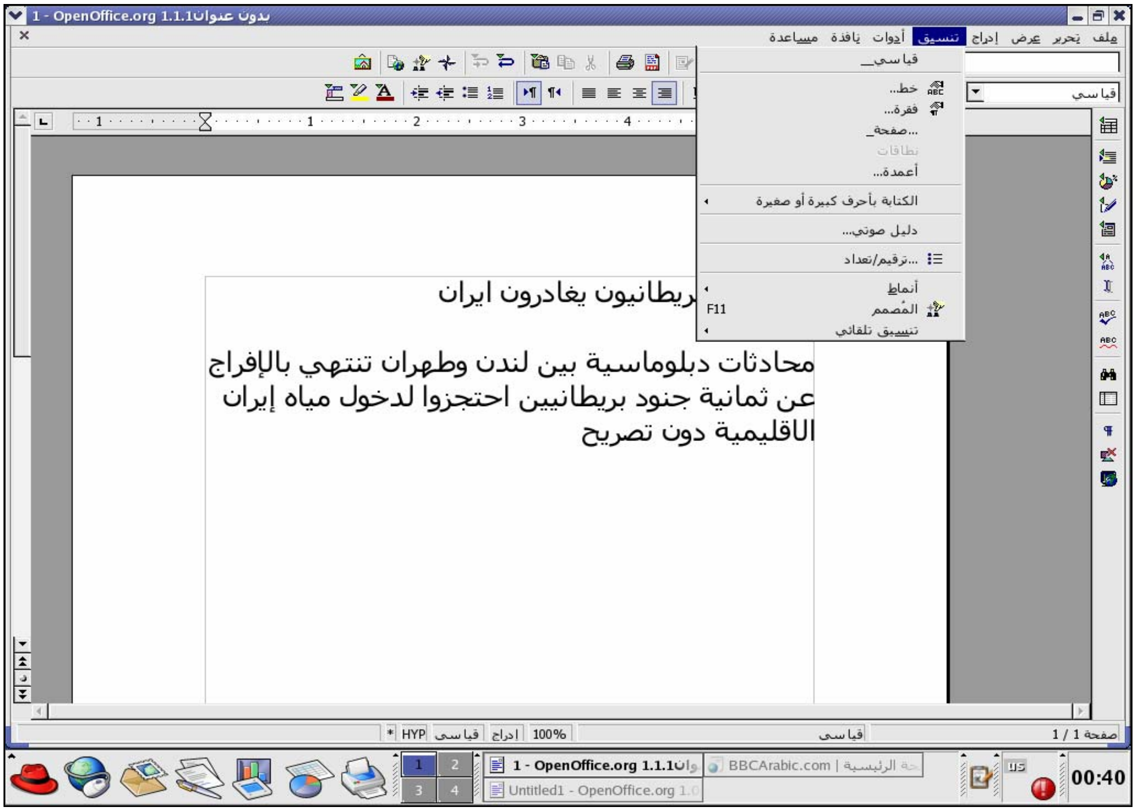
Figure 10: GNOME Start-Up Menus in Arrabix



```
bash-2.05b# cd ..
bash-2.05b# ls
autoexec.bat                 ldr              slides in linux
boot.ini                     agefile.sys      System Volume In
config.sys                   rogram Files     temp
Documents and Settin
hiberfil.sys
bash-2.05b# cd ra
bash-2.05b# ls
1
Arabic.pdf
Atif Bhai's Documents        Nafees Naskh (Pakistan) release notes.p
```

Figure 11: Terminal in Arrabix 0.8

Interface terminology translation for Arabic GNOME is 51% complete. Only basic desktop files, GNOME lib files, Nautilus browser files, etc. have been partially localized in Red Hat. Interface terminology translation for Arabic KDE is 99% complete. Mozilla 1.6 is partially localized in Arabic. However Mozilla version 1.4 provides maximum terminology translations for Arabic language. The latest version of Open Office has been completely localized in Arabic. All applications of Open Office, including Writer, Spread Sheet and Presentation etc. are available in Arabic.



ريطانيون يغادرون ايران

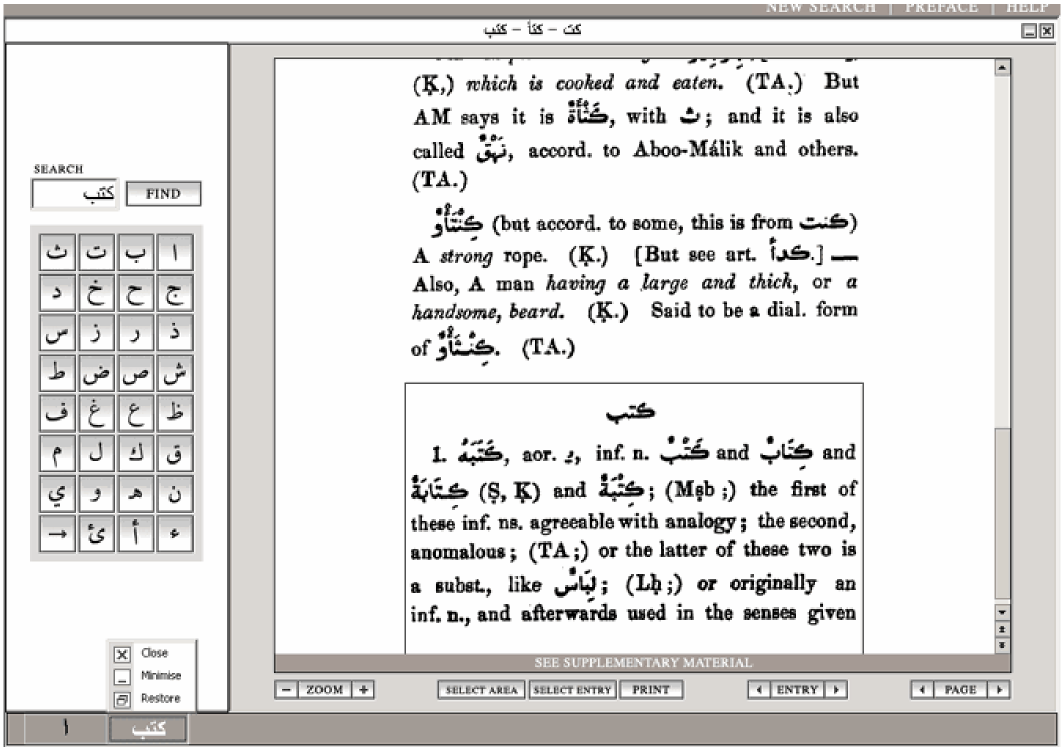
محادثات دبلوماسية بين لندن وطهران تنتهي بالإفراج
عن ثمانية جنود بريطانيين احتجزوا لدخول مياه إيران
الاقليمية دون تصريح

Figure 12: Open Office Writer

Status of Advanced Applications

A lot of work has been done and is currently underway across the world for development of Arabic language applications. Language applications available include encoding converters (e.g. see [8]), bidirectional writing utilities and spell checkers (e.g. Arabic Word Corrector, ARWOC [10], Duali (for Linux) [11], and products by Microsoft and Sakhr [12]).

Sakhr Automatic Reader [12] is a commonly used OCR. Other OCRs include Arabic OCR and ICR (intelligent character recognition) by CIYA [13]. Arabic lexicon, English-Arabic lexicon, Arabic text-to-speech system, Arabic speech recognition system, machine translation between English and Arabic and a variety of other applications are also available through Sakhr [12] and other vendors. For example, other vendor applications include text-to-speech systems by IBM, MBROLA, Lingvosoft and ArabTalk, Speech recognition systems by Natural Speech Communication, machine translation systems for English to Arabic by An-Nakel, ITRANS (Intelligent Translation System) WebTrans™ and Almisbar, and handwriting recognition systems by PackNet Arabizer Suite, IMAGiNET Arabic Writer and IFN/ENET. Some screen shots of these applications are given below.



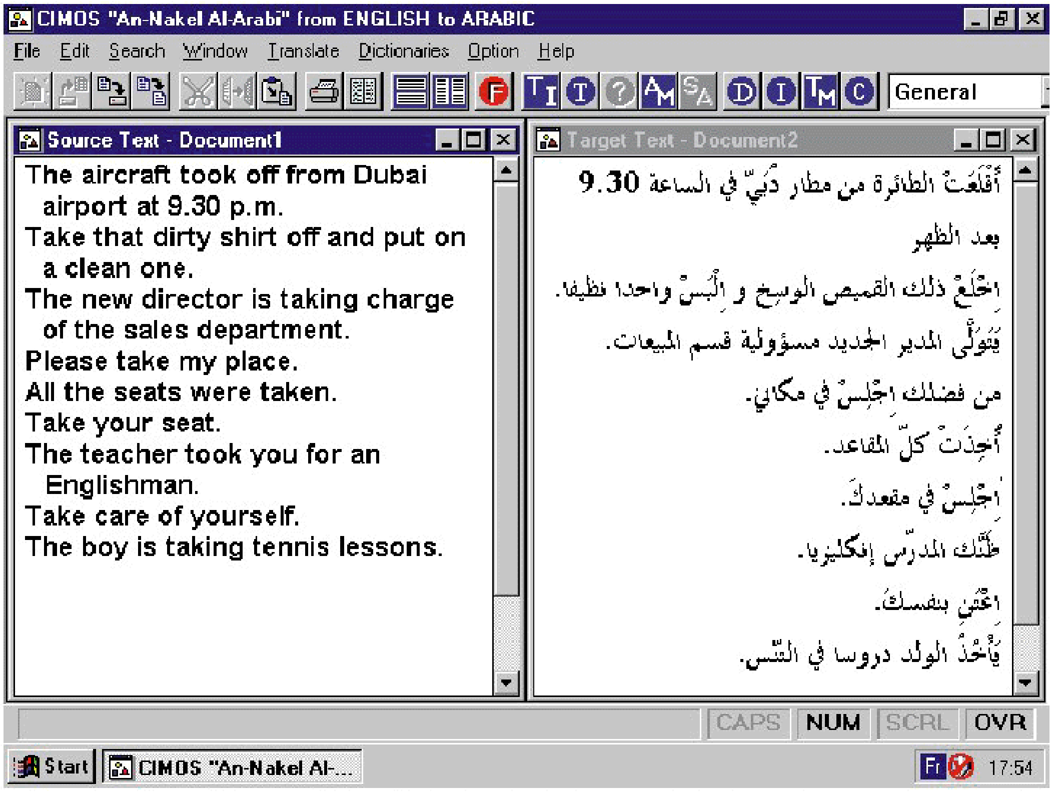
(Ḳ,) *which is cooked and eaten.* (TA.) But AM says it is كَثَاةٌ, with ث; and it is also called نَبْقٌ, accord. to Aboo-Málik and others. (TA.)

كِتْنَأْوٌ (but accord. to some, this is from كنت) A *strong* rope. (Ḳ.) [But see art. كدأ.] — Also, A man *having a large and thick,* or *a handsome, beard.* (Ḳ.) Said to be a dial. form of كِنْثَأْوٌ. (TA.)

كتب

1. كَتَبَهُ, aor. ـُ, inf. n. كَتْبٌ and كِتَابٌ and كِتَابَةٌ (Ṣ, Ḳ) and كِتْبَةٌ; (Mṣb ;) the first of these inf. ns. agreeable with analogy; the second, anomalous; (TA ;) or the latter of these two is a subst., like لِبَاسٌ; (Lḥ ;) or originally an inf. n., and afterwards used in the senses given

Figure 13: Lane's Arabic English Lexicon [14]


CIMOS "An-Nakel Al-Arabi" from ENGLISH to ARABIC
File Edit Search Window Translate Dictionaries Option Help
General
Source Text - Document1
Target Text - Document2
CAPS NUM SCRL OVR
Start
CIMOS "An-Nakel Al-...
Fr
17:54

The aircraft took off from Dubai airport at 9.30 p.m.
Take that dirty shirt off and put on a clean one.
The new director is taking charge of the sales department.
Please take my place.
All the seats were taken.
Take your seat.
The teacher took you for an Englishman.
Take care of yourself.
The boy is taking tennis lessons.

أَقْلَعَتْ الطائرة من مطار دُبَيّ في الساعة 9.30
بعد الظهر
اِخْلَعْ ذلك القميص الوسخ و اِلْبَسْ واحدا نظيفا.
يَتَوَلَّى المدير الجديد مسؤولية قسم المبيعات.
من فضلك اِجْلِسْ في مكاني.
أُخِذَتْ كلّ المقاعد.
اِجْلِسْ في مقعدكَ.
ظَنَّك المدرّس إنكليزيا.
اِعْتَنِ بنفسكَ.
يَأْخُذُ الولد دروسا في التّنس.

Figure 14: An-Nakel Machine Translation [15]

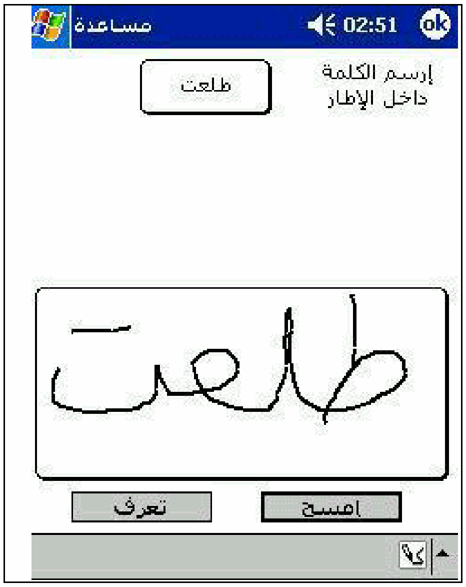


Figure 15: IMAGiNET Arabic Writer [16]

References

[1] http://www.ethnologue.com/show_language.asp?code=arb
[2] http://www.omniglot.com/writing/arabic.htm
[3] http://www.microsoft.com/globaldev/reference/iso/28596.mspx
[4] “Unicode and Arabic Script.” http://ead.staatsbibliothek-berlin.de/2003/unicode/arabic.pdf
[5] http://www.microsoft.com/globaldev/reference/sbcs/1256.mspx
[6] http://www.microsoft.com/globaldev/reference/keyboards.mspx
[7] http://www.unicode.org/cldr/version/1.3.html
[8] http://www.microsoft.com/middleeast/arabicdev/beta/converter/
[9] http://www.unicode.org
[10] http://www.coltec.net/new1/arwoc.html
[11] http://www.arabeyes.org/project.php?proj=duali
[12] http://www.sakh.com
[13] http://members.aol.com/gnhbos/ocroptions.htm
[14] http://www.fonsvitae.com/laneslexicon.html
[15] http://www.translation.net/nakel_translation.html
[16] http://www.pocketgear.com/software_detail.asp?id=10541

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.

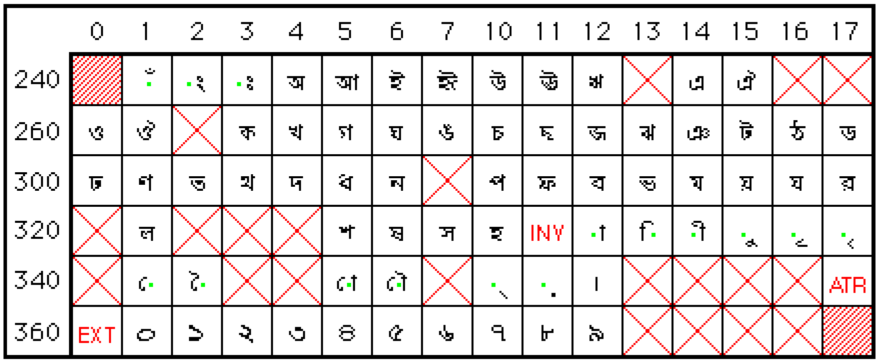
Indo-European
    Indo-Iranian
        Indo-Aryan
            Eastern zone
                Bengali-Assamese
                    BENGALI

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.



	0	1	2	3	4	5	6	7	10	11	12	13	14	15	16	17
240	▨	ঁ	ং	ঃ	অ	আ	ই	ঈ	উ	ঊ	ঋ	☒	এ	ঐ	☒	☒
260	ও	ঔ	☒	ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ	ট	ঠ	ড
300	ঢ	ণ	ত	থ	দ	ধ	ন	☒	প	ফ	ব	ভ	ম	য়	য	র
320	☒	ল	☒	☒	☒	শ	ষ	স	হ	INV	া	ি	ী	ু	ূ	ৃ
340	☒	ে	ৈ	☒	☒	ো	ৌ	☒	্	়	।	☒	☒	☒	☒	ATR
360	EXT	০	১	২	৩	৪	৫	৬	৭	৮	৯	☒	☒	☒	☒	▨

Figure 2: Language Family Tree for Bengali [19]

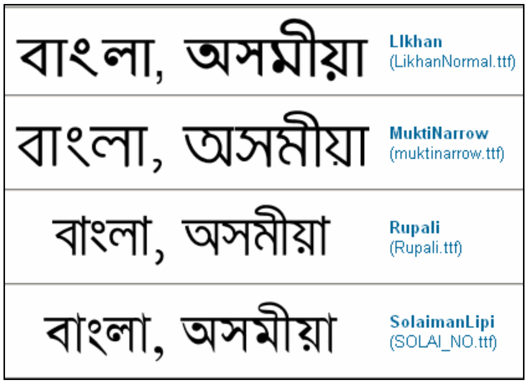
A Bangladesh national standard for Bengali text encoding, BSD 1520, has also been recommended by Bangladesh Standards and Testing Institute (BSTI) in 1995. However, this encoding is used infrequently. This standard was revised in 2000 and eventually Unicode was also formally adopted as national standard. 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.



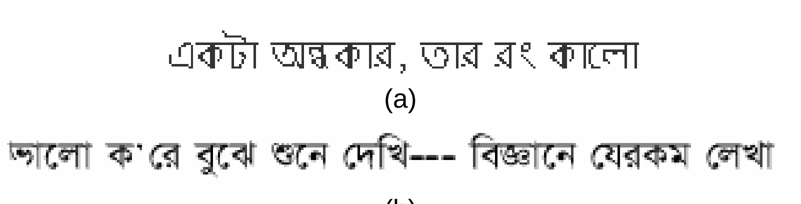
বাংলা, অসমীয়া	LIkhan (LikhanNormal.ttf)
বাংলা, অসমীয়া	MuktiNarrow (muktinarrow.ttf)
বাংলা, অসমীয়া	Rupali (Rupali.ttf)
বাংলা, অসমীয়া	SolaimanLipi (SOLAI_NO.ttf)

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.



একটা অন্ধকার, তার রং কালো

(a)

ভালো ক'রে বুঝে শুনে দেখি--- বিজ্ঞানে যেরকম লেখা

(b)

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.

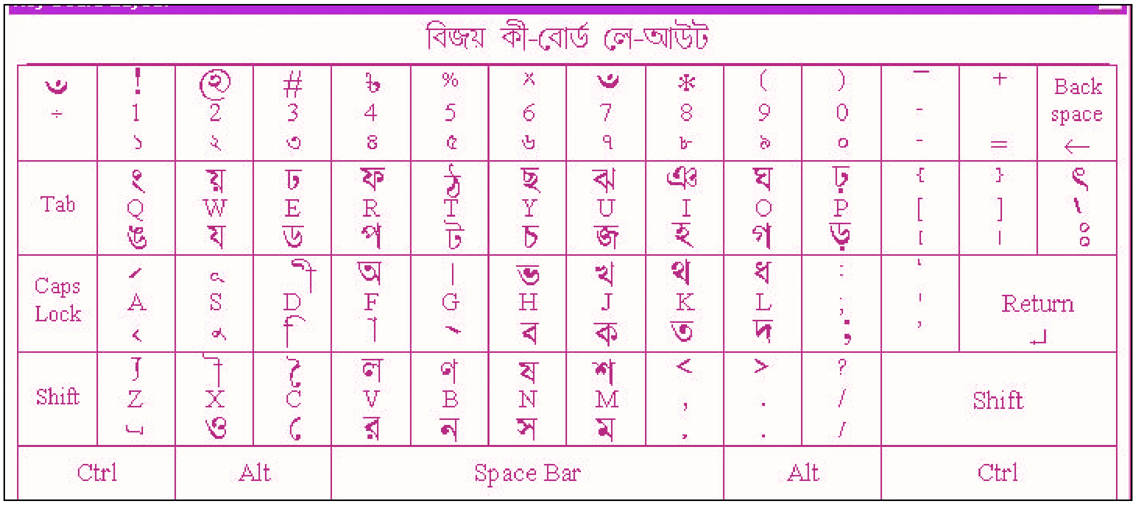
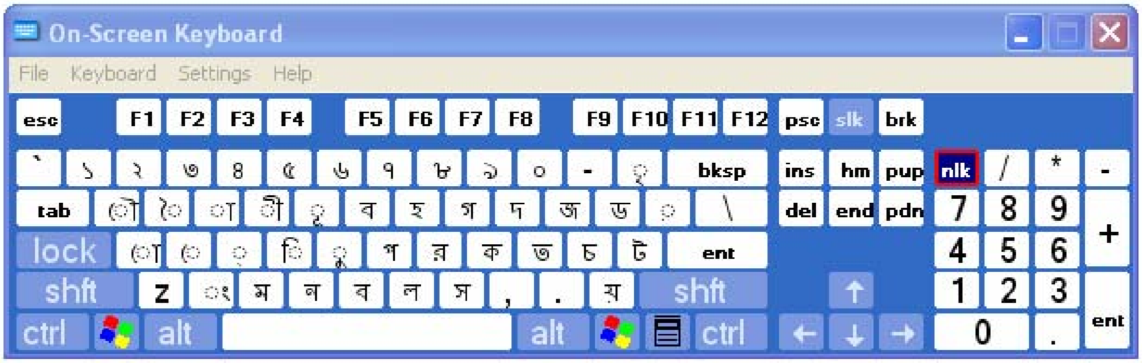


Figure 5: Bijoy Bengali Keyboard Layout [9]

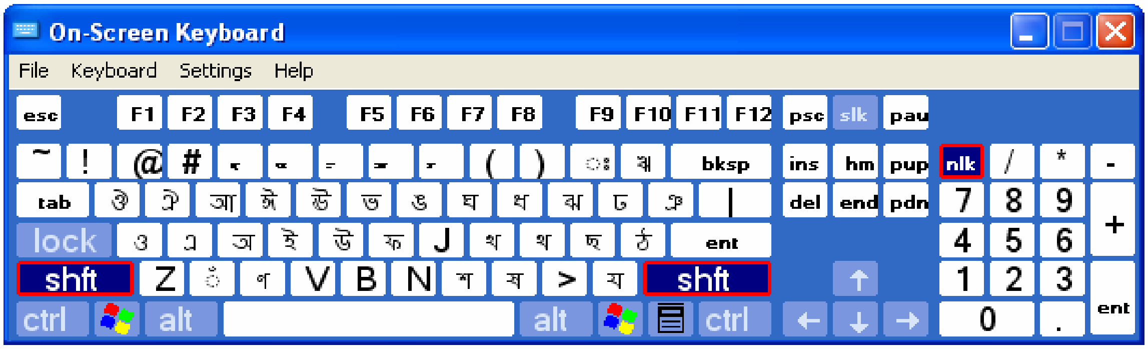
Bangladesh Computer Council published Bangladesh Jatiyo (national) keyboard standard in 2003, based on a phonetic layout. However it is reported to be used rarely [10].

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.

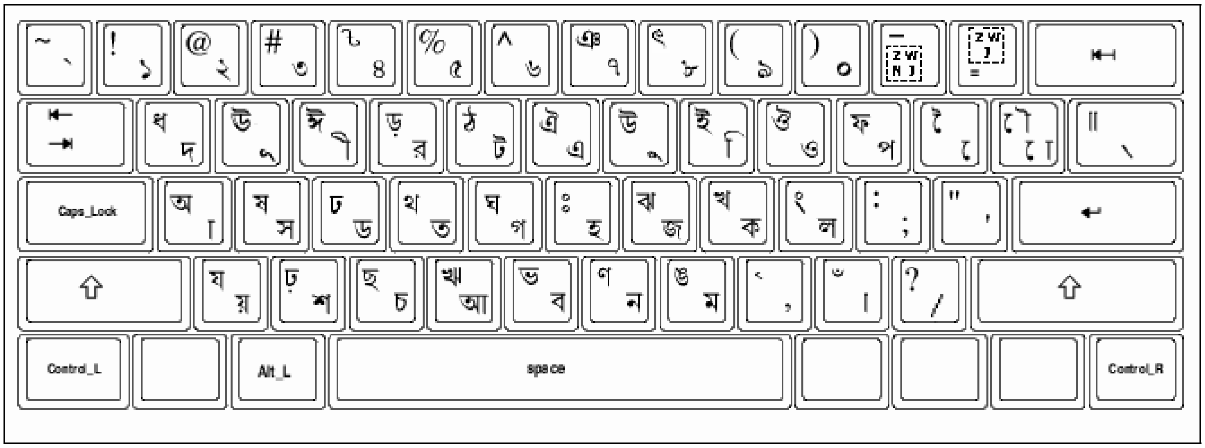


Figure 7: Probhat Phonetic Keyboard Layout [11]

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.

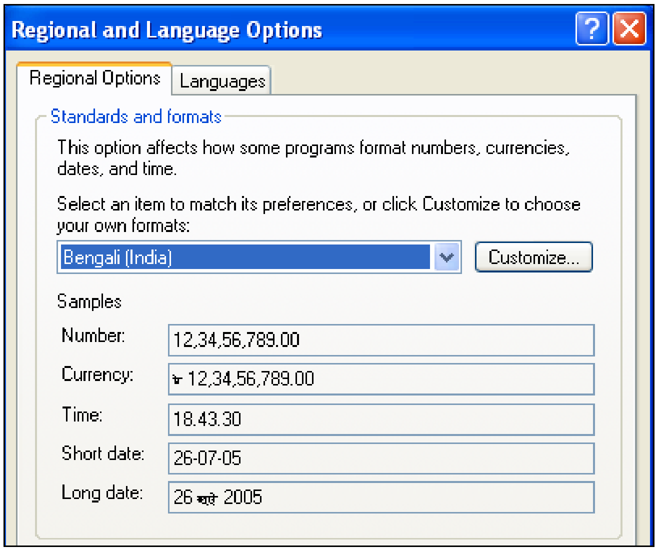
Bengali sorting is not supported by Microsoft or Linux platforms.

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.



Regional and Language Options

Regional Options | Languages

Standards and formats

This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Bengali (India) | Customize...

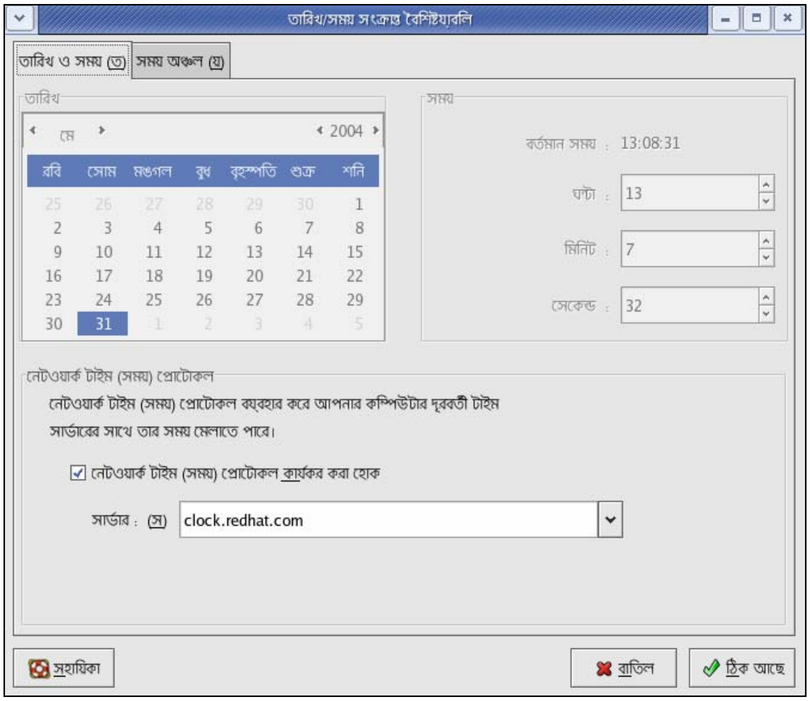
Samples

Number:	12,34,56,789.00
Currency:	৳ 12,34,56,789.00
Time:	18.43.30
Short date:	26-07-05
Long date:	26 জুলাই 2005

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.



তারিখ/সময় সংক্রান্ত বৈশিষ্ট্যাবলি

তারিখ ও সময় (ত) | সময় অঞ্চল (য)

তারিখ

মে 2004

রবি	সোম	মঙ্গল	বুধ	বৃহস্পতি	শুক্র	শনি
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

সময়

বর্তমান সময় : 13:08:31

ঘণ্টা : 13

মিনিট : 7

সেকেন্ড : 32

নেটওয়ার্ক টাইম (সময়) প্রোটোকল

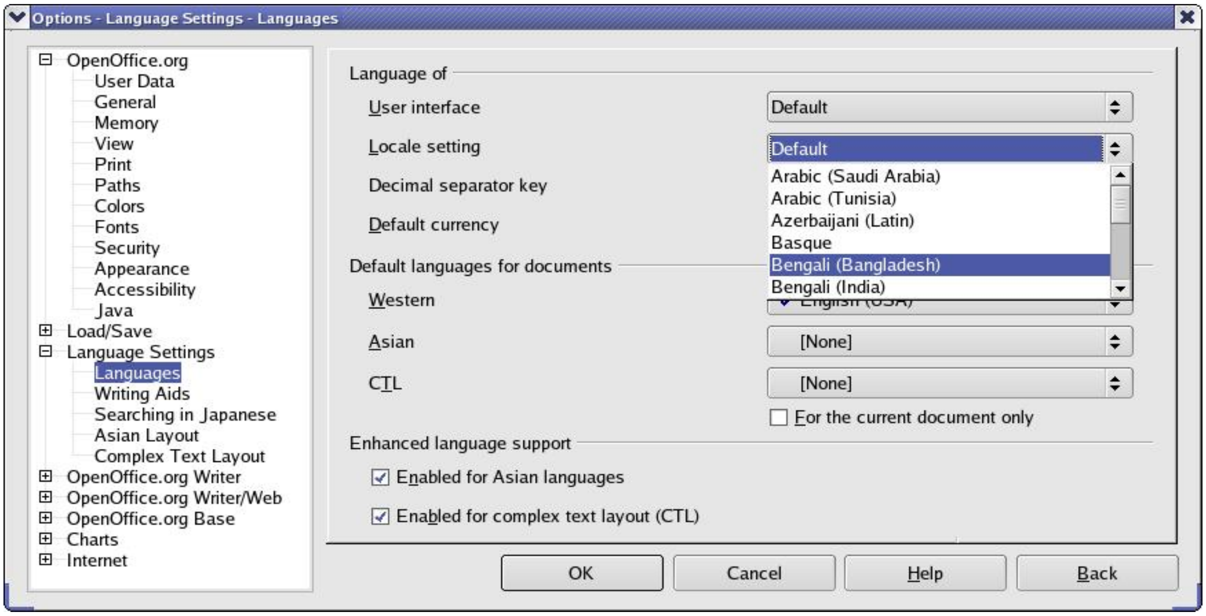
নেটওয়ার্ক টাইম (সময়) প্রোটোকল ব্যবহার করে আপনার কম্পিউটার দূরবর্তী টাইম সার্ভারের সাথে তার সময় মেলাতে পারে।

☑ নেটওয়ার্ক টাইম (সময়) প্রোটোকল কার্যকর করা হোক

সার্ভার : (স) clock.redhat.com

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).



Language of

User interface: Default

Locale setting: Default

Decimal separator key

Default currency

Default languages for documents

Western: English (USA)

Asian: [None]

CTL: [None]

□ For the current document only

Enhanced language support

☑ Enabled for Asian languages

☑ Enabled for complex text layout (CTL)

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].

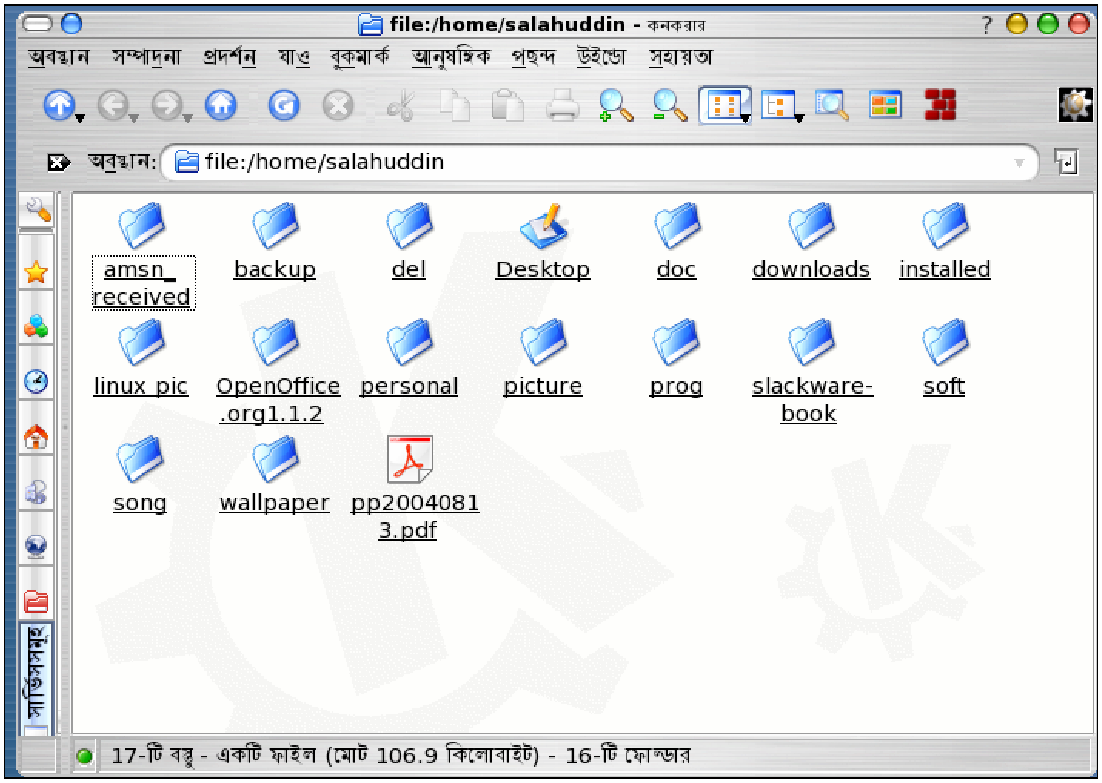
Microsoft Platform

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

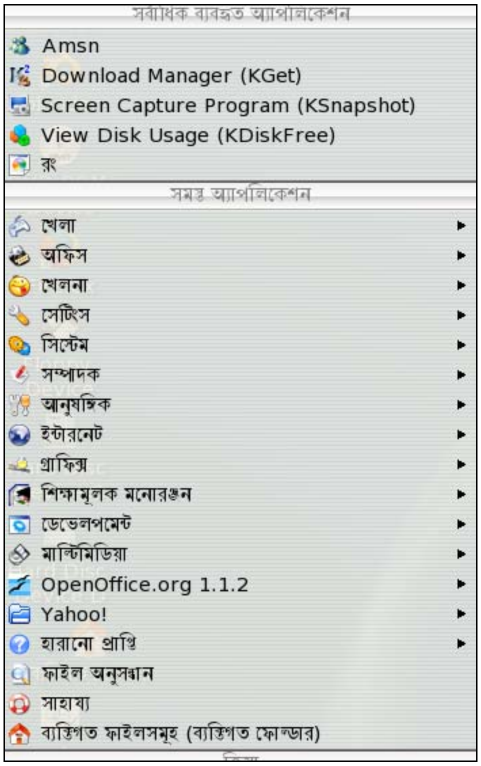
Red Hat has launched a beta version of Bengali Linux distribution. Red Hat Enterprise Linux Bengali version includes applications such as Office suite with a word processor, spread sheet, presentation tool, as well as a web browser and an e-mail client [15].

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.

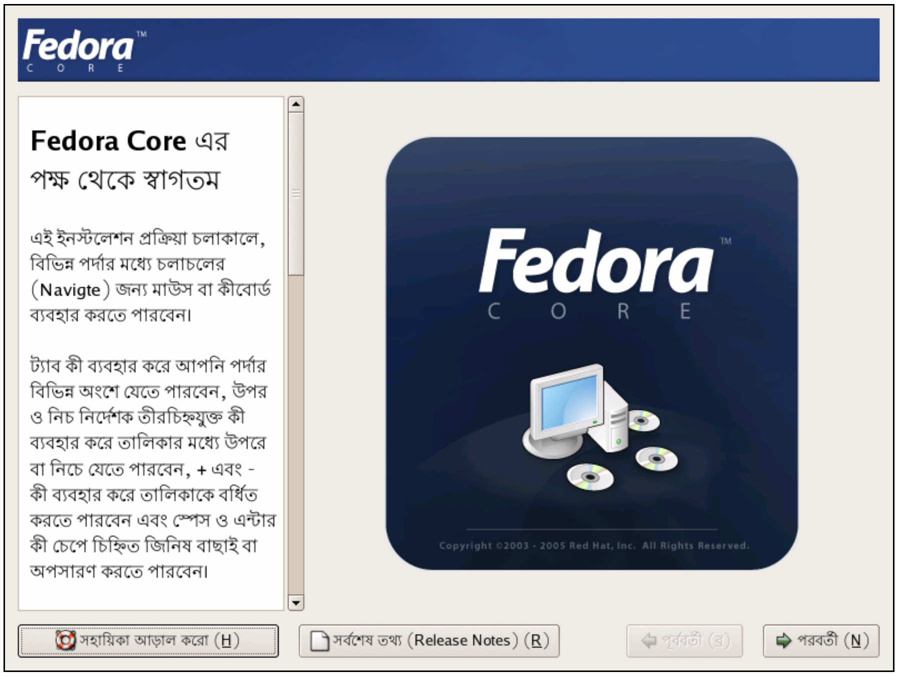


17-টি বস্তু - একটি ফাইল (মোট 106.9 কিলোবাইট) - 16-টি ফোল্ডার

(a)



(b)



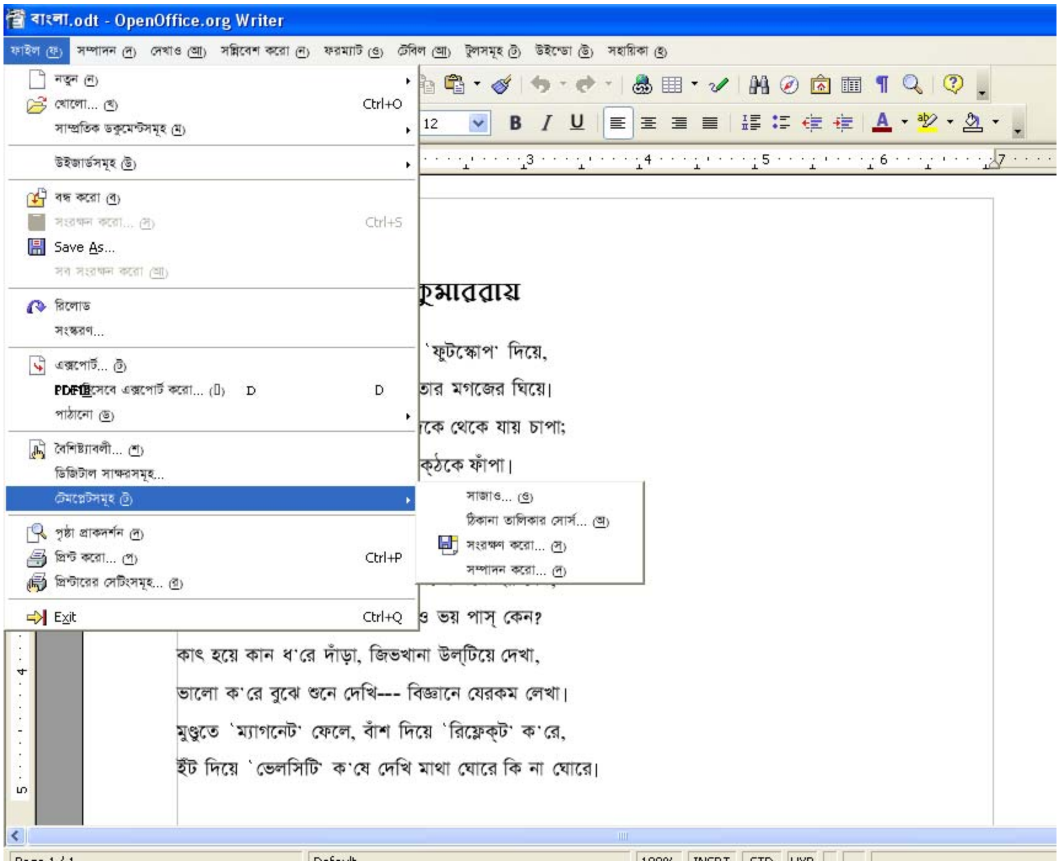
Fedora Core এর পক্ষ থেকে স্বাগতম

এই ইনস্টলেশন প্রক্রিয়া চলাকালে, বিভিন্ন পর্দার মধ্যে চলাচলের (Navigte) জন্য মাউস বা কীবোর্ড ব্যবহার করতে পারবেন।

ট্যাব কী ব্যবহার করে আপনি পর্দার বিভিন্ন অংশে যেতে পারবেন, উপর ও নিচ নির্দেশক তীরচিহ্নযুক্ত কী ব্যবহার করে তালিকার মধ্যে উপরে বা নিচে যেতে পারবেন, + এবং - কী ব্যবহার করে তালিকাকে বর্ধিত করতে পারবেন এবং স্পেস ও এন্টার কী চেপে চিহ্নিত জিনিষ বাছাই বা অপসারণ করতে পারবেন।

Fedora CORE
Fedora CORE
Copyright ©2003 - 2005 Red Hat, Inc. All Rights Reserved.
সহায়িকা আড়াল করো (H)
সর্বশেষ তথ্য (Release Notes) (R)
পূর্ববর্তী (B)
পরবর্তী (N)

(c)



কুমাররায়

`ফুটস্কোপ' দিয়ে,

তার মগজের ঘিয়ে।

কে থেকে যায় চাপা;

কঠকে ফাঁপা।

ও ভয় পাস্ কেন?

কাৎ হয়ে কান ধ'রে দাঁড়া, জিভখানা উল্‌টিয়ে দেখা,

ভালো ক'রে বুঝে শুনে দেখি--- বিজ্ঞানে যেরকম লেখা।

মুণ্ডুতে `ম্যাগনেট' ফেলে, বাঁশ দিয়ে `রিফ্লেক্ট' ক'রে,

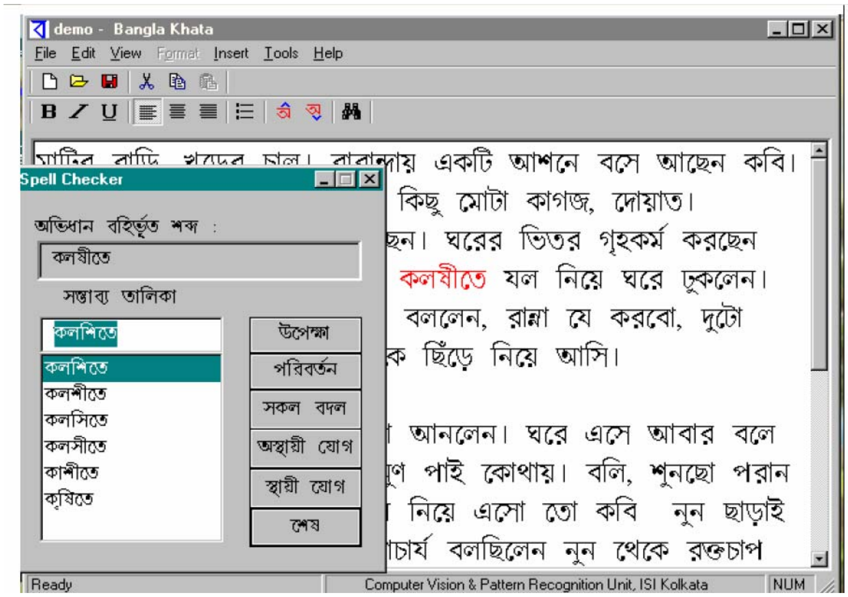
ইঁট দিয়ে `ভেলসিটি' ক'ষে দেখি মাথা ঘোরে কি না ঘোরে।

(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.


demo - Bangla Khata
File Edit View Format Insert Tools Help
B I U
Spell Checker
উপেক্ষা
পরিবর্তন
সকল বদল
অস্থায়ী যোগ
স্থায়ী যোগ
শেষ
Ready
Computer Vision & Pattern Recognition Unit, ISI Kolkata
NUM

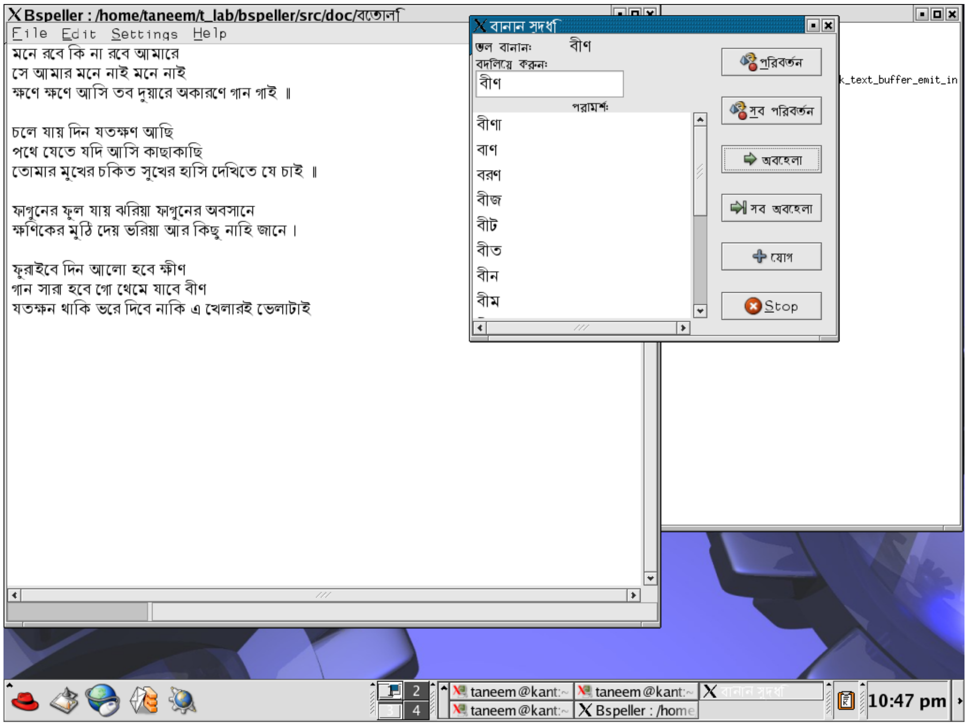
বারান্দায় একটি আশনে বসে আছেন কবি।
কিছু মোটা কাগজ, দোয়াত।
ছেন। ঘরের ভিতর গৃহকর্ম করছেন
কলষীতে যল নিয়ে ঘরে ঢুকলেন।
বললেন, রান্না যে করবো, দুটো
ক ছিঁড়ে নিয়ে আসি।
আনলেন। ঘরে এসে আবার বলে
ুণ পাই কোথায়। বলি, শুনছো পরান
নিয়ে এসো তো কবি নুন ছাড়াই
াচার্য বলছিলেন নুন থেকে রক্তচাপ

অভিধান বহির্ভূত শব্দ : কলষীতে

সম্ভাব্য তালিকা
কলশিতে
কলশিতে
কলশীতে
কলসিতে
কলসীতে
কাশীতে
কৃষিতে

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].



মনে রবে কি না রবে আমারে
সে আমার মনে নাই মনে নাই
ক্ষণে ক্ষণে আসি তব দুয়ারে অকারণে গান গাই ॥

চলে যায় দিন যতক্ষণ আছি
পথে যেতে যদি আসি কাছাকাছি
তোমার মুখের চকিত সুখের হাসি দেখিতে যে চাই ॥

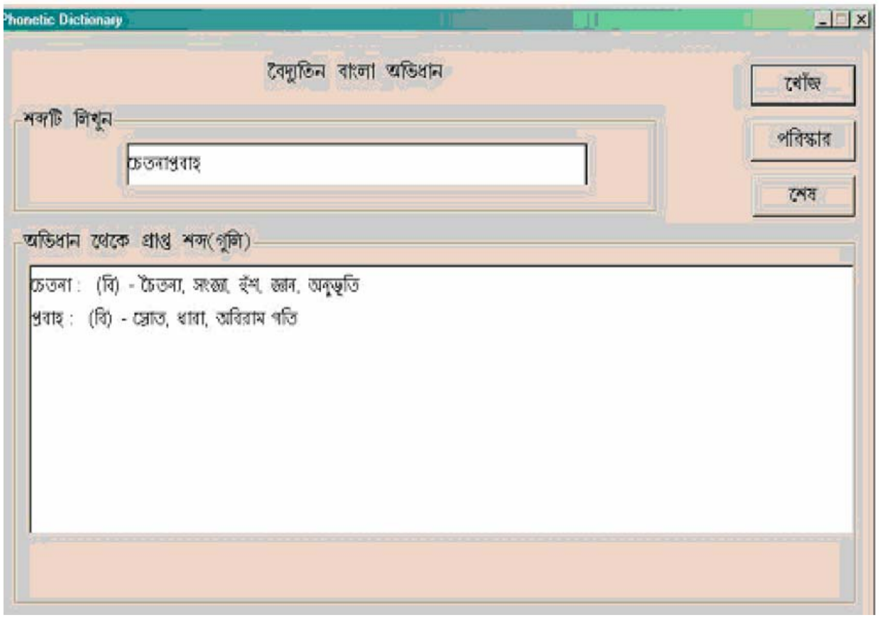
ফাগুনের ফুল যায় ঝরিয়া ফাগুনের অবসানে
ক্ষণিকের মুঠি দেয় ভরিয়া আর কিছু নাহি জানে।

ফুরাইবে দিন আলো হবে ক্ষীণ
গান সারা হবে গো থেমে যাবে বীণ
যতক্ষন থাকি ভরে দিবে নাকি এ খেলারই ভেলাটাই

ভুল বানান: বীণ
বদলিয়ে করুন: বীণ
পরামর্শ:
বীণা
বাণ
বরণ
বীজ
বীট
বীত
বীন
বীম

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].



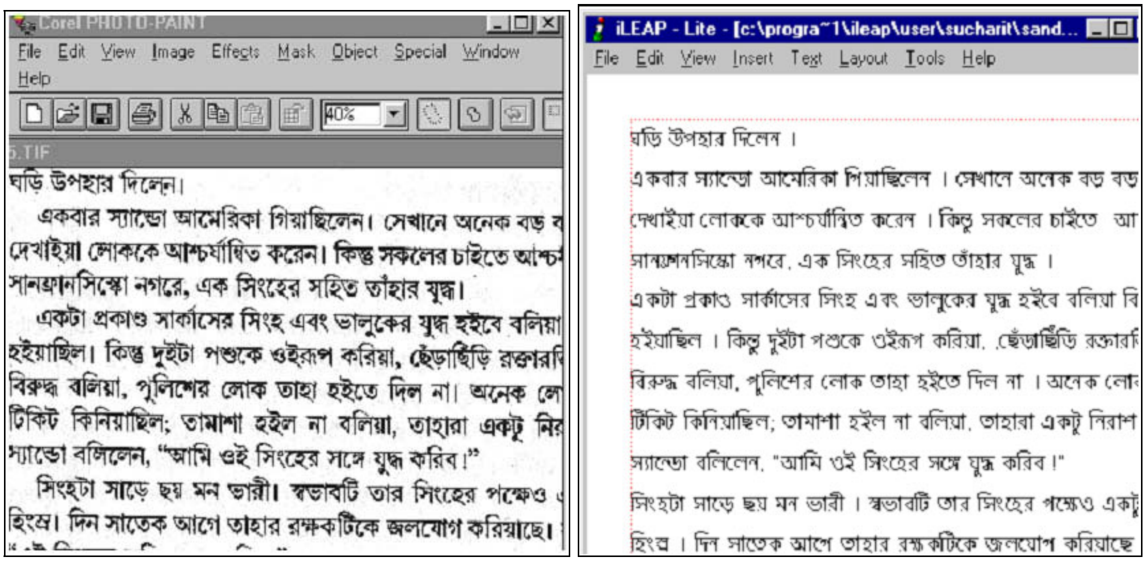
শব্দটি লিখুন: চেতনাপ্রবাহ

অভিধান থেকে প্রাপ্ত শব্দ(গুলি)

চেতনা : (বি) - চৈতন্য, সংজ্ঞা, হুঁশ, জ্ঞান, অনুভূতি
প্রবাহ : (বি) - স্রোত, ধারা, অবিরাম গতি

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.


Corel PHOTO-PAINT
File Edit View Image Effects Mask Object Special Window Help
40%
5.TIF
iLEAP - Lite - [c:\progra~1\ileap\user\sucharit\sand...
File Edit View Insert Text Layout Tools Help

ঘড়ি উপহার দিলেন।
একবার স্যান্ডো আমেরিকা গিয়াছিলেন। সেখানে অনেক বড় ব
দেখাইয়া লোককে আশ্চর্যান্বিত করেন। কিন্তু সকলের চাইতে আশ্চ
সানফ্রানসিস্কো নগরে, এক সিংহের সহিত তাঁহার যুদ্ধ।
একটা প্রকাণ্ড সার্কাসের সিংহ এবং ভালুকের যুদ্ধ হইবে বলিয়া
হইয়াছিল। কিন্তু দুইটা পশুকে ওইরূপ করিয়া, ছেঁড়াছিঁড়ি রক্তার
বিরুদ্ধ বলিয়া, পুলিশের লোক তাহা হইতে দিল না। অনেক লো
টিকিট কিনিয়াছিল; তামাশা হইল না বলিয়া, তাহারা একটু নির
স্যান্ডো বলিলেন, "আমি ওই সিংহের সঙ্গে যুদ্ধ করিব!"
সিংহটা সাড়ে ছয় মন ভারী। স্বভাবটি তার সিংহের পক্ষেও
হিংস্র। দিন সাতেক আগে তাহার রক্ষকটিকে জলযোগ করিয়াছে।

ঘড়ি উপহার দিলেন ।
একবার স্যান্ডো আমেরিকা গিয়াছিলেন । সেখানে অনেক বড় বড়
দেখাইয়া লোককে আশ্চর্যান্বিত করেন । কিন্তু সকলের চাইতে আ
সানফ্রানসিস্কো নগরে, এক সিংহের সহিত তাঁহার যুদ্ধ ।
একটা প্রকাণ্ড সার্কাসের সিংহ এবং ভালুকের যুদ্ধ হইবে বলিয়া বি
হইয়াছিল । কিন্তু দুইটা পশুকে ওইরূপ করিয়া, ছেঁড়াছিঁড়ি রক্তার
বিরুদ্ধ বলিয়া, পুলিশের লোক তাহা হইতে দিল না । অনেক লো
টিকিট কিনিয়াছিল; তামাশা হইল না বলিয়া, তাহারা একটু নিরাশ
স্যান্ডো বলিলেন, "আমি ওই সিংহের সঙ্গে যুদ্ধ করিব !"
সিংহটা সাড়ে ছয় মন ভারী । স্বভাবটি তার সিংহের পক্ষেও একটু
হিংস্র । দিন সাতেক আগে তাহার রক্ষকটিকে জলযোগ করিয়াছে

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.

Burmese

Burmese belongs to Tibeto-Burman language family and derives from Sino-Tibetan, as shown in Figure 1. It is the official language of Myanmar, where 32 million people speak it as their first language [1, 2]. Some people in China and India also speak Burmese.

Sino-Tibetan
Tibeto-Burman
Lolo-Burmese
Burmish
Southern
BURMESE [BMS]

Figure 1: Language Family Tree of Burmese [2]

Myanmar or Burmese script is used to write Burmese language. The script has been developed from the Mon script, adapted from southern Indian Pali script. The earliest known inscriptions in Burmese script date back to 11th century [3].

Character Set and Encoding

Unicode code chart 1000-109F is the internationally standardized character set encoding for Myanmar script [4] but is not frequently used. Two other ad hoc character set encoding schemes, MyaZedi (developed by Solveware Solutions) and Win/CE/Geocomp, are more frequently used at the national level [5].

Fonts and Rendering

Microsoft's support for TTF and OTF fonts is able to render Myanmar fonts but fonts shipped by Microsoft do not support Myanmar. Many Myanmar Unicode fonts have been developed by local vendors and are available, e.g. MyaZedi [6], Pdadauk (Graphite) [7], MyaZedi M17N [8] and Myanmar_OTF [8]. Work is under progress to provide support in Pango rendering engine for GNOME. Mozilla (Firefox and Thunderbird) builds are also partially available in Myanmar [17].

Keyboard

Many keyboard layouts have been developed for Burmese character set. Popularly used keyboard layouts are Win/CE/Geocomp, SCIM-KMFL–Unicode, WIN Myanmar and MyaZedi. In terms of usage WIN/CE/Geocomp and MyaZedi are used widely by business, publishers, and government agencies [9]. SCIM-M17N developed by Myanmar Development Lab is the national standard, but it is not widely used [5]. A comprehensive list is given in [9]. Figure 2 shows the MyaZedi keyboard layout.

If Unicode is used, a more complex input mechanism is required. Unicode has released a short technical note which explains these issues [14].

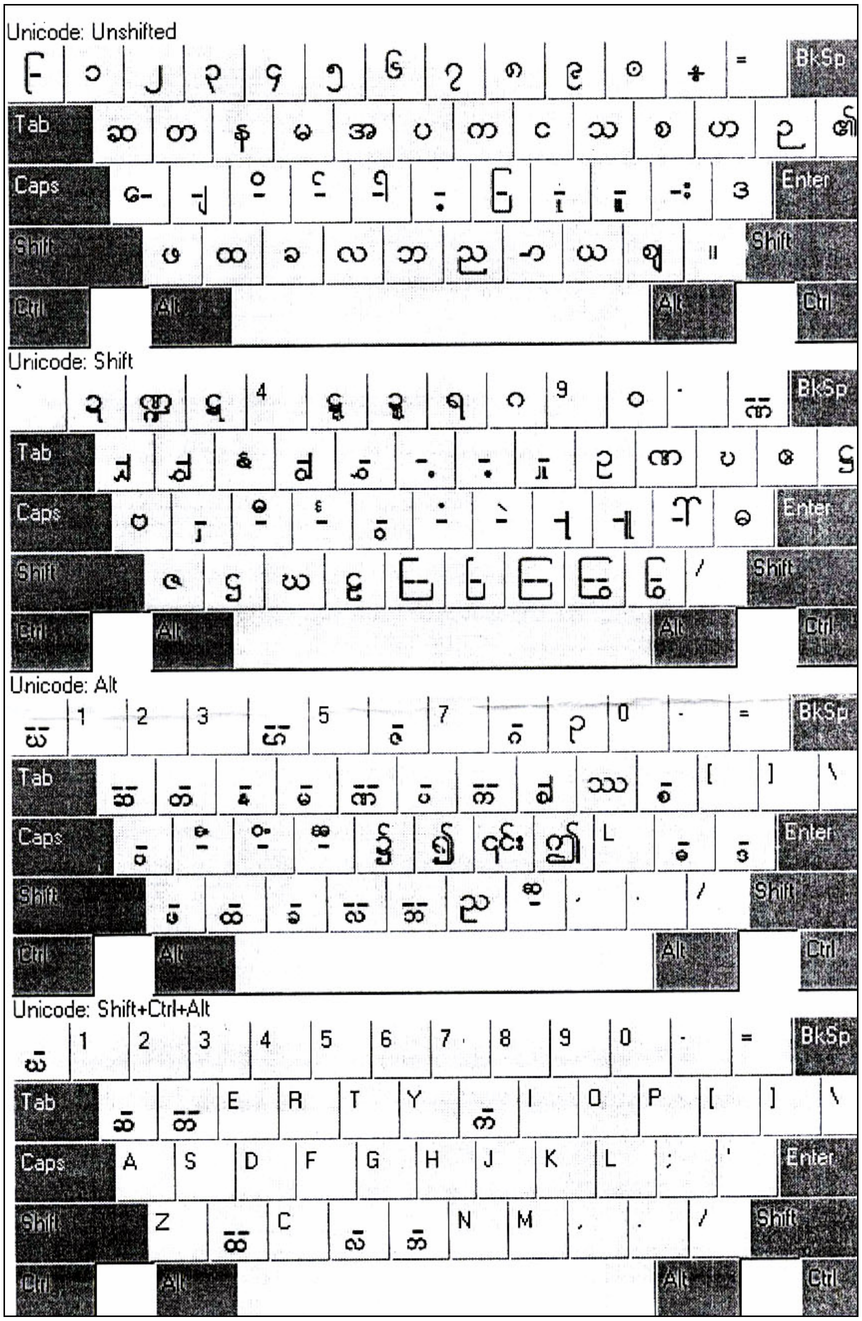


Figure 2: MyaZedi Keyboard Layout [10].

Microsoft Platform

Microsoft does not provide support for a Myanmar keyboard. However various keyboard layouts developed by local vendors can be used on Microsoft platform. A few of these keyboard layouts are based on layouts previously mentioned [9]. Following figures show the four different states of the SOAS Myanmar Keyboard layout, also available for Microsoft platform.

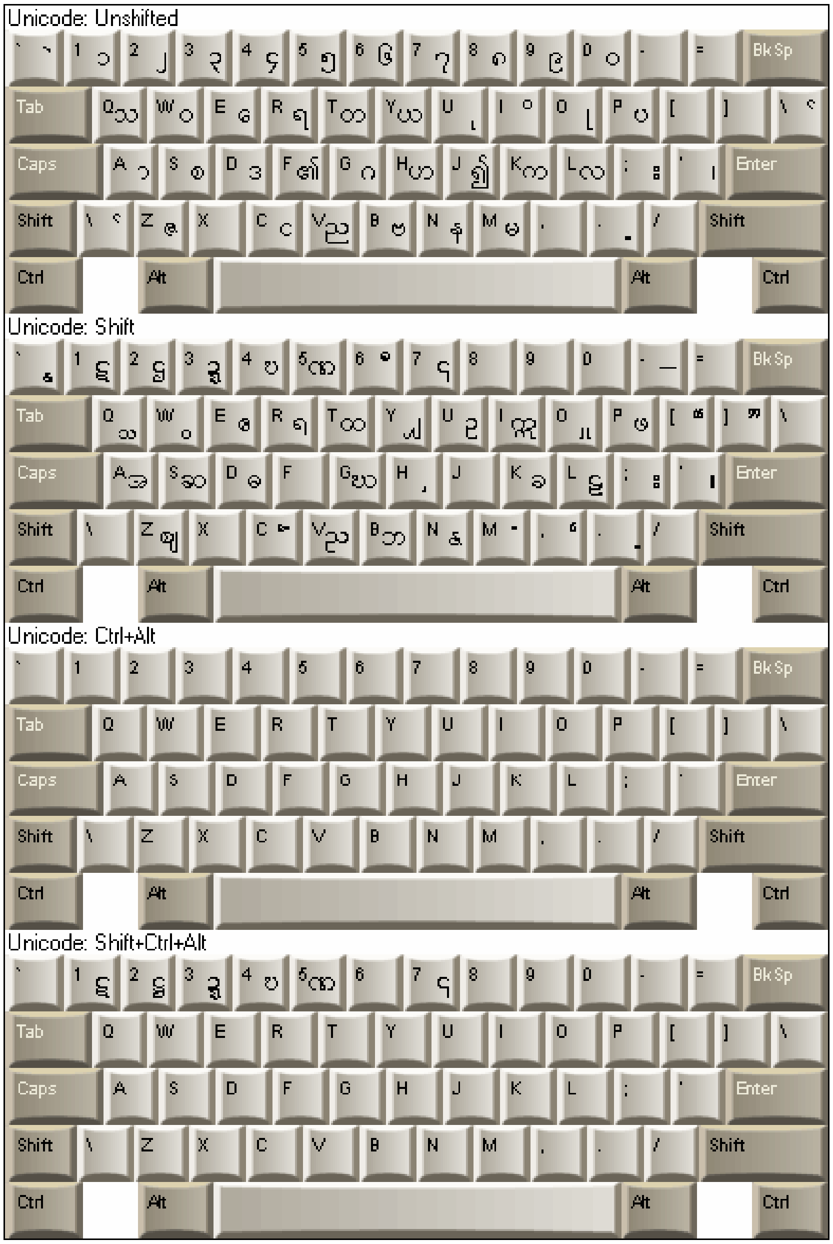


Figure 3: SOAS Myanmar Keyboard Layout [11]

Linux

SCIM-M17N keyboard (by Myanmar NLP Lab [8]) and Ava and traditional Win keyboards (by Myanmar Linux Users Group (LUG) [15]) are available for Linux but are rarely used [5].

Collation

There are two main collation sequences used for Myanmar, Pali order used for older dictionaries, and Spelling Book order used in modern dictionaries [16]. Non-Unicode fonts allow variable sequence of keystrokes to generate the same surface string, making it difficult to develop sorting sequences. However, Unicode enables a unique input sequence, on which collation can be built. Details of how to develop a collation sequence based on modern lexicographic order are available in [16]. A Myanmar collation sequence developed by Myanmar NLP has been standardized nationally but is not widely known and used yet [5].

Microsoft platform does not provide collation support for Burmese. Myanmar NLP Research Center has developed a Myanmar sorter, which can sort Myanmar text in Unicode. GeoComp has also developed a sorting engine based on GeoComp Myanmar font encoding [12].

Myanmar collation is defined in IBM ICU and Glibc for open source platforms [18].

Locale

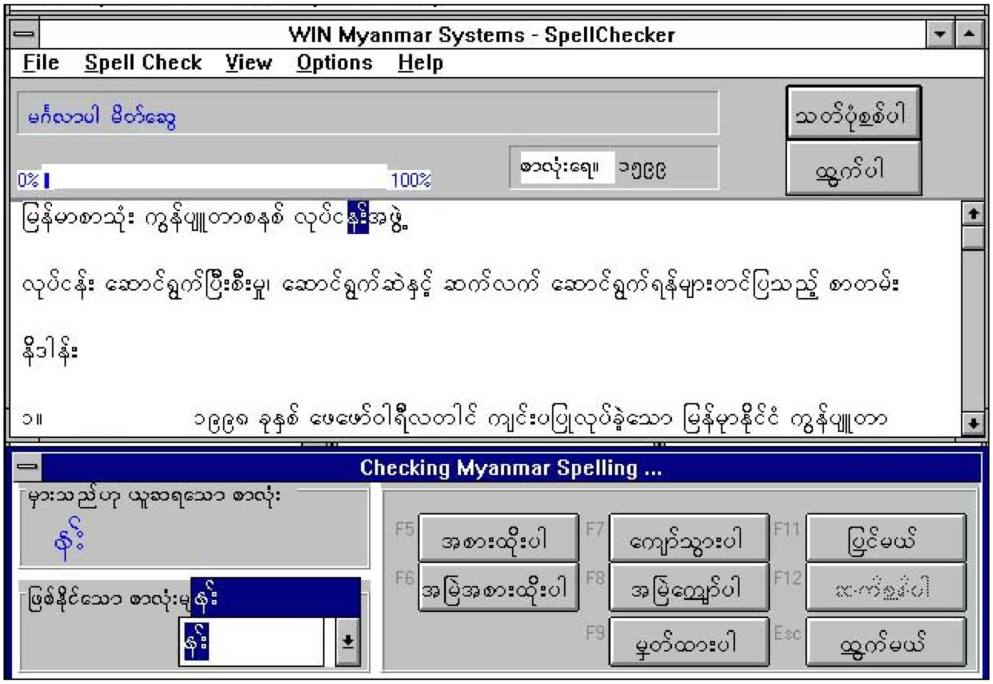
Burmese locale language name is “my” and country abbreviation is “mm” (earlier “bu” in ISO 3166). Myanmar locale data is not defined in the latest version of CLDR or IBM ICU. Microsoft does not provide support for Myanmar locale. Locale is being defined on Linux platform by Myanmar LUG and Myanmar NLP Research Center.

Interface Terminology Translation

Through the Myanmar Enabling Kit project [13], Myanmar LUG is developing interface terminology translations for the Linux based applications. Open Office 1.1.1 has already been released (with few reported errors) and work is in progress to translate applications on GNOME [17]. LIP for Myanmar on Microsoft platform has not been initiated but localized versions of Microsoft products are available from other vendors [18].

Status of Advanced Applications

Advanced local language applications for Myanmar language are being researched and developed by local vendors. Work has been done on defining line breaking [19] and developing utilities for its implementation (in Java [20]). Myanmar Spell Checker is also under development. Myanmar Unicode and NLP Research Center has released an initial version of Myanmar spell checker [5, 12]. Figure 4 shows the screen of this spell checker.



မင်္ဂလာပါ မိတ်ဆွေ

စာလုံးရေ။ ၁၅၉၉

မြန်မာစာသုံး ကွန်ပျူတာစနစ် လုပ်ငန်းအဖွဲ့

လုပ်ငန်း ဆောင်ရွက်ပြီးစီးမှု၊ ဆောင်ရွက်ဆဲနှင့် ဆက်လက် ဆောင်ရွက်ရန်များတင်ပြသည့် စာတမ်း

နိဒါန်း

၁။ ၁၉၉၈ ခုနှစ် ဖေဖော်ဝါရီလတွင် ကျင်းပပြုလုပ်ခဲ့သော မြန်မာနိုင်ငံ ကွန်ပျူတာ

မှားသည်ဟု ယူဆရသော စာလုံး: နိုး

ဖြစ်နိုင်သော စာလုံး:မ နိုး

နိုး

Figure 4: Myanmar Spell Checker [12]

Research on Myanmar OCR application has also been initiated. The current version of OCR, developed by Myanmar NLP Research Center, recognizes Myanmar digits. Work has also started on Myanmar speech recognition and text-to-speech systems [12].

References

[1] http://www.ethnologue.com/show_language.asp?code=mya
[2] http://en.wikipedia.org/wiki/Burmese_language
[3] http://www.omniglot.com/writing/burmese.htm
[4] http://www.unicode.org/charts/PDF/U1000.pdf
[5] Reported by PAN Localization project survey filled by Myanmar NLP Research Center.
[6] http://www.myazedi.com
[7] http://scripts.sil.org
[8] http://www.myanmars.net/unicode
[9] http://www.myanmars.net/unicode/keyboards/index.htm
[10] http://www.myanmars.net/unicode/doc/
[11] http://mercury.soas.ac.uk/wadict/burmese/SOASMyanmar_keyboard_and_font_user_manual.pdf
[12] http://www.myanmars.net/unicode/projects.htm
[13] http://www.iosn.net/country/myanmar/news/NLPTEAM
[14] Hosken, M and Tuntunlwin, M. “Representing Myanmar in Unicode: Details and Examples.” http://www.unicode.org/notes/tn11/myanmar_uni.pdf, 2004.
[15] http://www.thanlwinsoft.org
[16] Stribley, K. “Collation of Myanmar (Burmese) in Unicode: Sorting Myanmar in Unicode According to “Spelling Book Order” .” http://www.thanlwinsoft.org/ThanLWinSoft/MyanmarUnicode/Sorting/MyanmarCollation.pdf, 2005
[17] http://www.thanlwinsoft.org/ThanLWinSoft/MyanmarUnicode/Applications
[18] http://www.myanmars.net/winmyanmar/
[19] Stribley, K. “Syllable Based Dual Weight Algorithm for Line Breaking in Myanmar Unicode.” http://www.thanlwinsoft.org/ThanLWinSoft/MyanmarUnicode/Applications, 2005.
[20] http://www.thanlwinsoft.org/ThanLWinSoft/MyanmarUnicode/Parsing/MyanmarParser.java

Chinese

Chinese is a Sino-Tibetan language spoken by about 867 million people across the globe. One fifth of people in the world speak some dialect of Chinese. It is the national and the official language of China, Taiwan, Singapore and United Nations. Chinese is spoken in more than fifty different dialects within China and also in Brunei, Cambodia, Indonesia (Java and Bali), Laos, Malaysia (Peninsular), Mauritius, Mongolia, Philippines, Russia (Asia), Singapore, Taiwan, Thailand, United Kingdom, USA and Vietnam [1].

Figure 1: Language Family Tree for Chinese [1]

Chinese is written with characters known as Hanzi. Each Chinese character represents a syllable of spoken Chinese and also has a meaning. The characters were originally pictures of people, animals or other things, but over the centuries they have become increasingly stylized and no longer resemble the things they represent. Many characters are actually compounds of two or more characters [2]. The simplified script (Simplified Chinese) was officially developed in the People's Republic of China in 1949 in an effort to improve literacy. The simplified script is also used in Singapore but the older traditional characters are still used in Taiwan, Hong Kong, Macau and Malaysia. Further simplifications were published in 1977 but proved very unpopular and abandoned in 1986 [3].

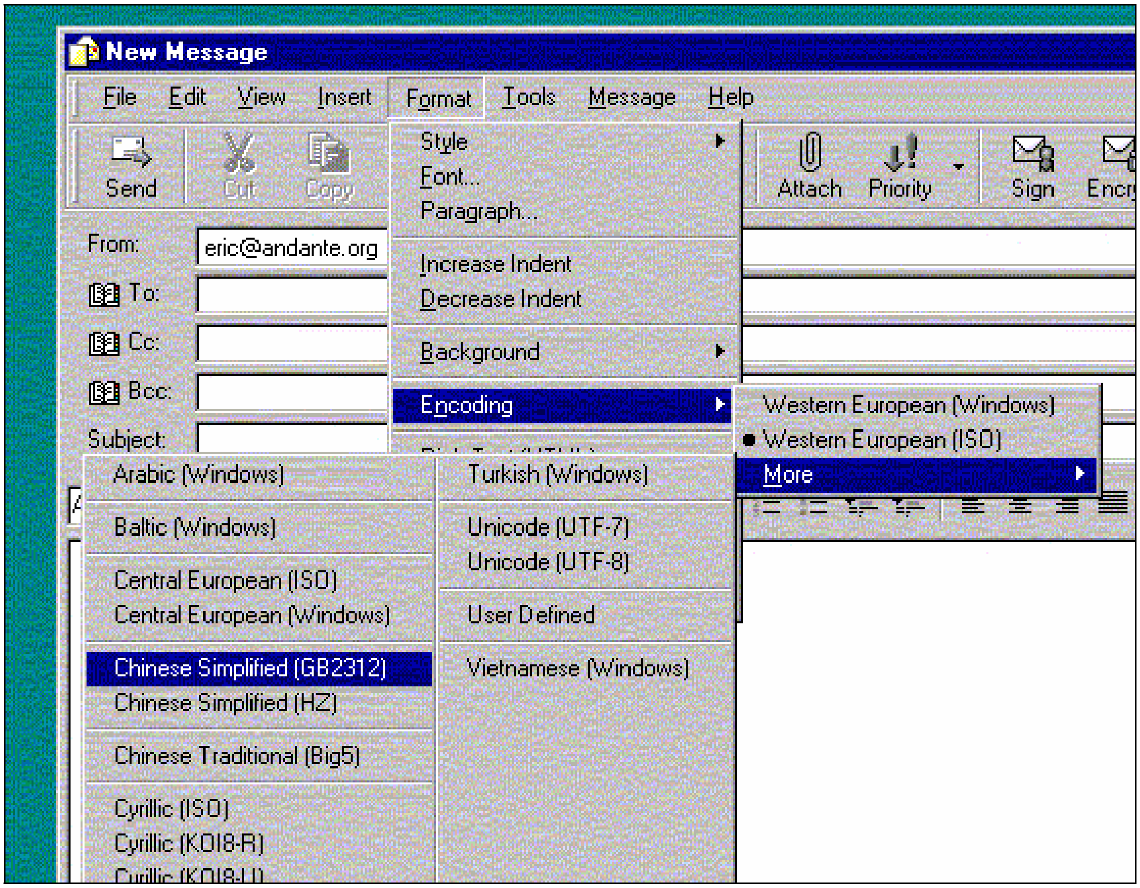
Character Set and Encoding

There are three different sets of character encodings for Chinese, (i) Guobiao code [4] for Simplified Chinese for Mainland China, (ii) Big5 for Traditional Chinese for Hong Kong and Taiwan [5], and (iii) Unicode, which combines the two Chinese forms. See [6] for an overview. Also, see [30] for more encodings.

Many different standards exist for Simplified Chinese, collectively known as Guobiao code [4]. GB 2312 is the official character set standard for China (GB abbreviates Guojia Biaozhun (national standard) in Chinese). GB 2312 (1980) includes 6,763 Chinese characters, symbols and punctuation, Japanese Kana, the Greek and Cyrillic alphabets, Zhuyin, and a double-byte set of Pinyin letters with tone marks. GB 2312 has a counterpart standard for Traditional Chinese (Fanti) forms replacing Simplified Chinese (Jianti) forms, known as GB/T 12345. GB based fonts are normally available for both forms. GB 2312 covers 99.75% of the characters used for Chinese input, but does not have effective coverage for historical texts and names, and is very widely used [7]. When Unicode 1.1 was announced with Traditional Chinese character set and Chinese characters simplified after 1980 (when GB 2312 was released), GB13000.1-93 was announced as the equivalent Chinese standard. The extra characters in these equivalent standards were also added to GB 2312 in empty slots resulting in GBK (or GB extension). This was implemented in Microsoft Windows 95 (Code Page 936), and thus became very popular. Encodings of GBK is compatible with GB 2312 but does not agree with Unicode 1.1 (or equivalent GB 13000.1-93) [9]. Eventually, GBK was also extended in 2000 to GB 18030-2000. This standard has more characters and also allows four bytes to represent a character (previously up to two bytes). It is similar to UTF-8 encoding of Unicode but has a different encoding scheme [8]. Since September 1, 2001, support of GB 18030 is mandatory for all operating systems sold in China [8].

Big5 encoding was developed by a consortium of five Taiwan based companies in 1984. This was later extended to include missing characters in 1995 by Hong Kong government. Further additions were made in 1999 for Hong Kong Supplementary character set. It is still widely used in Taiwan and Hong Kong. The extended form is sometimes referred to as Big5e [5].

ISO 2022 (or ISO IEC 2022) is also standardized for Chinese (ISO-2022-CN) which enables the characters to be represented as a sequence of 7-bit characters, similar to UTF-8 encoding mechanism for Unicode. It is different from UTF-8 because it enables switching to different encoding schemes depending on the escape sequence (while UTF-8 is limited to Unicode encoding). More details are given in [10, 11]. A related standard is EUC-CN (which is also based on ISO-2022) specifically to use GB standards [12]. Its variant HZ [13, 14] is also used.

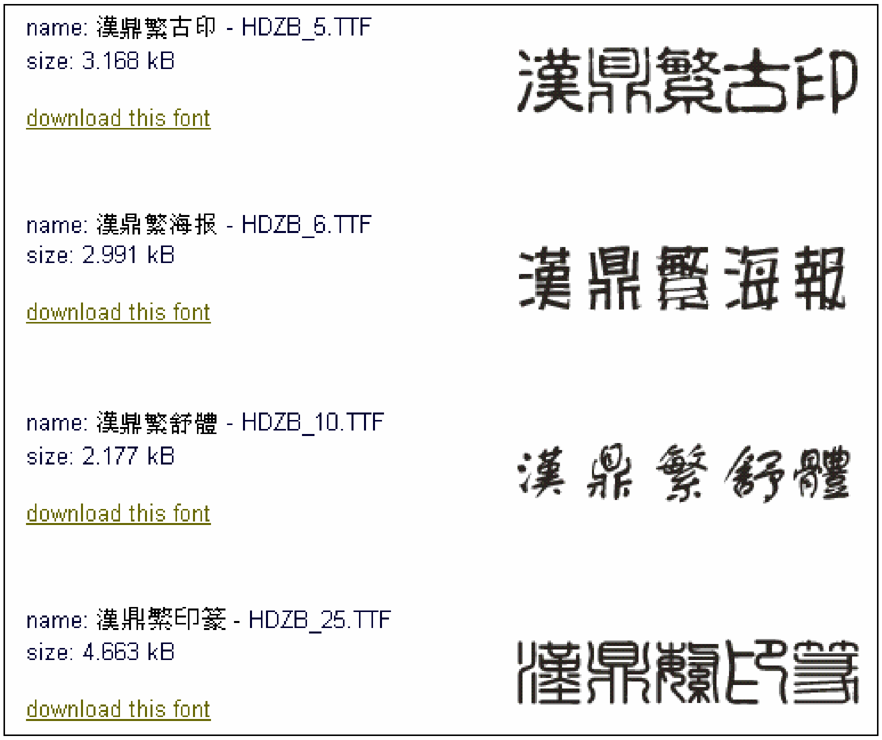


From: eric@andante.org

Figure 2: Screen Shot Showing Different Encoding Support in Applications [15]

Fonts and Rendering

Free and vendor fonts for different encodings are available for Chinese, e.g. see [3, 16]. Rendering is also supported well on many different platforms. A number of Chinese fonts are available for both Microsoft and Linux platforms. On Debian Arphic TT Chinese fonts, xfonts-intl-chinese, xfonts-cjk and Unifont, and on Red Hat taipeifonts, ttfonts-zh_CN, ttfonts-zh_TW, are among many Chinese fonts being used.



name: 漢鼎繁古印 - HDZB_5.TTF
size: 3.168 kB

download this font

name: 漢鼎繁海报 - HDZB_6.TTF
size: 2.991 kB

download this font

name: 漢鼎繁舒體 - HDZB_10.TTF
size: 2.177 kB

download this font

name: 漢鼎繁印篆 - HDZB_25.TTF
size: 4.663 kB

download this font

Figure 3: Chinese Fonts Used on Windows Platform [16]

Keyboard and Input Method Engines

With so many distinct characters, a keyboard having unique keys for them would be very large and not very efficient to use (e.g. [17]). Therefore, many input methods have been devised to input large number of Chinese characters effectively. These methods can be classified in three categories: input based on (i) encoding, (ii) pronunciation, (iii) character structure. Multiple methods in each category exist for inputting Chinese text. Going in detail of all the variety of methods is beyond the scope of this survey, however, a comprehensive overview is given at [18, 21] and associated links.

In the first category, input can be entered by directly entering the internal encoding (based on schemes discussed in the earlier section) or an indirect way of doing the same. The pronunciation based methods (e.g. the popular Pinyin method) requires input in Latin text to phonetically "spell" a word, which is then looked up in a lexicon to replace it with the appropriate Chinese character. The rules for Romanization have been defined, e.g. see [19]. Finally, structure based input methods require the character structure to be defined as a sequence of strokes and is input accordingly. For example, CKC Chinese Input system requires entry of a sequence of up to 4 digits (0 through 9), each digit representing a set of similar strokes. An internal engine determines the corresponding output character. This method only requires the numeric pad as the keyboard [20]. Extension of this uses graffiti handwriting recognition, in which the sequence of strokes allows user to see the possible input characters available. However, this method is less accurate.

Microsoft Platform

Microsoft Windows XP provides support for built-in Chinese keyboards. Once a keyboard is enabled it will facilitate Chinese typing on all Microsoft applications like Internet Explorer, Microsoft Office, Notepad, Word Pad, etc. The keyboard also depends on the choice of input method.

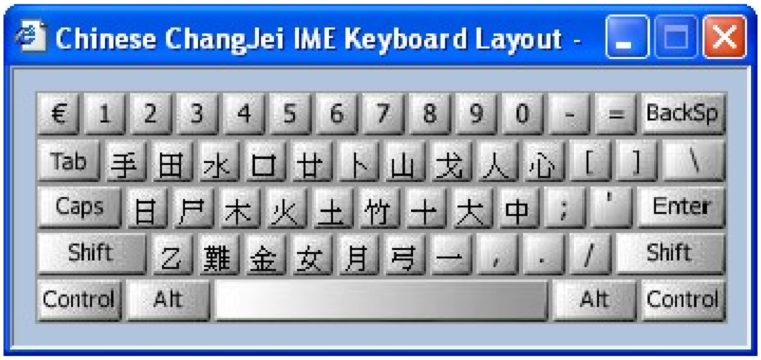

Chinese ChangJei IME Keyboard Layout -
€ 1 2 3 4 5 6 7 8 9 0 - = BackSp
Tab 手 田 水 口 廿 卜 山 戈 人 心 [ ] \
Caps 日 尸 木 火 土 竹 十 大 中 ; ' Enter
Shift 乙 難 金 女 月 弓 一 , . / Shift
Control Alt Alt Control

Figure 4: A Keyboard Layout for Chinese on Microsoft Platform

Microsoft has developed multiple IMEs for Simplified and Traditional Chinese character input [15, 21] including many of the methods listed in [20] based on multiple encoding systems, as discussed in previous sections. A detailed overview is provided in [21]. Further details are also provided in [29].

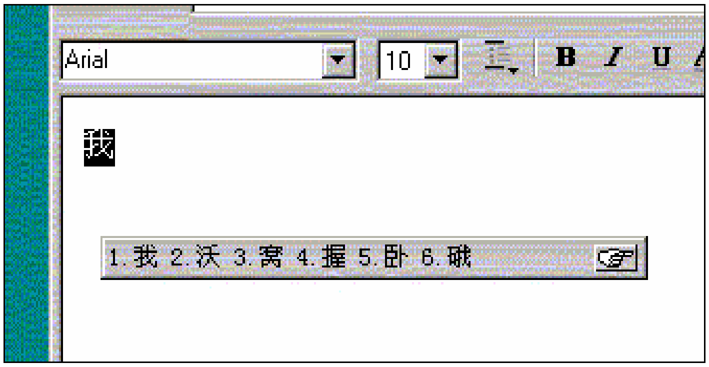
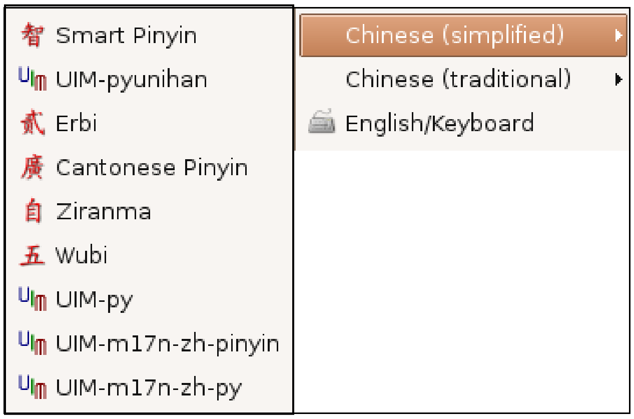


Figure 5: IME Pad for Chinese Character Input [15]

Linux Platform

A variety of input methods is also available on Linux platforms. Figure 6 below shows the available input methods available for Chinese based on Ubuntu [23].



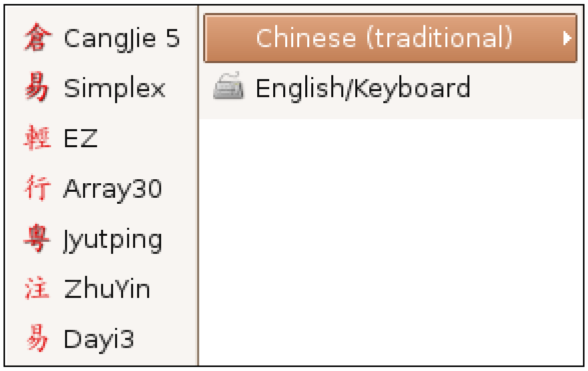


Figure 6: IME for Ubuntu CJK (Linux) for Simplified and Traditional Chinese [23]

Collation

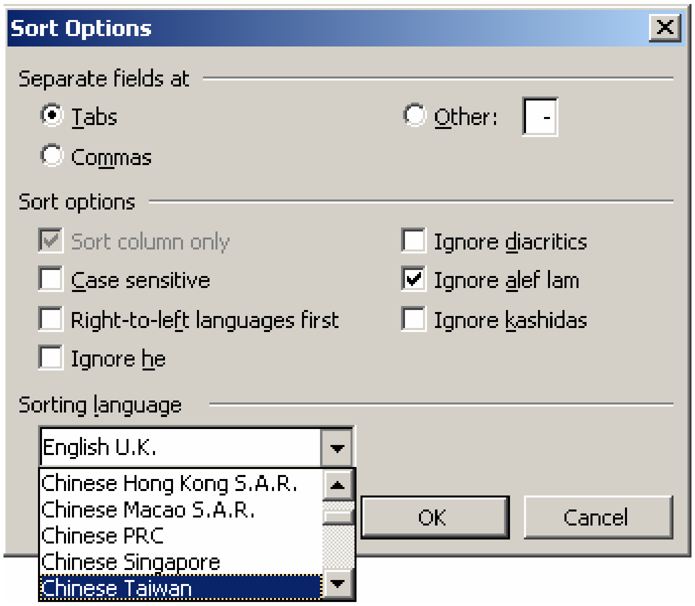
Like input methods, collation may also be done in a variety of ways, including phonetic, alphabetic (e.g. based on Latin input), stroke based or dictionary based, for example, Chinese Big5 order, PRC Chinese Phonetic order, Chinese Unicode order, PRC Chinese Stroke Count order and Traditional Chinese Bopomofo order [24]. Many of these methods are available on various platforms, including Linux and Microsoft. Also see [30] for a list of collation possibilities.

Windows XP provides built in support for sorting Chinese characters. The table below shows the Chinese collation options available on SQL Server.

Chinese (Taiwan)	0x30404	Chinese_Taiwan_Bopomofo	950
Chinese (Taiwan)	0x404	Chinese_Taiwan_Stroke	950
Chinese (People's Republic of China)	0x804	Chinese_PRC	936
Chinese (People's Republic of China)	0x20804	Chinese_PRC_Stroke	936
Chinese (Singapore)	0x1004	Chinese_PRC	936

Table 1: Locale, Local ID, Collation Designator and Code Page for Windows [25]

Figure 7 shows the Chinese sorting options in Windows XP for different locales.



Sort Options

Separate fields at

● Tabs ○ Other: -

○ Commas

Sort options

☑ Sort column only

□ Case sensitive

□ Right-to-left languages first

□ Ignore he

□ Ignore diacritics

☑ Ignore alef lam

□ Ignore kashidas

Sorting language

English U.K.

- Chinese Hong Kong S.A.R.
- Chinese Macao S.A.R.
- Chinese PRC
- Chinese Singapore
- Chinese Taiwan

Figure 7: Sort Options for Chinese in Office 2003

Locale

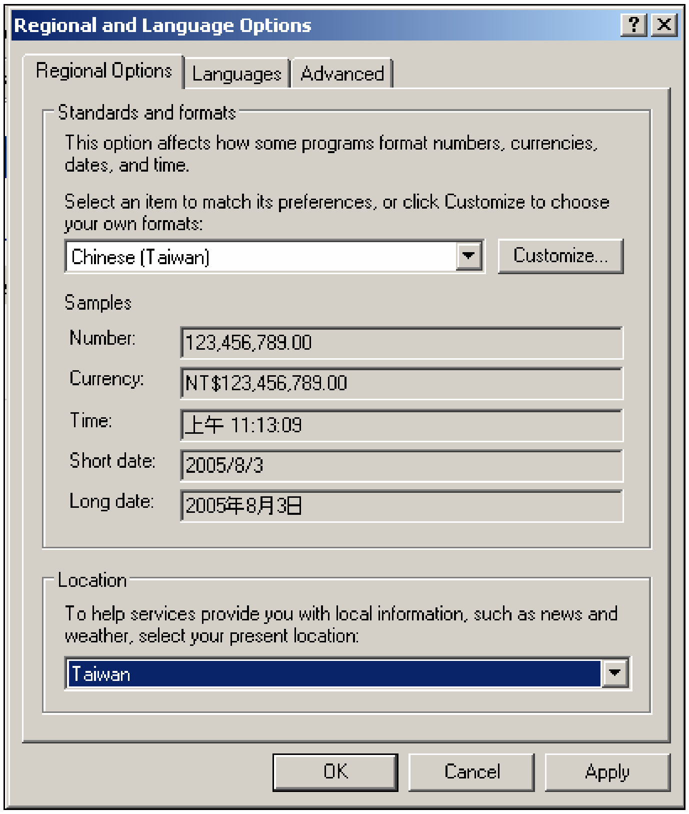
Locales for both Traditional and Simplified Chinese have been defined in IBM ICU [27]. Locales for Chinese are also available in the CLDR1.3. The locales are zh_CN (for China), zh_TW (for Taiwan), zh_HK (for Hong Kong) and zh_SP (for Singapore) (e.g. see [26, 30]).

Microsoft Platform

Microsoft provides support for Chinese locales in Windows. If the system locale is switched to Chinese, changes in date, time, and currency symbols is observed in all application of Microsoft. Figure 8 below shows Chinese settings for long date format and currency symbol within the Windows XP locale settings. Figure 7 above lists the locales available for Chinese.

Linux Platform

Chinese locale is available for all Chinese distributions for Linux, which also include KDE, GNOME, Mozilla and Open Office.



Standards and formats

This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Chinese (Taiwan)

Samples

Number:	123,456,789.00
Currency:	NT$123,456,789.00
Time:	上午 11:13:09
Short date:	2005/8/3
Long date:	2005年8月3日

Location

To help services provide you with local information, such as news and weather, select your present location:

Taiwan

Figure 8: Chinese (Taiwan) Locale on Windows XP

Interface Terminology Translation

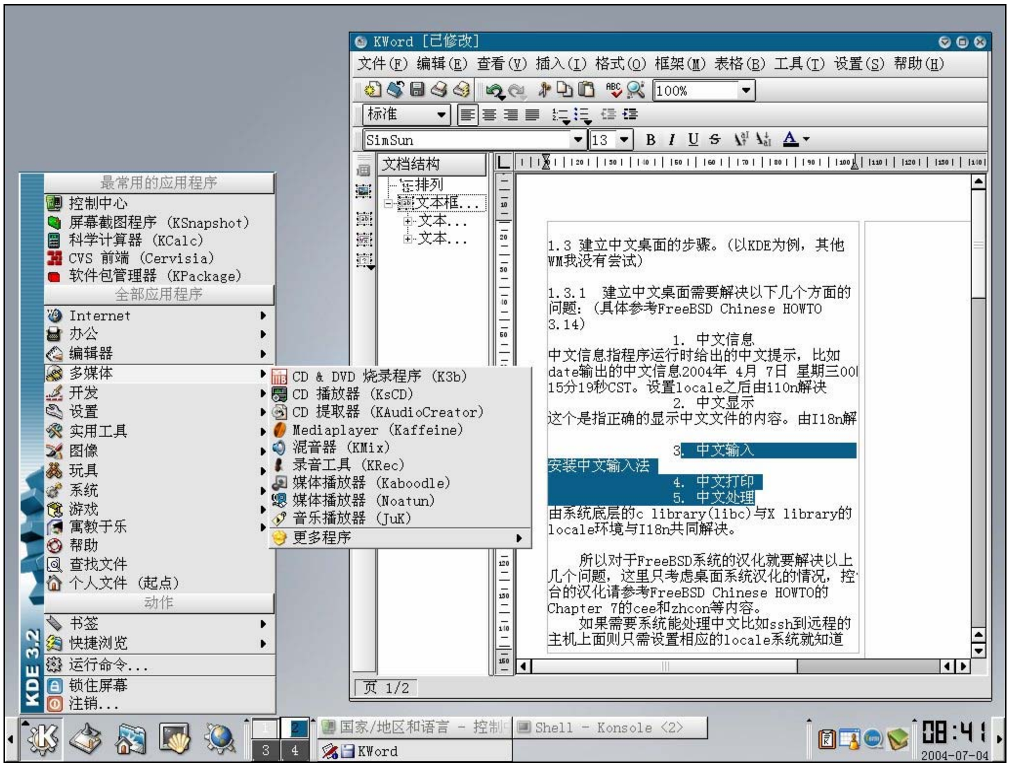
Interface terminology is available for Chinese on multiple platforms.

Microsoft Platform

Microsoft provides full support for Chinese language. The original version comes with Chinese fonts, locale, keyboard and input methods on top of which the MUI pack can be run to obtain Chinese interface [28, 31].

Linux Platform

A project on translation of KDE has been initiated. This group has recently started out and presently (July 2005) 34% KDE is translated for Traditional Chinese and 85.69% KDE is translated for Simplified Chinese [32]. For GNOME 99.23% strings have been translated for Simplified Chinese and 96.66% strings have been translated for Traditional Chinese [33]. A completely localized Chinese version of Open Office is also available to run on Microsoft Windows and Linux [34]. FireFox 1.0.6 is available in Chinese [35]. The figures given below show the localized interfaces of KDE, Mozilla and Mandarake in Chinese [36].



1.3 建立中文桌面的步骤。(以KDE为例，其他WM我没有尝试)

1.3.1 建立中文桌面需要解决以下几个方面的问题：(具体参考FreeBSD Chinese HOWTO 3.14)

1. 中文信息

中文信息指程序运行时给出的中文提示，比如date输出的中文信息2004年 4月 7日 星期三00l 15分19秒CST。设置locale之后由i10n解决

2. 中文显示

这个是指正确的显示中文文件的内容。由I18n解

3. 中文输入

安装中文输入法

4. 中文打印

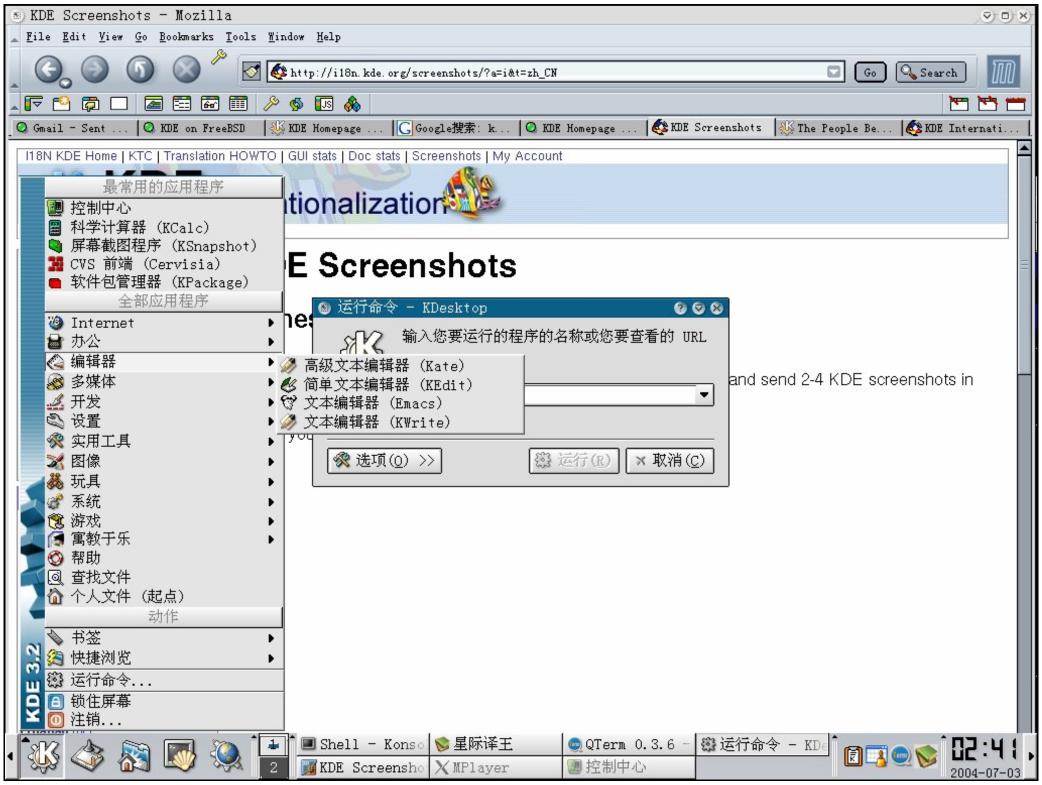
5. 中文处理

由系统底层的c library(libc)与X library的locale环境与I18n共同解决。

所以对于FreeBSD系统的汉化就要解决以上几个问题，这里只考虑桌面系统汉化的情况，控台的汉化请参考FreeBSD Chinese HOWTO的Chapter 7的cee和zhcon等内容。

如果需要系统能处理中文比如ssh到远程的主机上面则只需设置相应的locale系统就知道

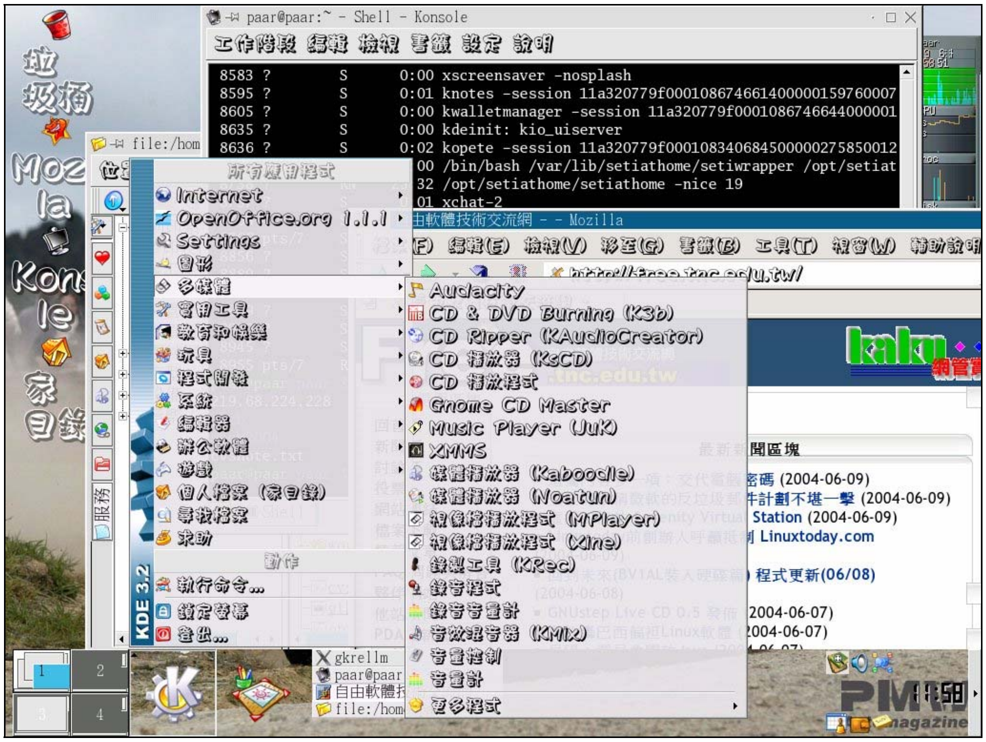
Figure 9: KDE K-Word for Simplified Chinese [36]



输入您要运行的程序的名称或您要查看的 URL

and send 2-4 KDE screenshots in

Figure 10: Mozilla Browser [36]



```
8583 ?        S      0:00 xscreensaver -nosplash
8595 ?        S      0:01 knotes -session 11a320779f000108674661400000159760007
8605 ?        S      0:00 kwalletmanager -session 11a320779f0001086746644000001
8635 ?        S      0:00 kdeinit: kio_uiserver
8636 ?        S      0:02 kopete -session 11a320779f00010834068450000027585001 2
                     00 /bin/bash /var/lib/setiathome/setiwrapper /opt/setiat
                     32 /opt/setiathome/setiathome -nice 19
                     01 xchat-2
```

Figure 11: KDE Traditional Chinese [36]

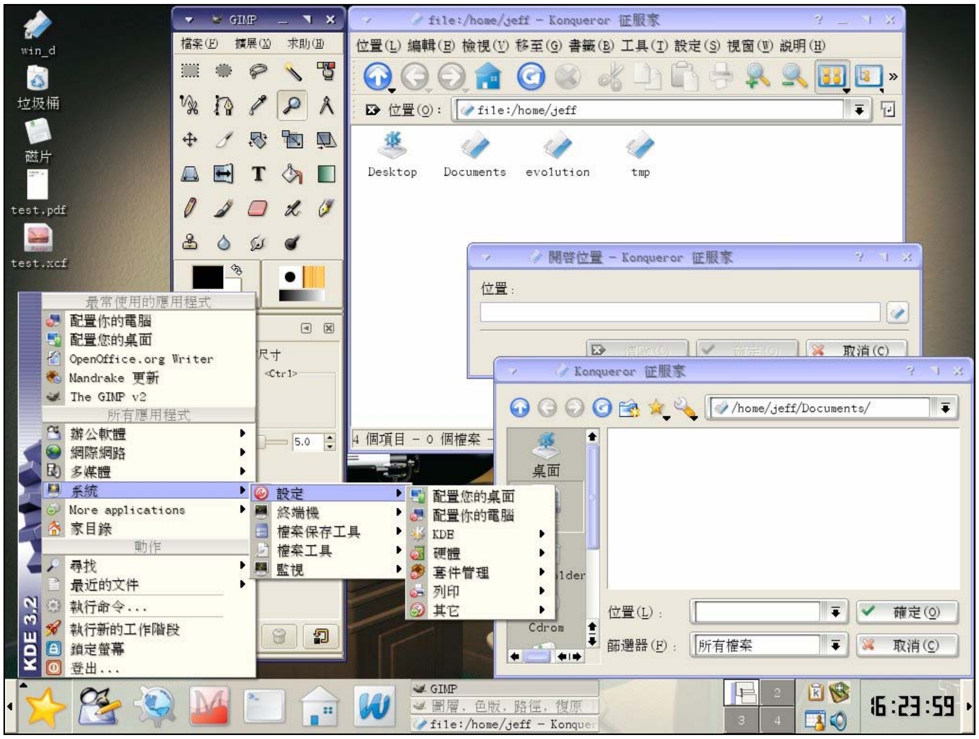


Figure 12: Simplified Chinese Mandrake [36]

Status of Advanced Applications

There has been significant work on Chinese language processing. Chinese and multilingual corpora and dictionaries are available (e.g. see [38, 39, 49]). There is also significant work done on Chinese word and line segmentation, spell checkers, and machine translation software. The work on MT continues to-date. Chinese to English MT software include CITAC, TransWhiz and Systran Professional Premium. Some Chinese to English MT web services are Babelfish, Transtar, Amikai, Netat and WorldLingo. English to Chinese MT web services include IBM AlphaWorks, Gist-in-Time System, Babelfish, ReadWorld, EWTransLite and WorldLingo [40]. Also see [41].

PenPower Chinese OCR Pro is a commercial product. It recognizes hand-written and print characters [42]. Other available OCRs are Han Wang Chinese OCR [43], SunmiPage ScanInsert Chinese OCR [44] and Dan Ching Gold [45].

MSRA Chinese text-to-speech (TTS) system is a product developed by MSRAsia Speech Group. MSRA is conducting research in voice technologies including speech recognition, speech synthesis and speech enabled information search in Asia [46]. Other TTS systems include those developed by Bell Labs [47], and IBM [48]. Similarly, significant work is being done on Chinese speech recognition and speech to speech translation (e.g. see [50, 51]).

Chinese online handwriting recognition systems including Twin-Bridge, Pen Power and QuickStroke systems are also available (see [37]).

References

[1] http://www.ethnologue.com/show_language.asp?code=cmn
[2] http://www.omniglot.com/writing/chinese.htm
[3] http://www.geocities.com/dtmcbride/tech/charsets/chinese.html
[4] http://en.wikipedia.org/wiki/Guobiao_code
[5] http://en.wikipedia.org/wiki/Big5
[6] http://en.wikipedia.org/wiki/Chinese_character_encoding
[7] http://en.wikipedia.org/wiki/GB2312
[8] http://en.wikipedia.org/wiki/GB18030
[9] http://en.wikipedia.org/wiki/GBK
[10] http://en.wikipedia.org/wiki/ISO_2022
[11] ftp://sunsite.uio.no/pub/rfc/rfc1922.txt
[12] http://en.wikipedia.org/wiki/Extended_Unix_Code
[13] http://en.wikipedia.org/wiki/HZ_%28character_encoding%29
[14] http://www.cse.ohio-state.edu/cgi-bin/rfc/rfc1843.html
[15] http://www.andante.org/ime.html
[16] http://www.sino.uni-heidelberg.de/edv/sinopc/chinese_fonts.htm
[17] http://en.wikipedia.org/wiki/Image:Large_chinese_keyboard.jpg
[18] http://en.wikipedia.org/wiki/Chinese_input_methods_for_computers
[19] http://en.wikipedia.org/wiki/Pinyin
[20] http://en.wikipedia.org/wiki/CKC_Chinese_Input_System
[21] http://zsigri.tripod.com/fontboard/cjk/input.html#chinese
[22] http://www.microsoft.com/globaldev/handson/user/IME_Paper.mspx
[23] http://www.mrbass.org/linux/ubuntu/scim/
[24] http://www.simple-sw.com/collation.htm
[25] http://msdn.microsoft.com/library/default.asp?url=/library/en-us/instsql/in_collation_6gfn.asp
[26] http://www.mpi-sb.mpg.de/~pesca/locales.html
[27] http://www-950.ibm.com/software/globalization/icu/demo/locales/en/?_=gu&d_=en&_l=zh
[28] http://www.microsoft.com/office/editions/prodinfo/language/availability.mspx
[29] http://www.microsoft.com/downloads/details.aspx?FamilyID=B91AC197-FFA7-45A7-B1E1-C3457E1B0C1F&displaylang=EN

[30] http://www.uwm.edu/cgi-bin/IMT/wwwman?topic=Chinese(5)&msection=
[31] http://www.microsoft.com/downloads/details.aspx?FamilyID=DEB6731A-CA36-47A3-B4A0-2A1D19EEFA05&displaylang=EN
[32] www.GNOME.org
[33] www.KDE.org
[34] http://zh.openoffice.org/
[35] http://moztw.org/
[36] http://i18n.KDE.org/screenshots/
[37] http://www.cgcmall.com/SearchResults.asp?Cat=2
[38] http://www.ldc.upenn.edu/Catalog/
[39] http://www.pristine.com.tw/lexicon.php
[40] http://www.chinesecomputing.com/nlp/mt.html
[41] http://www.worldlingo.com/en/products_services/worldlingo_translator.html
[42] http://www.china-guide.com/software/ocr.html
[43] http://www.chinesesoftware.com/d_hanwang_ocr.html
[44] http://www.cyberway.com.sg/~computek/sicapp.htm
[45] http://www.worldlanguage.com/Products/18.htm
[46] https://research.microsoft.com/speech/tts.asp
[47] http://www.bell-labs.com/project/tts/mandarin.html
[48] http://www.research.ibm.com/tts/
[49] http://www.d-ear.com/CCC/
[50] http://office.microsoft.com/en-us/assistance/HA010347511033.aspx
[51] http://www.slt.atr.jp/slc-e/

Dzongkha

Dzongkha is a Sino-Tibetan language related to Tibetan. It has 0.13 million first-language speakers [1] and approximately 0.5 million total speakers [3] in Bhutan. Dzongkha is the native language of eight western districts of Bhutan (Thimphu, Paro, Punakha, Wangdue, Phodrang, Gasa, Ha, Dhakana, and Chukha) and also recognized as the national and official language of the country. Dzongkha speakers also reside in India (specifically West Bengal) and Nepal [2].

- Sino-Tibetan
  - Tibeto-Burman
    - Himalayish
      - Tibeto-Kanauri
        - Tibetic
          - Tibetan
            - Southern
              - DZONGKHA

Figure 1: Language Family Tree of Dzongkha [1]

Dzongkha is written in Tibetan script, which was modeled on Devanagari script [3].

Character Set and Encoding

Dzongkha character set was not standardized prior to the release of Unicode 4.0. Unicode block 0F00-0FFF is the standard character set encoding used for Tibetan script in computers [4], which is also accepted as national as well as international standard for encoding Dzongkha text [5].

Fonts and Rendering

Tibetan script is complex and not possible to implement using True Type fonts. Open Type fonts have been developed for Tibetan script for Tibetan and Dzongkha by different people and organizations. These fonts include fonts developed by Department of IT and Dzongkha Development Authority in Bhutan, e.g. Tsuyig, Joyig, Tashi, xTashi, Uchen, Wangdi, [5, 6, 7].

Microsoft Platform

Until recently Dzongkha was not supported on the Microsoft platform. However, the latest version of Uniscribe (USP10.dll, version 1.453.3665.0) supports layout tables for Tibetan script. This version is shipped with Office 2003 Service Pack 1. Inclusion of Tibetan in the latest version of Uniscribe has facilitated typing and web-browsing in Tibetan script for Dzongkha.

Microsoft does not ship fonts for Tibetan script. However, third-party Tibetan script fonts can be used on Microsoft platform for Dzongkha text input and display [6, 8]. Rendering results of some of these fonts are shown in Figure 2 below.

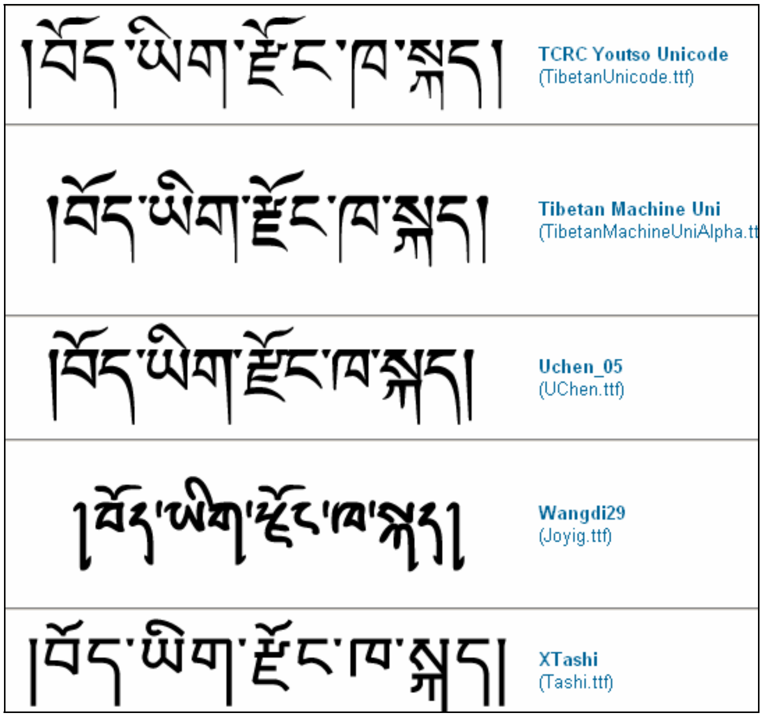


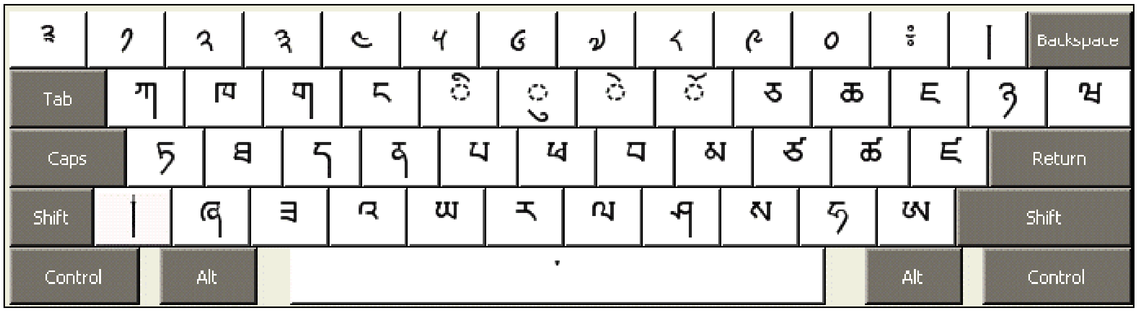
Figure 2: Unicode Dzongkha Fonts on Microsoft Platform [6]

Linux Platform

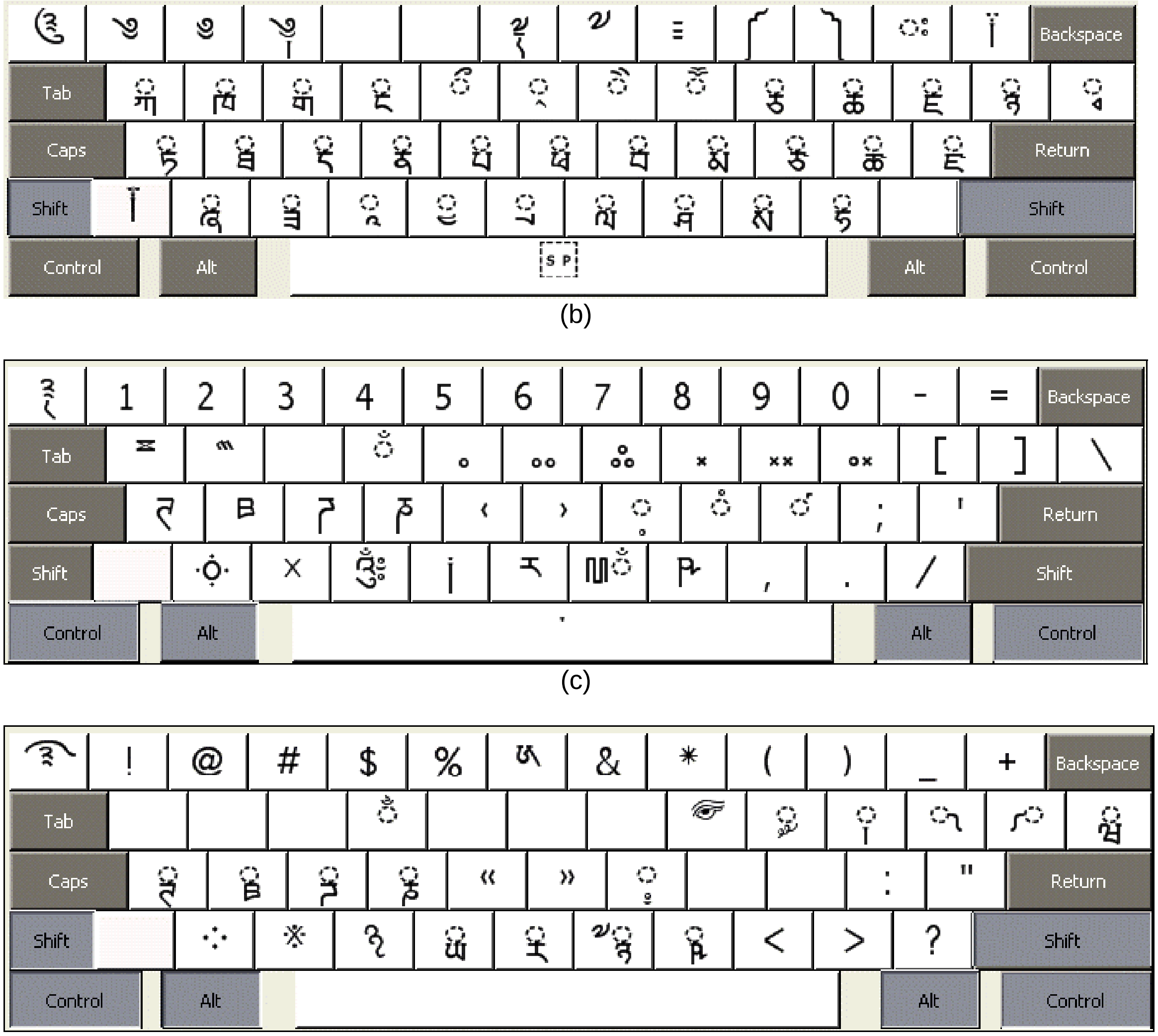
Dzongkha Development Authority, Department of Information Technology and Sherubste College are working together to enable Dzongkha computing on Linux operating system [9, 10] through the PAN Localization project [12]. Up till now the project has developed support for inclusion and rendering of Dzongkha Open Type fonts in X-Windows, Red Hat Linux, Fedora Core2 and Open Office 2.0 (though some technical challenges are still faced). Dzongkha Open Type fonts developed have also been successfully rendered through Pango in GNOME [11].

Keyboard

The Royal Government of Bhutan has nationally standardized a keyboard layout for Dzongkha. This has been designed by the Dzongkha Development Authority following consultation with all the major Dzongkha users and the Department of Information Technology. Figure 3 shows the standard keyboard layout [5].



(a)



(b)

(c)

(d)

Figure 3: Standardized Dzongkha Keyboard in (a) Normal, (b) Shift, (c) Alt+Ctrl and (d) Alt-Ctrl-Shift States [5]

Microsoft Platform

Based on layout designed jointly by the Dzongkha Development Authority (DDA) and Department of Information Technology (DoIT), Royal Government of Bhutan, Tibetan and Himalayan Digital Library (THDL) project has created keyboards for Dzongkha using MSKLC for Microsoft platform. It can be used to input Dzongkha or Tibetan Unicode text [13, 15].

Linux Platform

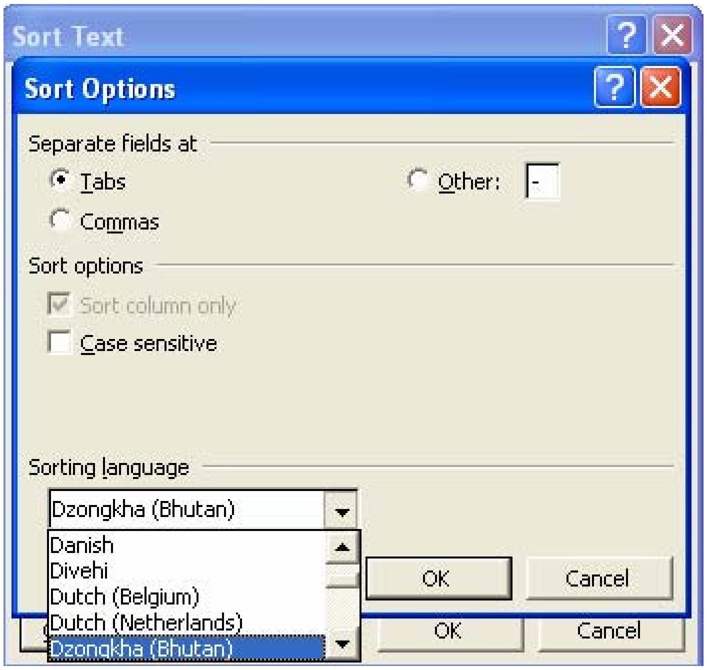
Keyboard support for Dzongkha on Linux platform has also been developed using the standard and is being distributed with the Linux distribution being developed by Department of IT through PAN Localization project [12].

Collation

Collation rules are being finalized through collaboration of Dzongkha Development Authority (DDA) and Department of IT in Bhutan. They are based on the dictionary published by DDA.

Microsoft Platform

Microsoft lists Dzongkha in the Sort Options in its latest releases, but the sort is based on DUCET. Thus, Dzongkha sort is not realized. Figure 4 shows the Sort Options on MS Office 2003.



Separate fields at

● Tabs ○ Other: -

○ Commas

Sort options

☑ Sort column only

☐ Case sensitive

Sorting language

Dzongkha (Bhutan)

Danish
Divehi
Dutch (Belgium)
Dutch (Netherlands)
Dzongkha (Bhutan)

Figure 4: Sort Options for Dzongkha on Microsoft Platform

Linux Platform

Collation rules developed by DDA have been implemented on Linux platform. They are supported in Dzongkha version of Open Office 2.0 developed by Department of IT of Government of Bhutan, through PAN Localization project [12].

Locale

A nationally standardized locale definition for Dzongkha for Bhutan (dz_BT) has been compiled by Dzongkha Development Authority (DDA) in consultation with major computer vendors and language experts. Dzongkha (also known as Bhutani) is now included as a distinct language/culture within ISO 639, with the language codes "dz" and "dzo" [16]. Dzongkha locale definition has been included in CLDR 1.3 in 2004. The locale definitions described include date, time, calendar formats, name of days, months and numbers, etc.

Microsoft Windows XP does not include locale definition for Dzongkha.

Through the research efforts of Bhutan team of PAN Localization project, Dzongkha locale is now supported on Linux platform [11]. Locale for GNU C library has been created and implemented in

Linux operating system. Support for locale and collation rules has also been added to Open Office 2.0.

Interface Terminology Translation

Microsoft Platform

Dzongkha version of Microsoft is currently not available. However development of a localized version of Windows in Dzongkha is on the short term agenda of Microsoft [14].

Linux Platform

Glossary translation of KDE for Dzongkha has been initialized but there is still no significant progress [17]. However, GNOME desktop is complete and work is underway for complete translation of Open Office through PAN Localization project [18].

Status of Advanced Applications

There is little progress on the development of advanced applications in Dzongkha. Most of the work under progress is on localization of Linux, specifically GNOME and Open Office platforms.

References

[1] http://www.ethnologue.com/show_language.asp?code=dzo
[2] http://en.wikipedia.org/wiki/Dzongkha
[3] http://www.omniglot.com/writing/tibetan.htm
[4] http://www.unicode.org/charts/PDF/U0F00.pdf
[5] "Technology Standards and Resources for Computing in Dzongkha." Department of IT, Royal Govt. of Bhutan. http://www.dit.gov.bt/guidelines/dzongkhastandard.pdf, 2004.
[6] http://salrc.uchicago.edu/resources/fonts/tibetanfonts.html
[7] http://www.dit.gov.bt/downloads/dzongkhafonts.zip
[8] Fynn, C. and Garson, T. "Tibetan fonts." http://iris.lib.virginia.edu/tibet/xml/show.php?xml=/tools/tibfonts.xml
[9] http://www.iosn.net/country/bhutan/news/dzongkha-on-linux
[10] http://sourceforge.net/projects/dzongkha/
[11] http://dzongkha.sourceforge.net/
[12] www.PANL10n.net
[13] http://iris.lib.virginia.edu/tibet/tools/dzkeyboard.html
[14] http://archives.cnn.com/2002/BUSINESS/asia/08/06/bhutan.windows/
[15] http://iris.lib.virginia.edu/tibet/tools/dzkeylayout.html
[16] "Oficial Nacional Standard of Dzongkha-Bhutan Locale." Dzongkha Development Authority. http://www.dit.gov.bt/guidelines/locale_culture.pdf, 2004.
[17] http://i18n.KDE.org/teams/index.php?a=i&t=dz
[18] http://l10n.openoffice.org/languages.html

Farsi (Persian)

Farsi belongs to Iranian branch of Indo-European language family. Persian or Western Farsi is spoken by about 22 million people residing in Iran. A close variant, Dari (Eastern Farsi), is spoken by seven million people in Afghanistan, Iran and Pakistan. Both varieties are also spoken in some other countries as well [1, 2]. Figure 1 shows the language family tree of Farsi.

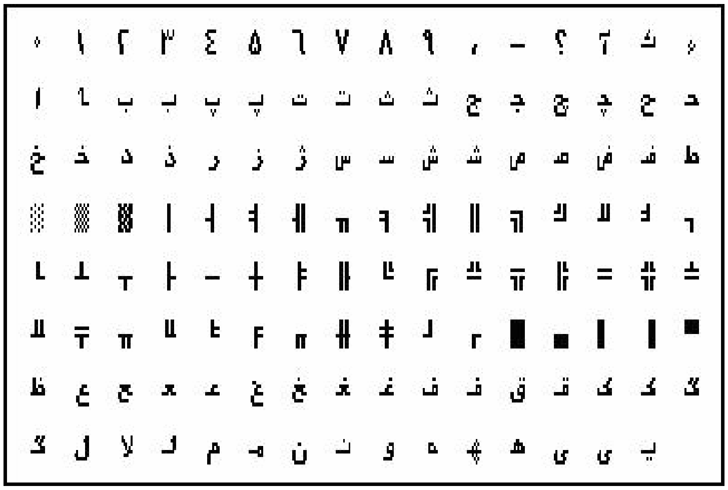
```
Indo-European
          Indo-Iranian
                    Iranian
                              Western
                                        South Western
                                                     Persian
                                                              FARSI
```

Figure 1: Language Tree Farsi of Language [1]

Farsi has been written with a number of different scripts, including Old Persian Cuneiform, Pahlavi, Aramaic, and Avestan. However after 642 AD Arabic script has been used for writing Farsi [2]. Nastalique style for Arabic script is used for writing Farsi.

Character Set and Encoding

Unicode Arabic script block from 0600-06FF is the standard character set encoding used for Farsi. A national standard based on relevant Unicode character subset within Arabic script block is also defined by Institute of Standards and Industrial Research in Iran (ISIRI) [3]. Earlier popular Farsi character set used for encoding was “Iran System”. The figure below shows this character set encoding.



۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	،	ـ	؟	آ	ﺋ	ء
ا	ﺎ	ب	ﺑ	پ	ﭘ	ت	ﺗ	ث	ﺛ	ج	ﺟ	چ	ﭼ	ح	ﺣ
خ	ﺧ	د	ذ	ر	ز	ژ	س	ﺳ	ش	ﺷ	ص	ﺻ	ض	ﺿ	ط
░	▒	▓	│	┤	╡	╢	╖	╕	╣	║	╗	╝	╜	╛	┐
└	┴	┬	├	─	┼	╞	╟	╚	╔	╩	╦	╠	═	╬	╧
╨	╤	╥	╙	╘	╒	╓	╫	╪	┘	┌	█	▄	▌	▐	▀
ظ	ع	ﻊ	ﻌ	ﻋ	غ	ﻎ	ﻐ	ﻏ	ف	ﻓ	ق	ﻗ	ک	ﮐ	گ
ﮔ	ل	ﻻ	ﻟ	م	ﻣ	ن	ﻧ	و	ه	ﻬ	ﻫ	ﯽ	ی	ﯾ	

Figure 2: Farsi Character Set Encoding for “Iran System”

Fonts and Rendering

Microsoft Platform

Microsoft Windows fonts Tahoma and Microsoft Sans Serif can be used for typing Farsi text. In addition to this there are other Unicode Farsi fonts available, some of which have been shown in the figure below. All these fonts follow the Naskh style of Arabic script. No Nastalique style font is available by Microsoft. Nastalique Open Type fonts are available from other organizations, which can be used for Persian, e.g. Nafees Nastalique [4].

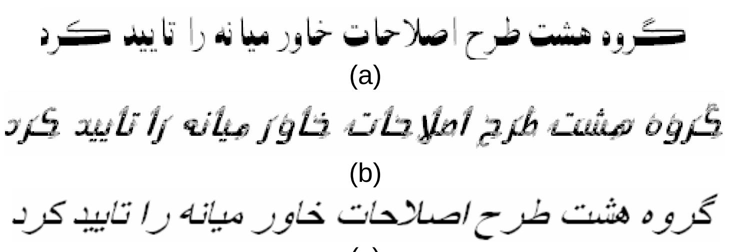
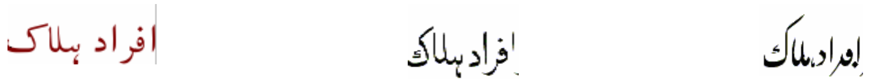

گروه هشت طرح اصلاحات خاور میانه را تایید کرد
(a)
گروه هشت طرح اصلاحات خاور میانه را تایید کرد
(b)
گروه هشت طرح اصلاحات خاور میانه را تایید کرد
(c)

Figure 3: Farsi Text Written in Arash, Sorkhpust and Times New Roman Fonts

Linux Platform

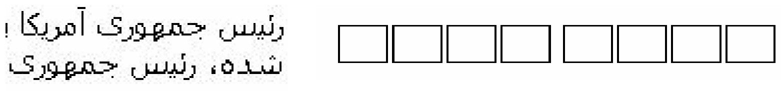
Open Type support for Nastalique style is not available on Linux platform as shown in Figure 4, which shows three fonts rendered on GNOME.



Persian (TTF) Nafees Naskh (OTF) Nafees Nastalique (OTF)

Figure 4: Farsi Fonts on GNOME

KDE only displays True Type fonts properly and does not display Open Type fonts, like Nafees Nastalique as shown in Figure 5.



Persian Naskh (TTF) Nafees Nastalique (OTF)

Figure 5: Farsi Fonts on KDE

Persian live CD Shabdix provides localized distribution of Persian on KDE 3.1.2. It has many built in Persian fonts but the rendering results are similar to those obtained on any KDE distribution [5].

Keyboard

ISIRI has published keyboard standard ISIRI 2901:1994 [6], shown in Figure 6.

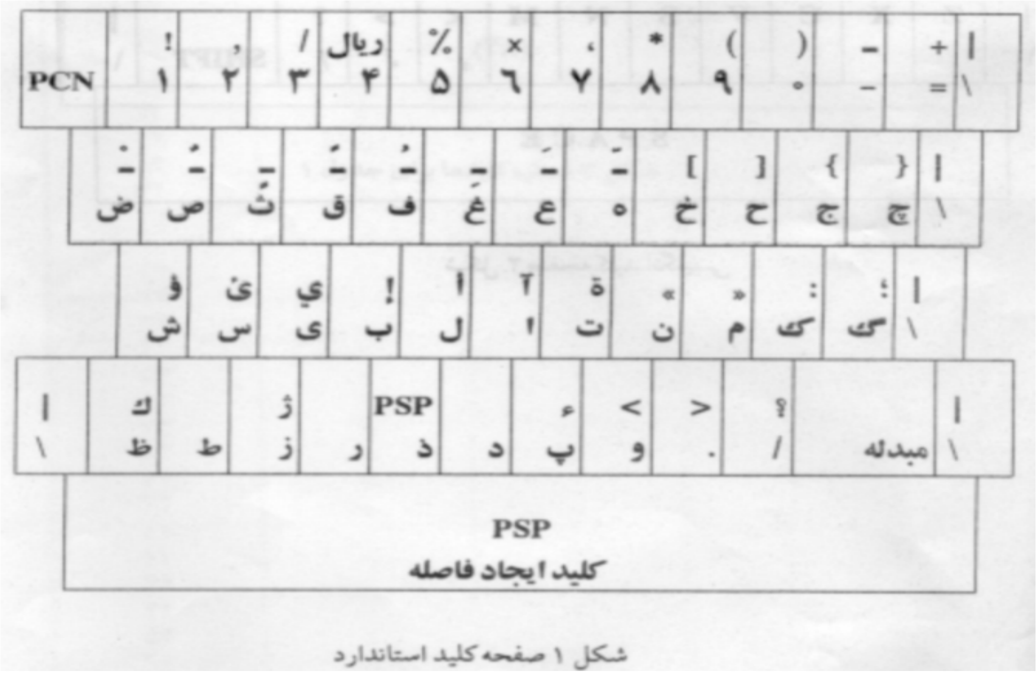


Figure 6: Iranian Standard ISIRI 2901 for Keyboard Layout [6]

Microsoft Platform

Microsoft Windows XP provides a Farsi keyboard, slightly different from ISIRI 2901. This can be used with all Unicode compatible applications. Figure 7 given below shows Farsi on-screen keyboard layout on Microsoft platform (see [7]). ISIRI 2901 compatible keyboards for Microsoft platform are also available from other organizations, e.g. see [8].

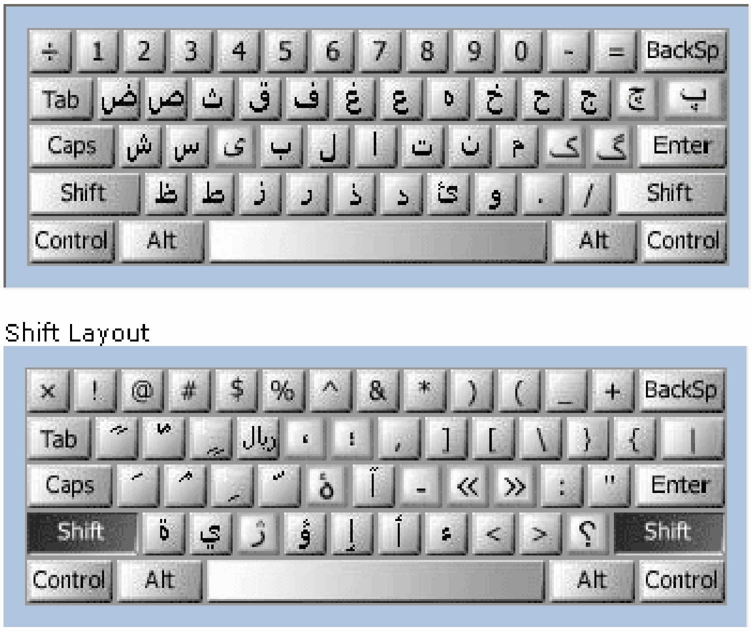


Figure 7: Microsoft Onscreen Keyboard Layout for Farsi [7]

Linux Platform

Red Hat 9 provides a Farsi keyboard layout which can be used both with KDE and GNOME. Keyboard support is part of input locale and runs at XWindows in Linux. Therefore, once enabled it works for all Unicode based technologies like Open Office and Mozilla. GNOME and KDE editors also support Farsi text. Farsi live distribution Shabdix also provides Farsi keyboard [5].

Collation

Collation sequence for Farsi has been defined for GNU C Library, but it has not been standardized.

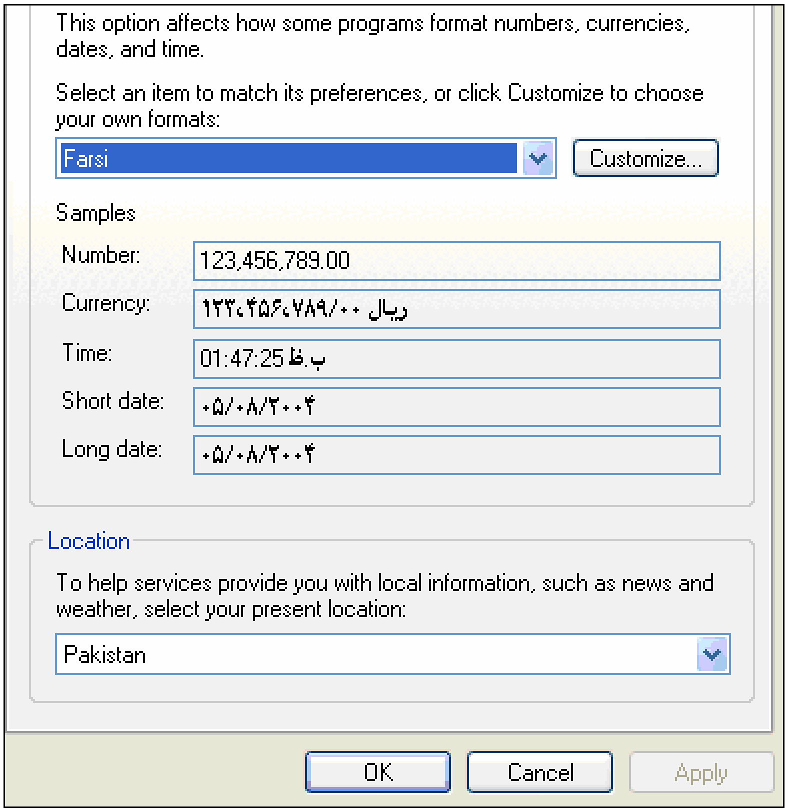
Farsi collation is supported by Microsoft through Farsi Language Interface Pack (LIP) [9]. Farsi collation is also supported on Linux platform.

Locale

Locale for Persian (fa_IR) is defined in CLDR 1.3.

Microsoft Platform

Microsoft XP provides Farsi locale enabling appropriate changes, as shown in Figure 8.



This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Farsi

Samples

Number: 123,456,789.00

Currency: ۱۲۳،۴۵۶،۷۸۹/۰۰ ریال

Time: 01:47:25 ب.ظ

Short date: ۰۵/۰۸/۲۰۰۴

Long date: ۰۵/۰۸/۲۰۰۴

Location

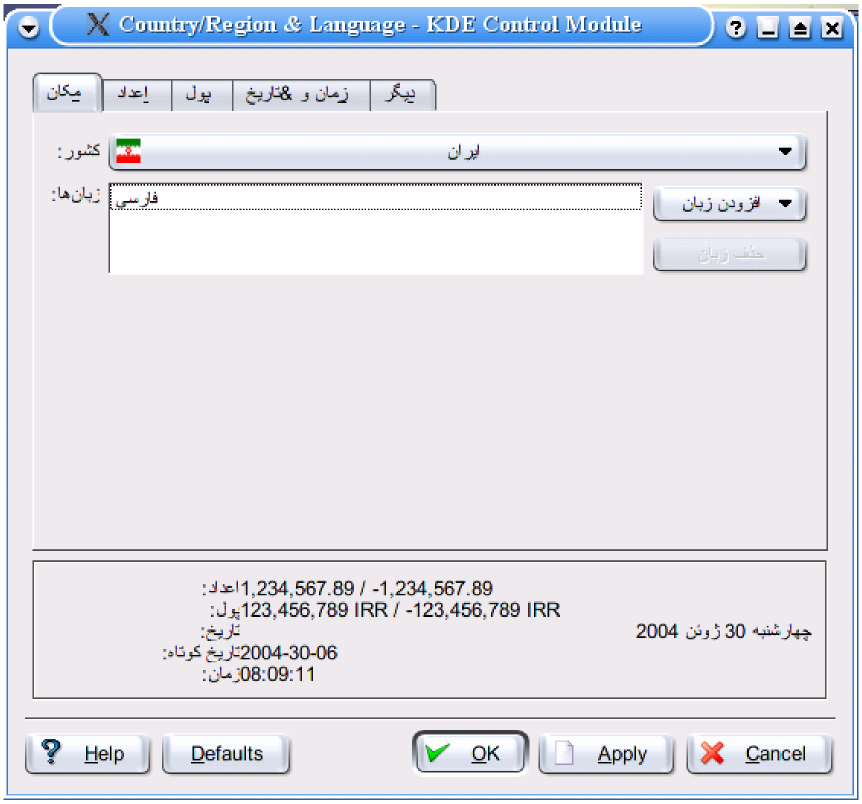
To help services provide you with local information, such as news and weather, select your present location:

Pakistan

Figure 8: Farsi Locale in Windows XP

Linux Platform

Locale definition for Farsi has been defined in Red Hat Linux 9. Farsi locale is enabled by default in Shabdix. Figure 9 shows KDE module localized in Shabdix.



کشور: ایران

زبان‌ها: فارسی

اعداد: 1,234,567.89 / -1,234,567.89
پول: 123,456,789 IRR / -123,456,789 IRR
تاریخ: چهارشنبه 30 ژوئن 2004
تاریخ کوتاه: 2004-30-06
زمان: 08:09:11

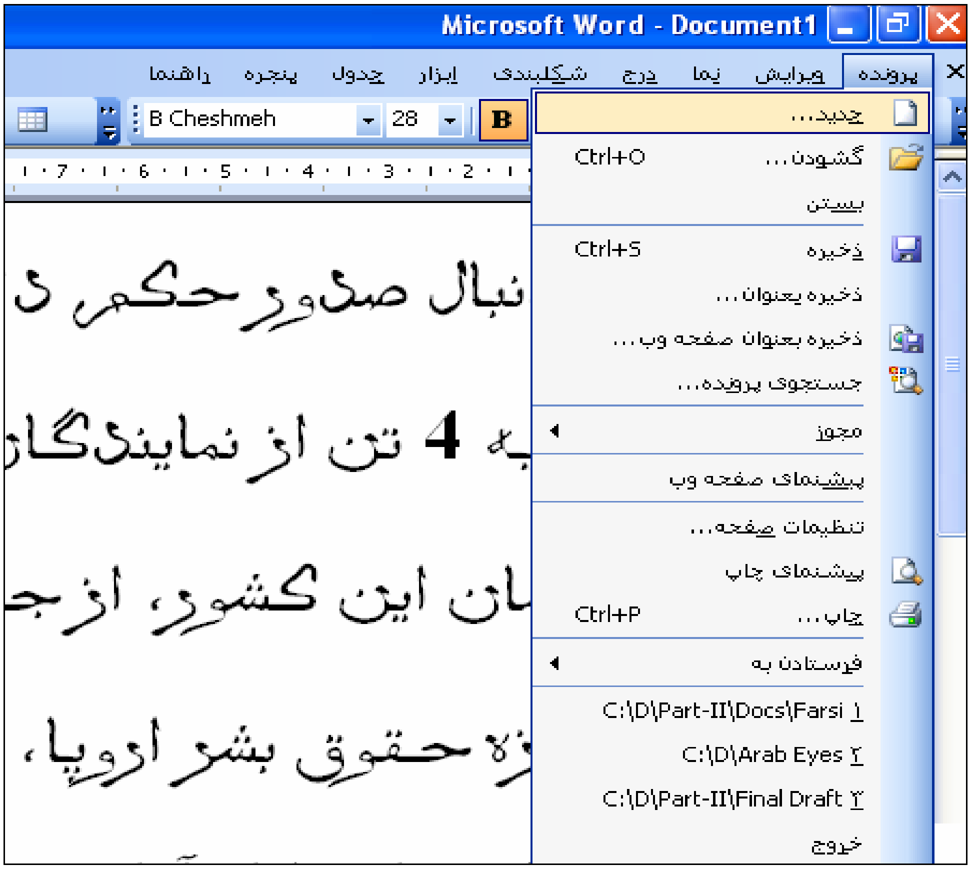
Figure 9: Farsi Date and Time Settings in Shabdix [4]

Interface Terminology Translation

Interface terminology for computers exists in Farsi, but has not been standardized.

Microsoft Platform

Microsoft provides complete Farsi interface through Farsi LIP (except Microsoft Office help files) for Windows 2000 and XP and Office 2003 [10]. Figure 10 shows the localized interface of Microsoft Word and Outlook in Farsi.



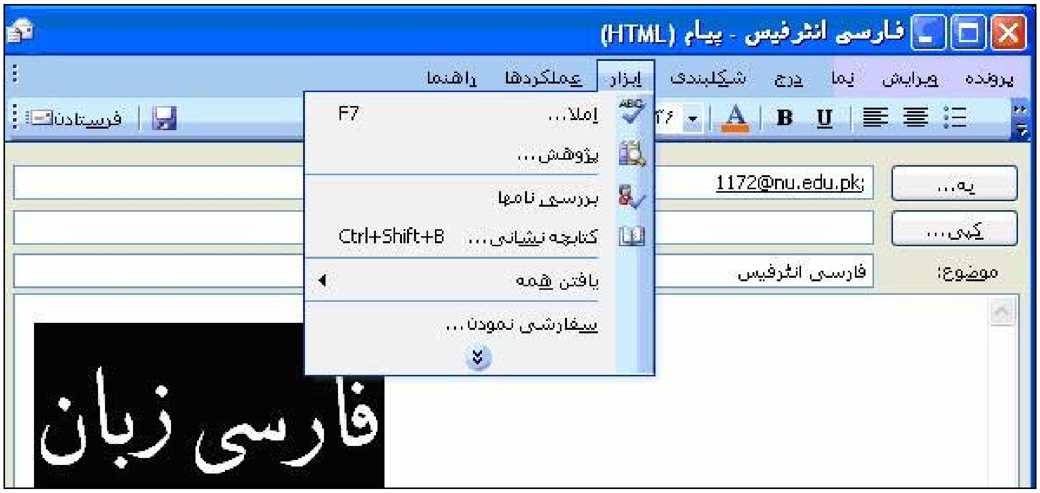
نبال صدور حکم د

به 4 تن از نمایندگار

ان این کشور، از ج

زه حقوق بشر اروپا،

(a)



1172@nu.edu.pk;

فارسی انترفیس

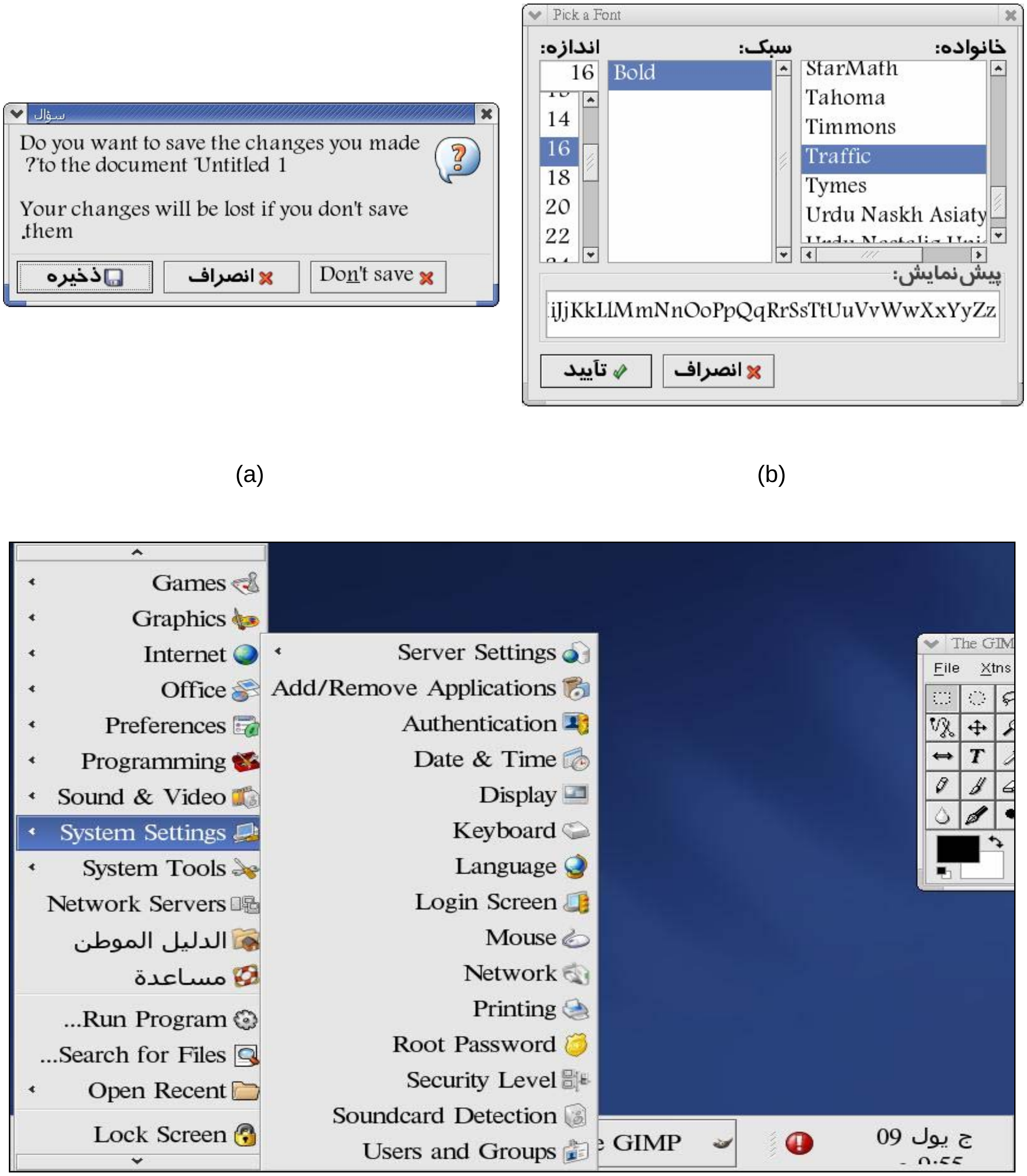
فارسی زبان

(b)

Figure 10: Localized Microsoft (a) Word, and (b) Outlook in Farsi

Linux Platform

Red Hat does not provide a localized Farsi interface for GNOME or KDE, but efforts are underway for the translation. GNOME 2.12 has been translated up to 45.89%, GNOME 2.10 is about 48% complete, nd KDE has been done up to 20%. Figure 11 below shows the Farsi translation in KDE on Red Hat distribution.



سؤال

Do you want to save the changes you made ?to the document Untitled 1

Your changes will be lost if you don't save .them

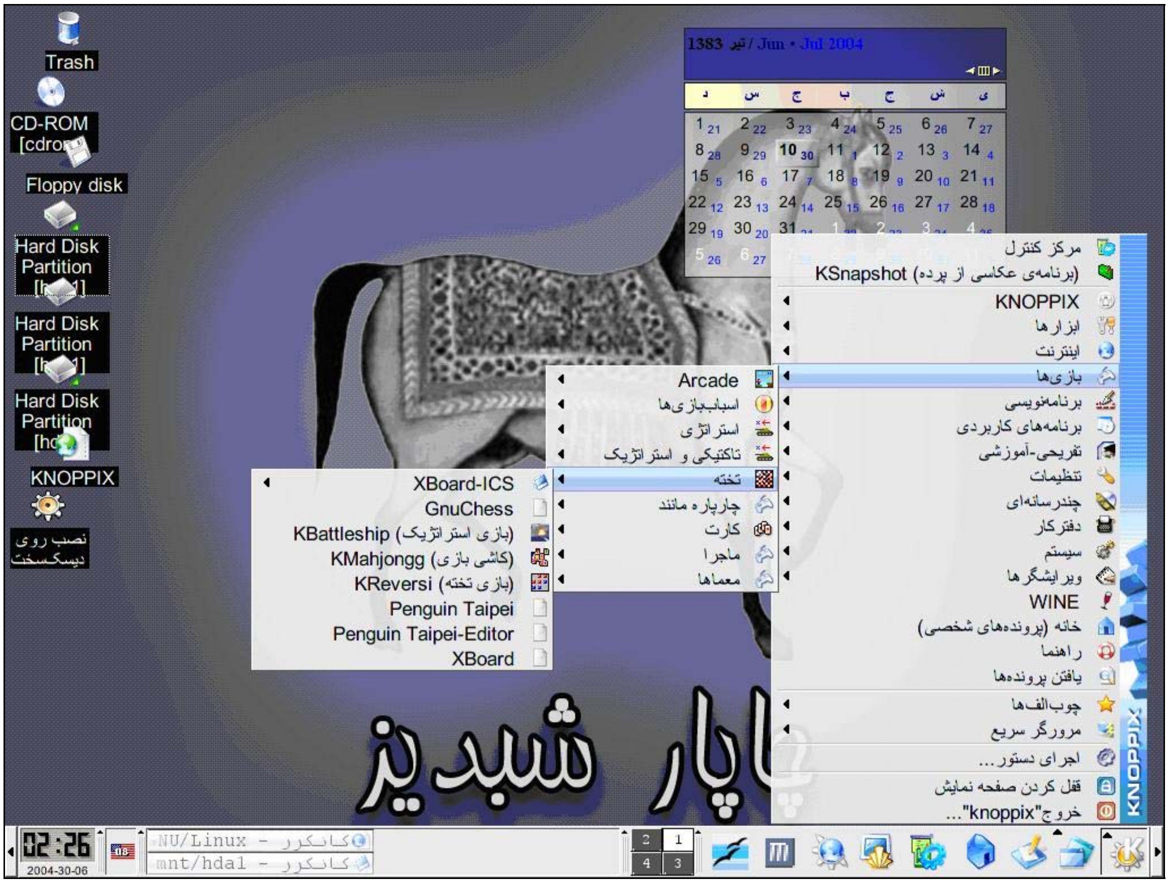
(a)

(b)

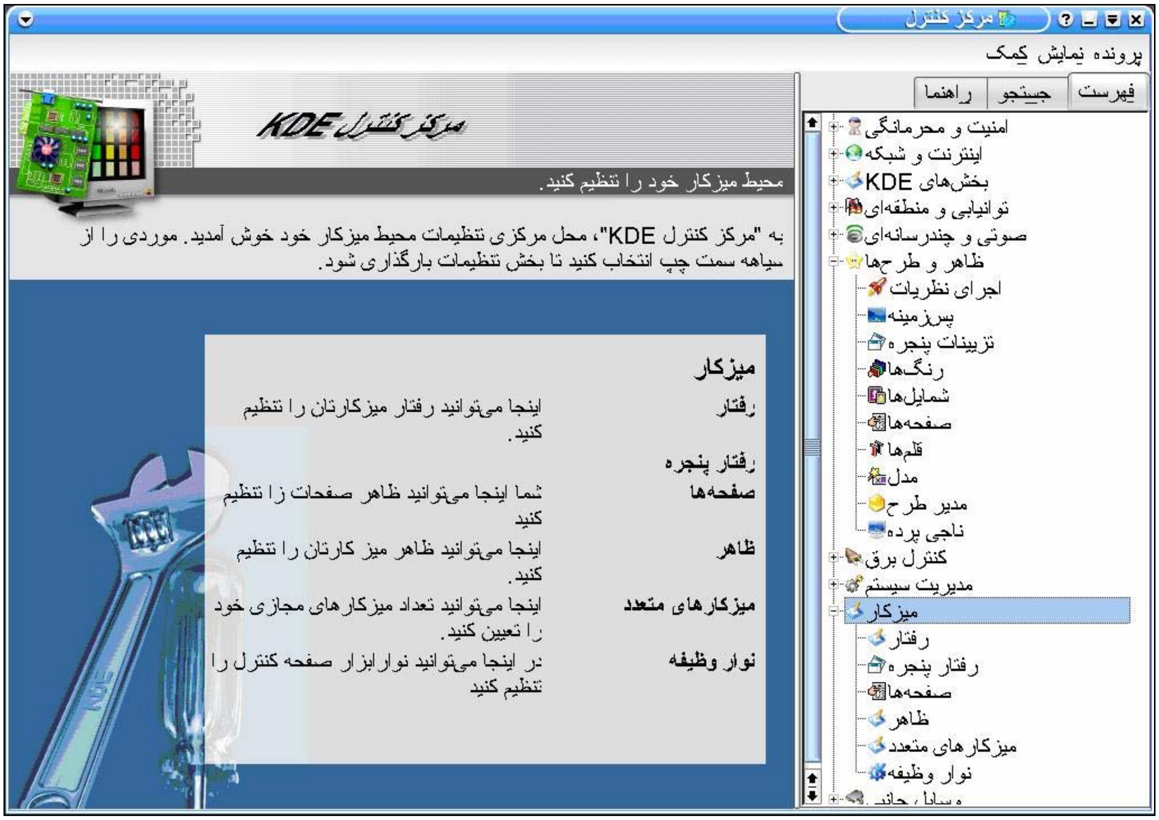
(c)

Figure 11: Partially Localized Red Hat 9 (a) File Save Dialogue, (b) Font Dialogue Box, and (c) Start-Up Menus

Shabdix provides a partial Farsi graphical user interface. Only the base KDE desktop menu files, an editor and Konqueror web browser are fully localized. Figure 12 shows the screen shots of Shabdix distribution. Farsi interface, desktop menus, editor menus all are designed to move from right-to-left, which is conventional for Arabic script based languages. This release is based on KNOPPIX 3.4 and includes an updated and modified Farsi KDE [5].



(a)



محیط میزکار خود را تنظیم کنید.

به "مرکز کنترل KDE"، محل مرکزی تنظیمات محیط میزکار خود خوش آمدید. موردی را از سیاهه سمت چپ انتخاب کنید تا بخش تنظیمات بارگذاری شود.

میزکار

رفتار	اینجا می‌توانید رفتار میزکارتان را تنظیم کنید.
رفتار پنجره	
صفحه‌ها	شما اینجا می‌توانید ظاهر صفحات زا تنظیم کنید
ظاهر	اینجا می‌توانید ظاهر میز کارتان را تنظیم کنید.
میزکارهای متعدد	اینجا می‌توانید تعداد میزکارهای مجازی خود را تعیین کنید.
نوار وظیفه	در اینجا می‌توانید نوارابزار صفحه کنترل را تنظیم کنید

(b)

Figure 12: (a) Start-Up Menu, and (b) Control Panel in Shabdix KDE

Status of Advanced Applications

Work is under progress to develop Farsi text-to-speech systems, lexicon, spell checker and thesaurus. However, these products are not commercially available.

References

[1] http://www.ethnologue.com/14/show_language.asp?code=PRS
[2] http://www.omniglot.com/writing/persian.htm
[3] "Information Technology – Persian Information Interchange and Display Mechanism, using Unicode." http://www.isiri.org/std/6219.htm
[4] http://www.crulp.org
[5] http://shabdix.berlios.de/
[6] "Keyboard Layout for Farsi." http://www.isiri.org/std/2901.htm
[7] http://www.microsoft.com/globaldev/keyboards/kbdfa.htm
[8] http://www.farsiweb.info/howto/win2keyb/
[9] http://www.microsoft.com/middleeast/arabicdev/farsi/wPaper.asp
[10] http://www.microsoft.com/middleeast/arabicdev/farsi/wPaper.aspx

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.

- Indo-European
    - Indo-Iranian
        - Indo-Aryan
            - Central zone
                - Western Hindi
                    - Hindustani
                        - HINDI

Figure 1: Language Family Tree for Hindi [1]

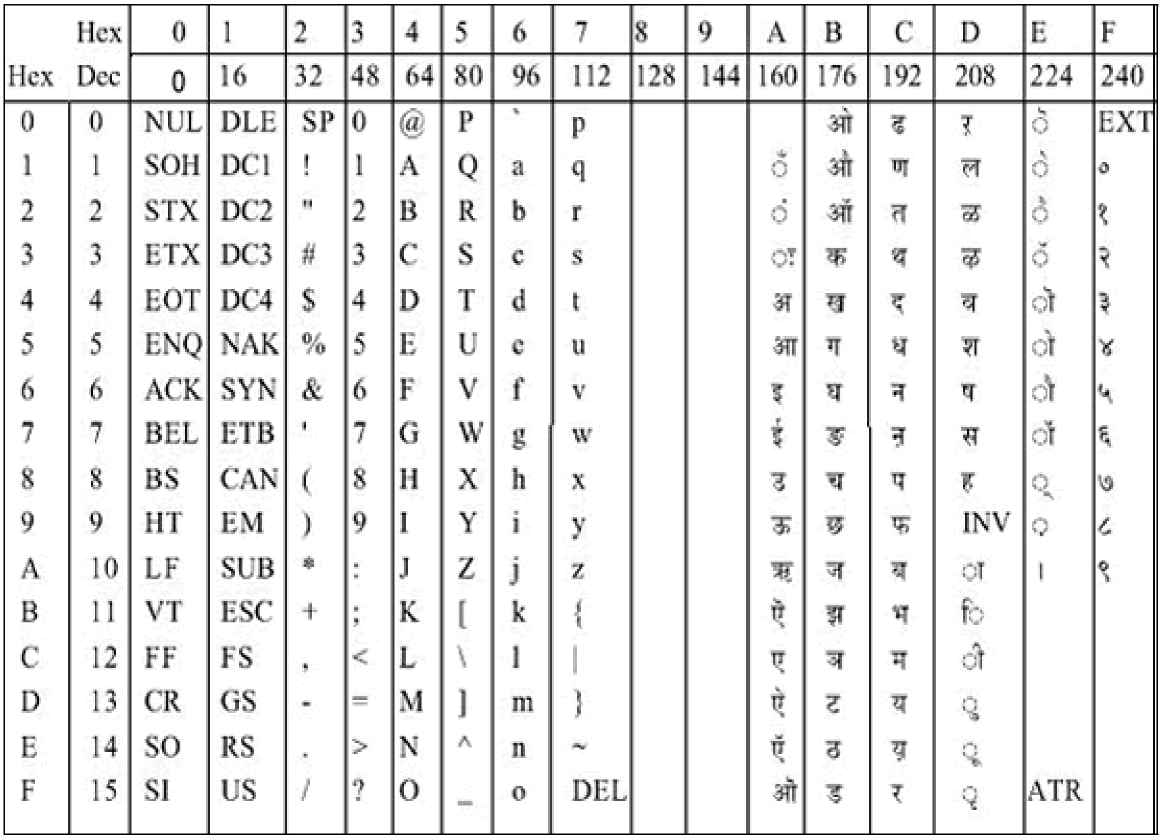
Hindi is written with Devanagari script, which was derived from Brahmi script [2].

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	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	ऴ	◌ॅ	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	{			ऎ	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			ऒ	ड	र	◌ृ	ATR	

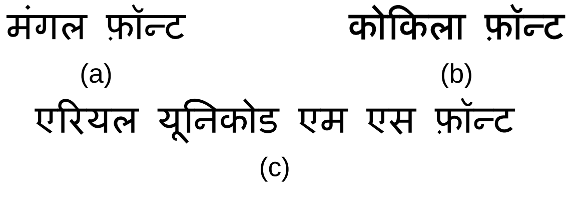
Figure 2: ISCII Code Chart IS 13194:1991 [3]

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.



मंगल फ़ॉन्ट

(a)

कोकिला फ़ॉन्ट

(b)

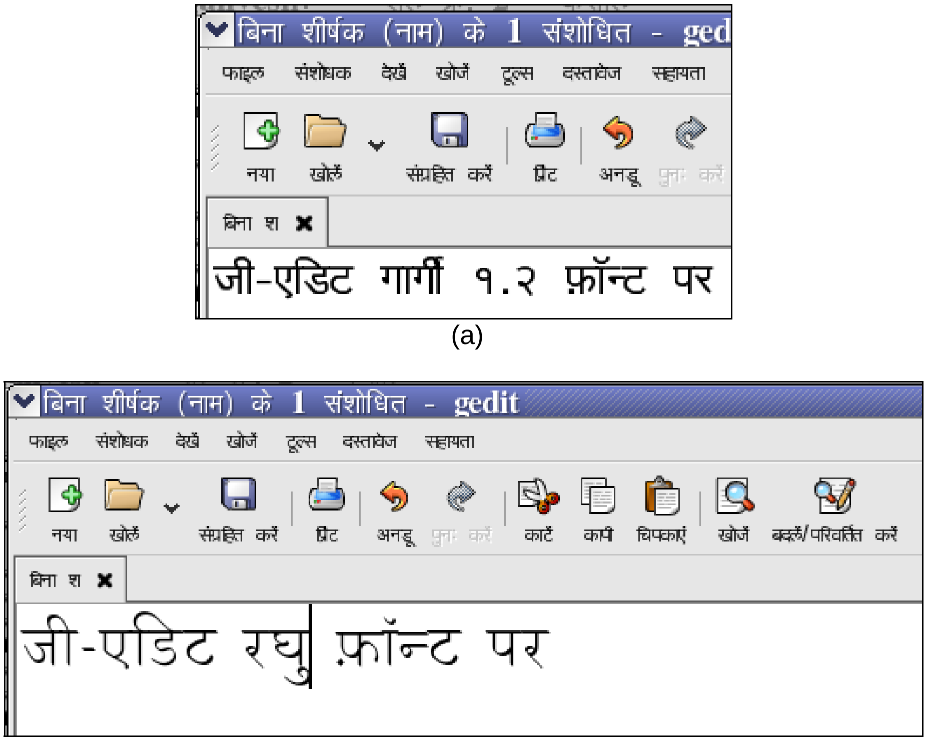
एरियल यूनिकोड एम एस फ़ॉन्ट

(c)

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.



जी-एडिट गार्गी १.२ फ़ॉन्ट पर

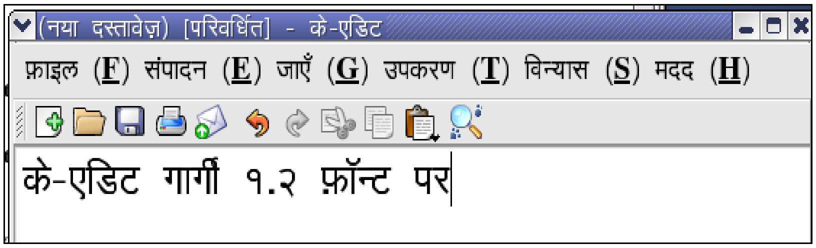
(a)

जी-एडिट रघु फ़ॉन्ट पर

(b)

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

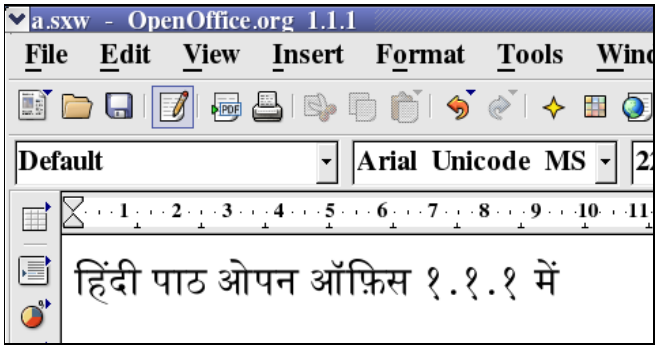
Pango rendering engine gives better results as it is more aware of substitution and positioning rules but these needs to be improved further. KDE does not provide any support for rendering Open Type fonts. Below is the demonstration of typing Hindi text using Gargi, on K-edit that runs Qt.



के-एडिट गार्गी १.२ फ़ॉन्ट पर

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.



हिंदी पाठ ओपन ऑफ़िस १.१.१ में

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.

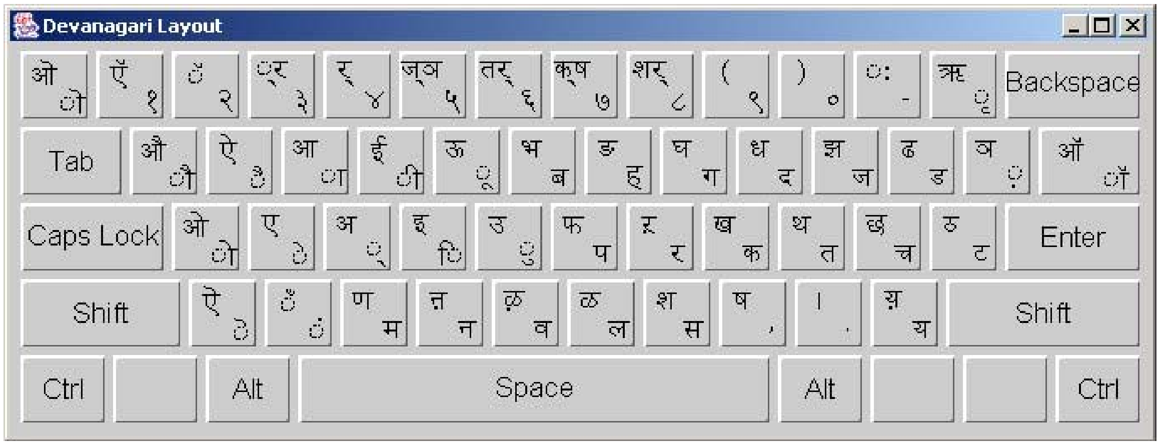


Figure 7: Inscript Keyboard for Hindi [11]

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].

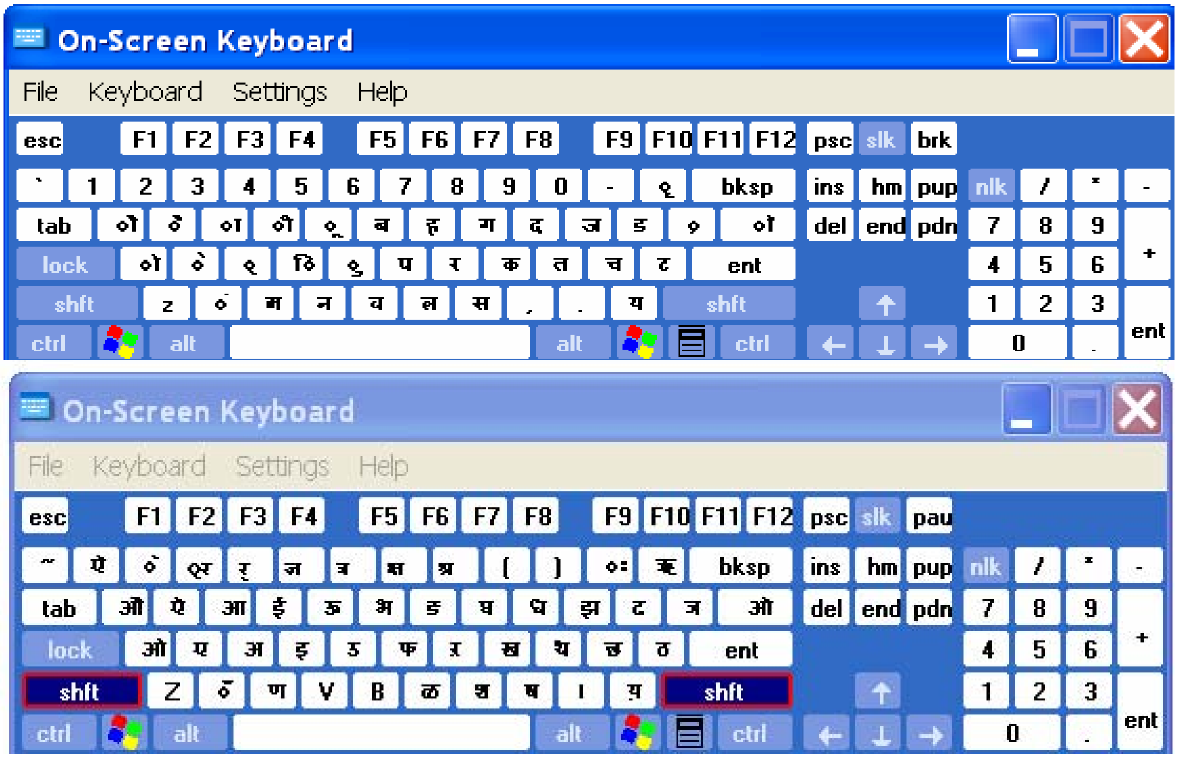


Figure 8: Traditional Keyboard Layout for Microsoft Platform

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

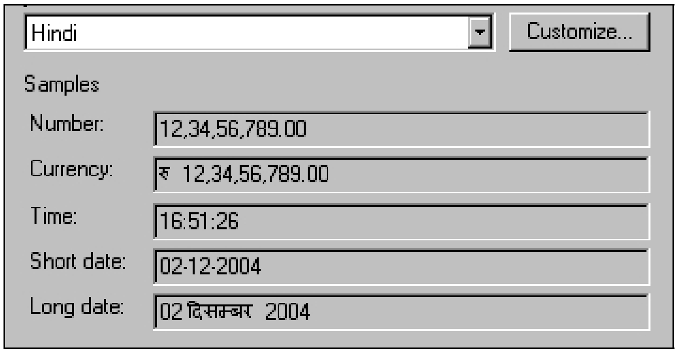
Collation on Linux platform is done using LC_COLLATE in the locale definition. The current hi_IN locale file does not have collation data included, so default sort order is used. As such this suffices for basic Hindi sorting, because Devanagari range in Unicode is based on ISCII-88 document, which does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



Hindi

Samples

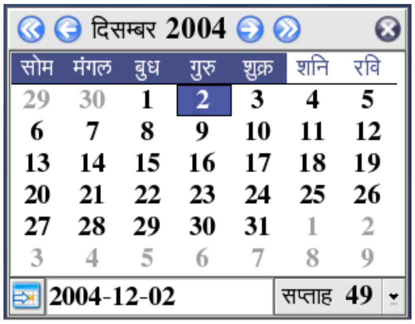
Number:	12,34,56,789.00
Currency:	रु 12,34,56,789.00
Time:	16:51:26
Short date:	02-12-2004
Long date:	02 दिसम्बर 2004

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.



दिसम्बर 2004

सोम	मंगल	बुध	गुरु	शुक्र	शनि	रवि
29	30	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

2004-12-02 सप्ताह 49

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.

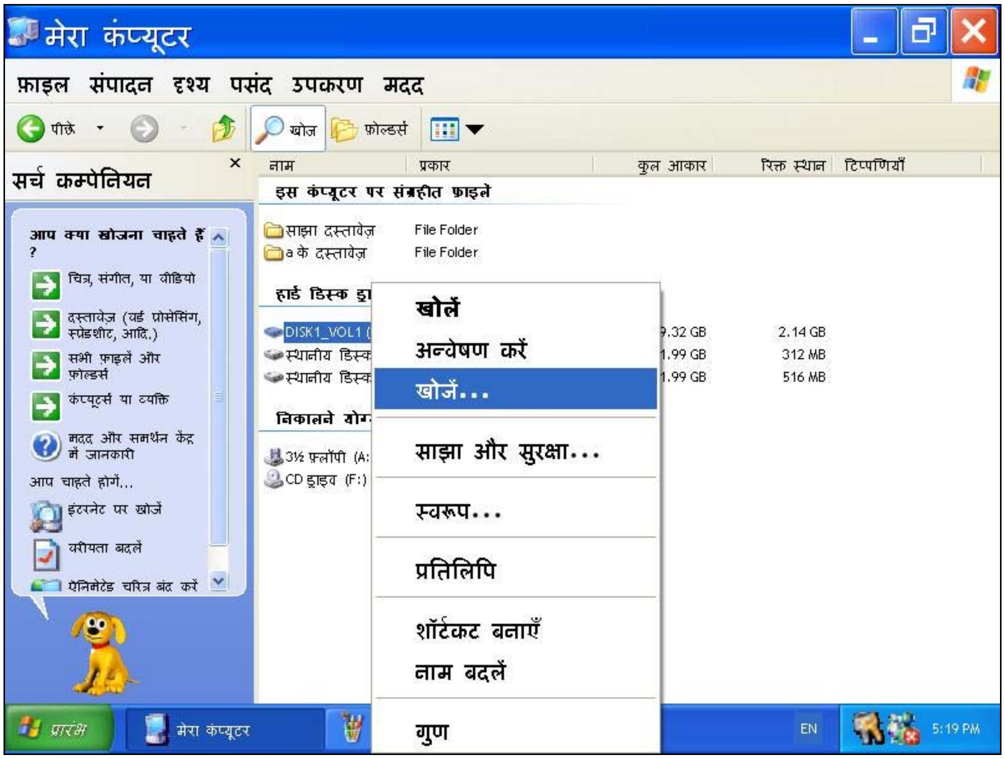


Figure 11: Localized Microsoft Windows XP

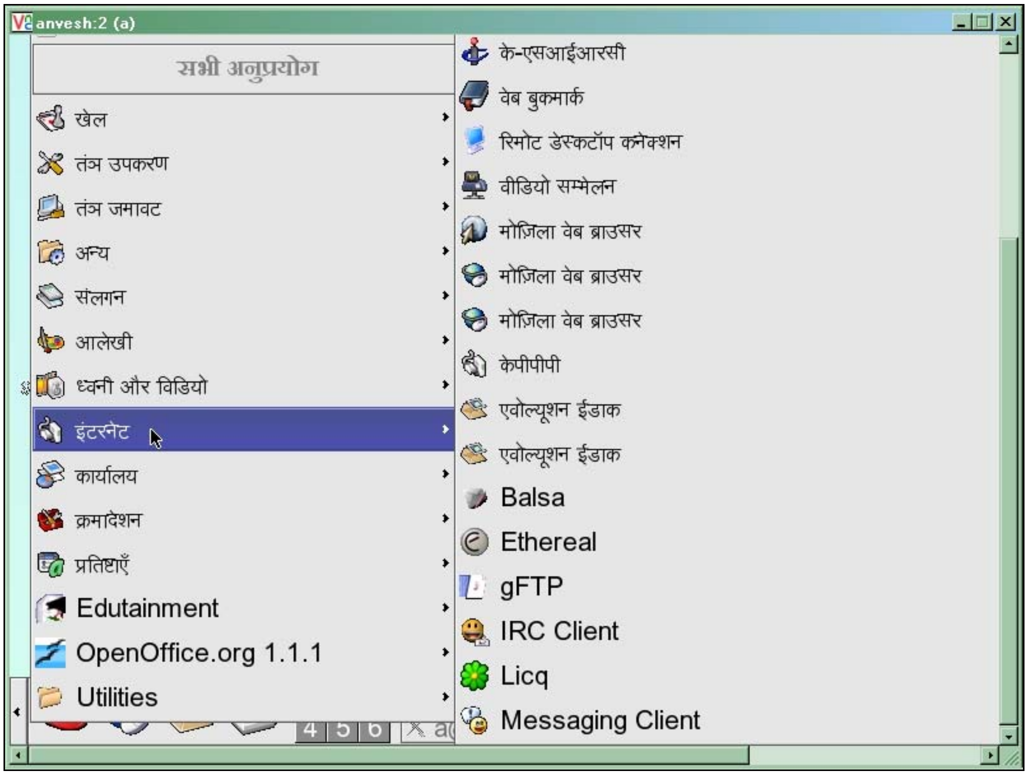


Figure 12: Localized Start-Up Menus for Hindi Linux Distribution

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

Indonesian

Indonesian or Bhasha Indonesia is similar to Malay language. It is the lingua franca and official language of Indonesia [1, 3]. Indonesian is spoken by 30 million people in Indonesia and other countries [2]. Indonesian vocabulary constitutes many foreign words from Arabic, English, Portuguese, etc. The language family tree of Indonesian is shown below.

```
Austronesian
     Malayo-Polynesian
          Malayic
               Malayan
                    Local Malay
                         INDONESIAN
```

Figure 1: Language Family Tree for Indonesian [1]

Indonesian is written in Latin script, though sometimes Arabic script is also used [1, 2]. Latin script was first introduced during the period of Dutch colonization, and later after independence in 1930, the script was formally adopted for writing Indonesian [2].

Character Set and Encoding

Unicode code values for Latin script character set are also used for Indonesian. Eight-bit Latin standard ISO 8859 is also used.

Fonts and Rendering

Usual Latin fonts and rendering is used for Indonesian.

Keyboard

Usual Latin characters based keyboard layout is used for Indonesian.

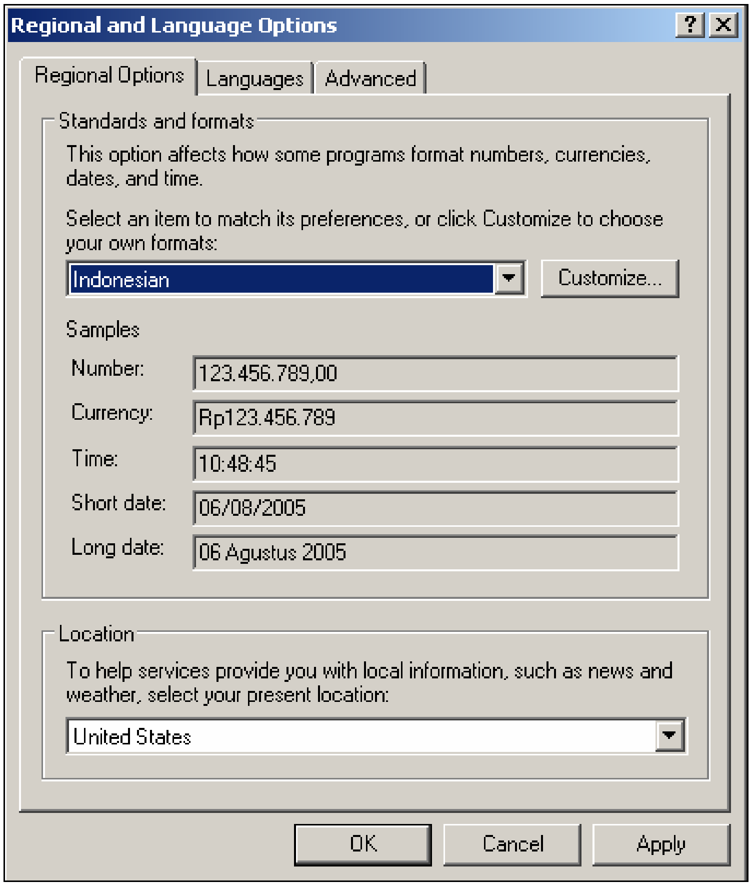
Collation

English collation order is used for Indonesian, which is available on most platforms.

Locale

Standard Locale for Indonesian has been defined in IBM ICU (id_ID). This locale definition has localized settings [4]. CLDR 1.3 also includes a locale for Indonesian.

Microsoft Windows XP provides support for Indonesian locale, as shown in Figure 2 below. The definition is also available in the Language Interface Pack for Indonesian [5]. Locale is also being incorporated in the Linux distributions.



Standards and formats

This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Indonesian

Samples

Number:	123.456.789,00
Currency:	Rp123.456.789
Time:	10:48:45
Short date:	06/08/2005
Long date:	06 Agustus 2005

Location

To help services provide you with local information, such as news and weather, select your present location:

United States

Figure 2: Indonesian Locale on Microsoft Platform

Interface Terminology Translation

Microsoft provides LIP for Indonesian which includes local language interface [5,6]. A complete Linux distribution in Indonesian is also available [11]. A team working on localizing GNOME in Indonesian language is currently active and GNOME glossary translation work has almost completed. Almost 87% of developers.lib and 84% of desktop strings have been translated [7]. In addition there are localized open source tools available through web portals in Indonesian (e.g. [8, 9, 10]).

The government of Indonesia has launched "Indonesia Go Open Source" program in 2004. This project is collaboration between Ministry of Research and Technology, Ministry of Communication and Information and the Indonesian Institute of Sciences. Through this project it is aimed to strengthen national information technology system and develop and freely disseminate OSS in Indonesian [12]. Following the objectives of the project it is believed that very soon a number of localized OSS applications will be available in Indonesian.

Status of Advanced Applications

Work is being done in language processing of Indonesian. There are many Indonesian and Indonesian-English dictionaries available online (e.g. see [13]). Spell checkers and morphological parsers for Indonesian are also available. There is very limited work on corpus [14, 15]. Indonesian text-to-speech system based on MBROLA was released in 2003 [16] and work has started on speech recognition [17] and machine translation with other languages.

References

[1] http://www.ethnologue.com/show_language.asp?code=ind
[2] http://www.omniglot.com/writing/indonesian.htm
[3] http://en.wikipedia.org/wiki/Indonesian_language
[4] http://www-950.ibm.com/software/globalization/icu/demo/locales/en/?_=gu&d_=en&_l=id
[5] http://www.microsoft.com/downloads/details.aspx?displaylang=id&FamilyID=0db2e8f9-79c4-4625-a07a-0cc1b341be7c
[6] http://www.microsoft.com/office/editions/prodinfo/language/availability.mspx
[7] http://l10n-status.GNOME.org/GNOME-2.8/id/index.html
[8] http://opensource-indonesia.com/kioss.php/index.php
[9] http://www.pandu.org/
[10] http://www.gerbanglinux.com/mod.php?mod=katalog&op=NewLinks&menu=26
[11] http://www.linux.or.id/
[12] http://www.igos.web.id/english/english.htm
[13] http://indonesian.dictionary.kamous.com/translator/reference.asp
[14] http://torvald.aksis.uib.no/corpora/2004-1/0231.html
[15] Nazief, B. “Development of Computational Linguistic Research: A Challenge for Indonesia,” available at http://acl.ldc.upenn.edu/P/P00/P00-1075.pdf
[16] http://lss.ee.itb.ac.id/~aa/indotts/
[17] Martin, T., Svendsen, T., Sridharan, S. (2003): "Cross-lingual pronunciation modeling for Indonesian speech recognition", In EUROSPEECH-2003, 3125-3128.

Japanese

Japanese, the national language of Japan, is spoken by about 127 million people across the globe [1]. Two forms of Japanese language are popularly in use, standard Japanese and common Japanese. The former is used as a medium of instruction in schools, print media and official documentation [3].

Japanese
Japanese
JAPANESE

Figure 1: Language Family Tree for Japanese [1]

Japanese script is used to write Japanese language. The script was formerly written using classical Chinese and eventually a hybrid Japanese-Chinese style. Currently, three styles are used collectively. First is a set of characters used to write Chinese loan words or similar Japanese words, called Kanji. In addition two syllabic (moraic) systems are used; Hiragana (used to write words either not present or too obscure in Kanji script) and Katakana (used to write sounds, scientific words, foreign words, etc.) [2, 3].

Character Set and Encoding

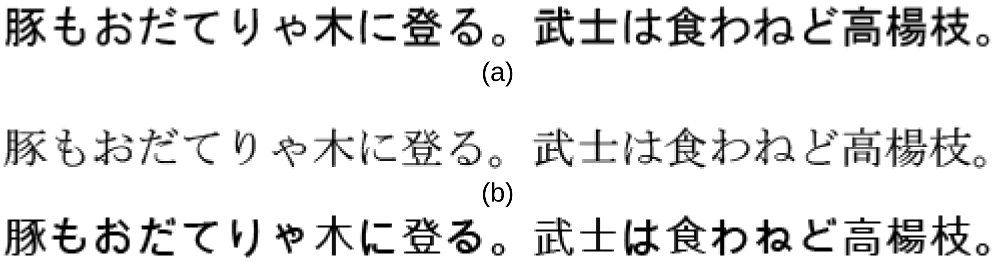
Japan has developed many Japanese Industrial Standards for character set encodings [4]. JIS X 0208-1990, is the most popular character set. It includes 6879 characters, Hiragana and Katakana syllabaries, 6355 Kanji, the Roman, Greek, and Cyrillic alphabets, numerals, and a number of typographic symbols. Normally, if JIS is not followed by a number, it refers to the JIS X 0208-1990 character set. This is not a single byte standard, and there are mechanisms to represent it in a sequence of two bytes. This is done by ISO 2022-JP [15] and EUC standards. Shift JIS is Microsoft's encoding for a Japanese character set containing approximately 7000 characters [4]. Unicode also supports Japanese characters. A detailed overview is given in [7].

Fonts and Rendering

There are several fonts which properly render Japanese script. Rendering is appropriately supported on most platforms.

Microsoft Platform

Currently, the Japanese version of Microsoft Windows comes with a set of fonts. Some Japanese fonts are Heisei Kaku Gothic, Heisei Maru Gothic, Heisei Mincho, MS Gothic, MS Mincho and Arial Unicode. The following figures show Japanese font rendering on Microsoft platform.


豚もおだてりゃ木に登る。武士は食わねど高楊枝。
豚もおだてりゃ木に登る。武士は食わねど高楊枝。
豚もおだてりゃ木に登る。武士は食わねど高楊枝。

(a)

(b)

(c)

Figure 2: (a) MS Gothic, (b) MS Mincho and (c) Arial Unicode [5]

Linux Platform

Commercial distributions of Linux use commercial fonts. The costs for these fonts are covered by the package price of the distribution. GNOME, KDE, Open Office and Mozilla use the font support from the underlying operating system. There are also a few open source Japanese fonts available [6].

Keyboard and Input Method Engines

There is currently one national standard for keyboard layout, called JIS X 6002-1980, "Keyboard layout for information processing using the JIS 7 bit coded character set". Although the standard describes itself as a keyboard layout for seven bit coded characters, it is widely and commonly used for the input of any Japanese encoding. This standard was later improved to JIS X 6004-1986 [7]. Another national standard, JIS X 4063-2000, "Keystroke to KANA Transfer Method Using Latin Letter Key for Japanese Input Method" defines the rules for input method for conversion from a combination of Latin characters to Japanese characters [8].

Microsoft Platform

Input Method Engines (IMEs) are widely used for converting key combinations of Latin letters into Japanese characters and then into Chinese characters. Microsoft Office provides a wide variety of IMEs for Japanese text input, which are explained in [9], and summarized below.

Keyboard and related functionality for Japanese is supported by Microsoft operating systems and .NET framework. The figure below shows the Hiragana/Katakana (JIS layout) IME used as a default. Once characters are entered conversion routines can be used to change characters from Katakana to the Kanji. The figure given below shows the normal state Hiragna/Katakana keyboard layout for Japanese.

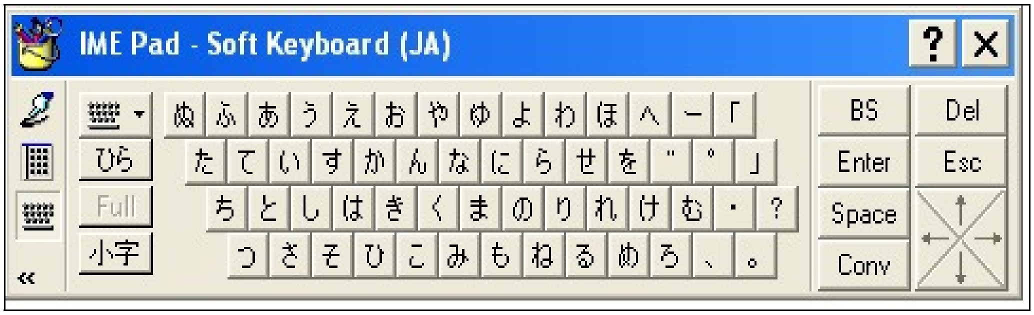

IME Pad - Soft Keyboard (JA)
?
X
ひら
Full
小字
«
ぬ ふ あ う え お や ゆ よ わ ほ へ ー 「
た て い す か ん な に ら せ を ゛ ゜ 」
ち と し は き く ま の り れ け む ・ ?
つ さ そ ひ こ み も ね る め ろ 、 。
BS
Del
Enter
Esc
Space
Conv

Figure 3: Hiragana/Katakana Keyboard Layout [9]

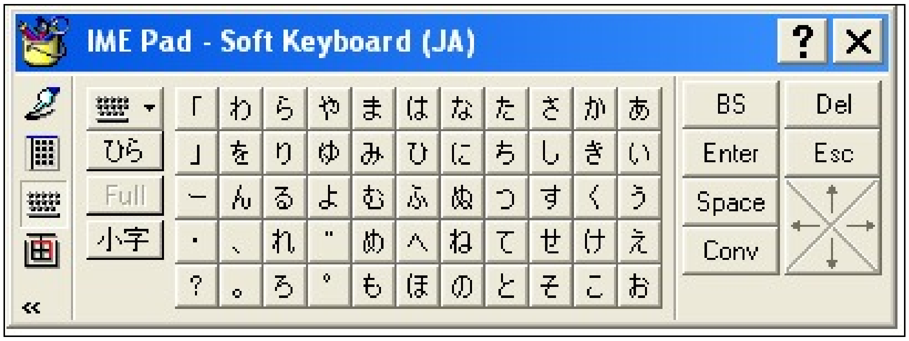

IME Pad - Soft Keyboard (JA)
?
X
ひら
Full
小字
«
「 わ ら や ま は な た さ か あ
」 を り ゆ み ひ に ち し き い
ー ん る よ む ふ ぬ つ す く う
・ 、 れ ゛ め へ ね て せ け え
? 。 ろ ゜ も ほ の と そ こ お
BS
Del
Enter
Esc
Space
Conv

Figure 4: Hiragana/Katakana Syllabic Keyboard Layout [9]

The Japanese numeric and date keyboard layout gives the user access to the Kanji used for dates and time.

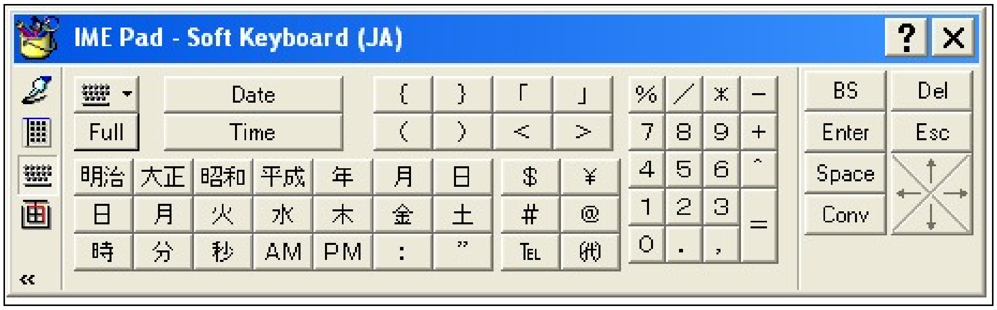
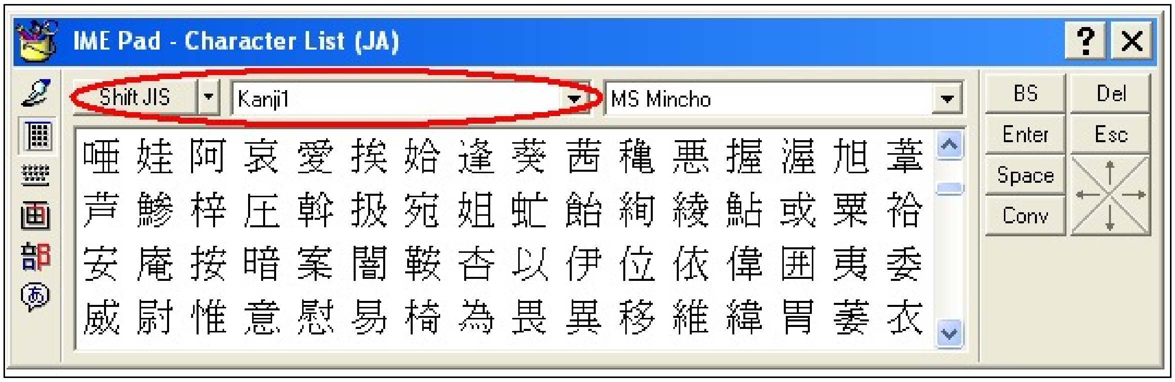


Figure 5: Date and Time Keyboard Layout [9]

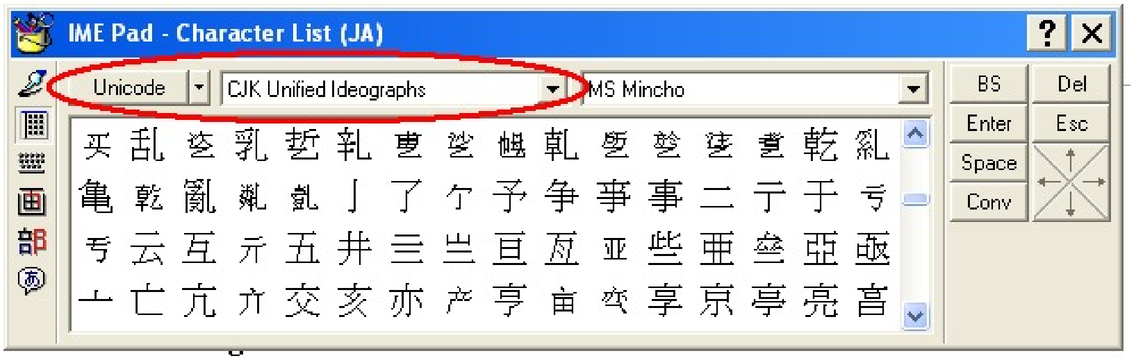
Another form of character input mechanism developed for Japanese character input facilitates users to select characters from two lists: one being the Shift JIS list which displays the characters based on JIS standard, and the second is based on Unicode Japanese characters.



Shift JIS | Kanji1 | MS Mincho

唖 娃 阿 哀 愛 挨 姶 逢 葵 茜 穐 悪 握 渥 旭 葦
芦 鯵 梓 圧 斡 扱 宛 姐 虻 飴 絢 綾 鮎 或 粟 袷
安 庵 按 暗 案 闇 鞍 杏 以 伊 位 依 偉 囲 夷 委
威 尉 惟 意 慰 易 椅 為 畏 異 移 維 緯 胃 萎 衣

Figure 6: JIS Based Character List [9]



Unicode | CJK Unified Ideographs | MS Mincho

买 乱 乲 乳 乴 乵 乶 乷 乸 乹 乺 乻 乼 乽 乾 乿
亀 亁 亂 亃 亄 亅 了 亇 予 争 亊 事 二 亍 于 亏
亐 云 互 亓 五 井 亖 亗 亘 亙 亚 些 亜 亝 亞 亟
亠 亡 亢 亣 交 亥 亦 产 亨 亩 亪 享 京 亭 亮 亯

Figure 7: Character List Based on Unicode CJK Characters [9]

Microsoft additionally provides a stroke utility within the Japanese IME Pad. It provides the user to look up a Japanese character by the strokes it takes to write the character. Figure 8 given below shows the Japanese strokes utility of the Japanese Input Method Engine.

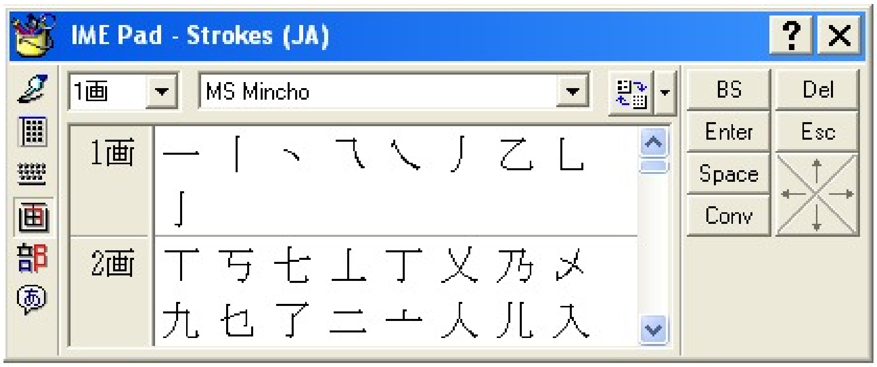


Figure 8: Strokes Utility in IME Pad [9]

Similar to the strokes utility, the radical utility for Japanese characters allows the user to look up a Japanese character in groups of main radicals of the Japanese language. The following figure shows the Japanese radical utility of the Japanese IME Pad.

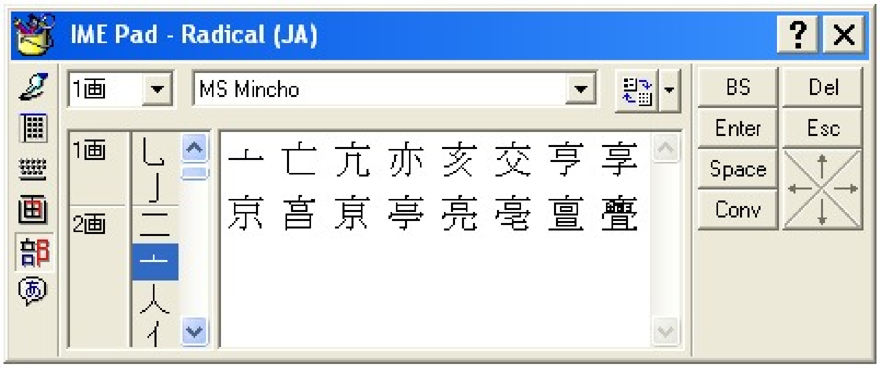
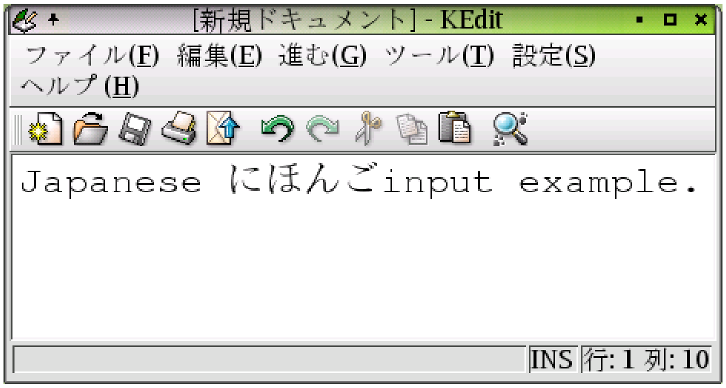


Figure 9: Radical Utility in IME Pad [9]

Linux Platform

Qt3/KDE3 support different styles of input methods and editing as defined in the XIM standard (X Input Method). They include “on the spot”, the pre-edit, over the spot, off the spot and root IME layout. The following figure shows the “on the spot” Japanese IME in KDE.


[新規ドキュメント] - KEdit
ファイル(F) 編集(E) 進む(G) ツール(T) 設定(S)
ヘルプ(H)
INS
行: 1 列: 10

Japanese にほんごinput example.

Figure 10: X-Input Method in K-Edit [10]

Collation

The JIS X 4061:1996, "Collation of Japanese character string" standardizes collation rules for Japanese characters [7]. However this standard is not widely implemented in computer software.

Microsoft Platform

Collation according to the Japanese character order as appears in Unicode has been supported in Microsoft platform. Microsoft Office components MS Word/Excel/Access and MS Outlook support dictionary order for Japanese. In the Microsoft .Net framework, string comparison is based on character sequence.

Linux Platform

Applications on Linux platform do not sort Japanese text in lexicographic order.

Locale

Japanese locale is defined in IBM ICU [11] and CLDR 1.3. The locale for Japanese in Japan is ja_JP.

Locale for Japanese has been implemented on Microsoft platform, and is available for all applications on this platform. On the Linux platform locale for Japanese (ja_JP.eucJP) has been implemented. GNOME, KDE, Open Office and Mozilla also support Japanese locale.

Interface Terminology Translation

There is no standardized terminology or glossary for the user interface of software. However, over the past decade of software development and localization, interface terminology has become fairly consistent among various software vendors, which has created a considerable similarity in the translation of interface across various software applications.

Microsoft Platform

Microsoft ships a complete localized version of Windows in Japanese. In addition to the applications and Microsoft Office suite, the .Net Framework has been completely localized in Japanese.

Linux Platform

In GNOME almost 81% of the glossary has been localized [12], in KDE 99% of the glossary has been translated [13], and Mozilla and Open Office have been completely localized.

Status of Advanced Applications

On Microsoft platform line breaking algorithm is implemented at the application level. Office applications including Word and PowerPoint provide a more sophisticated line breaking mechanism with user customization enabled, while Excel and Access provide only limited line breaking. Line breaking algorithm is also supported in Microsoft Outlook and Internet Explorer, but it supports only one level, although there are multiple levels of line breaking in Japanese text. Microsoft .Net framework also supports line breaking for Japanese.

Automatic line breaking algorithm is also implemented at the level of individual applications in KDE, but desktop environment does not provide any support. In Open Office, line breaking algorithm is implemented at different levels for different components (Writer, Calc, Impress, Draw and Database). Writer and Impress provide more sophisticated line breaking with user customization compared to other components.

Microsoft spell checker has been implemented for Japanese, which is available in Word and Outlook. On Linux platform, KDE spell checker is implemented at the level of individual applications, and desktop environment does not provide any support. In Mozilla no spell checking is provided for Japanese. In Open Office applications no Japanese spell checking is implemented. However, there is spell checking done in Japanese in the course of Kana-Kanji conversion, which is provided by the system.

Japanese and multilingual dictionaries (with Japanese) are also widely available.

Text-to-speech synthesis systems have also been developed but are still being studied for enhanced naturalness. Consumer products, such as voice browsers are also available. Speech recognition systems have also been implemented for Japanese and associated consumer products such as voice input method engines (IME) are also available. Latest versions of Microsoft Windows also provide voice input. Machine translation systems have also been implemented for a variety of language pairs, including English, Thai and some other languages. Work is also in progress on speech to speech translation.

Optical character recognition systems have been developed. Handwriting (graffiti) recognition for Japanese is also available. Some localized handheld devices that include PalmOS devices, Pocket PCs and Apple iPod, are available with Japanese graffiti recognition. Other than these some domestic devices are also available, such as Sharp Zaurus. The handwriting recognition applet of the IME Pad has been developed by Microsoft. The utility allows user to put a stroke to search for next possible Kanji. The IME tries to recognize the character after each stroke is written. The following figure shows the complete process to enter the Kanji character.

Links to these applications are not provided because there are many such applications by different organizations and universities, using various techniques, and may be found through most search engines (e.g. ATR laboratories [14]).

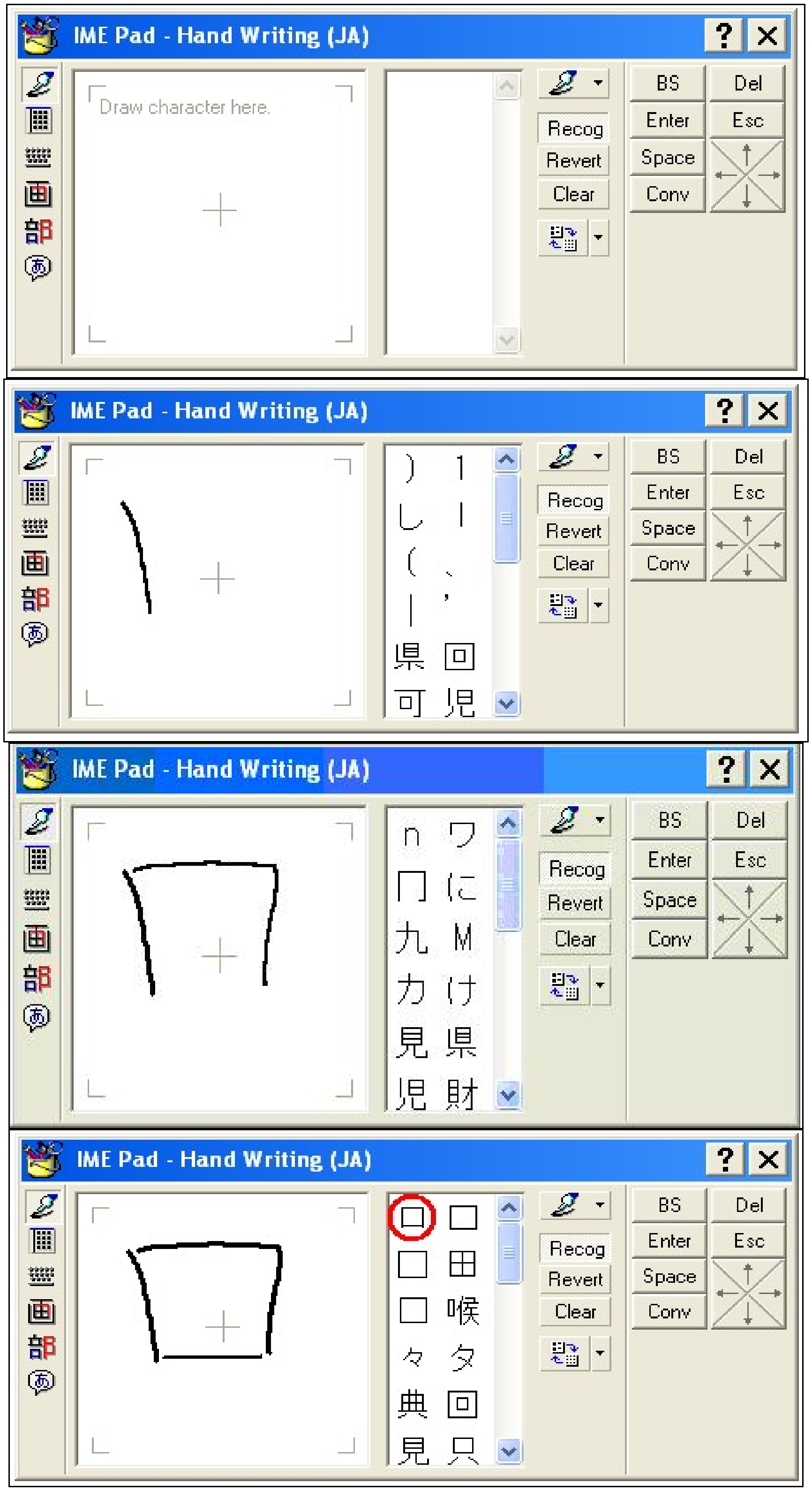

IME Pad - Hand Writing (JA)
Draw character here.
Recog
Revert
Clear
BS
Del
Enter
Esc
Space
Conv
) 1 し | (、 | ， 県 回 可 児
n ワ 冂 に 九 M 力 け 見 県 児 財
口 口 口 田 口 喉 々 夕 典 回 見 只

Figure 11: Handwriting IME for Kanji Characters [9]

References

[1] http://www.ethnologue.com/show_language.asp?code=jpn
[2] www.omniglot.com
[3] http://en.wikipedia.org/wiki/Japanese_language
[4] http://lfw.org/text/jp.html
[5] http://www.ascendercorp.com/msfonts/msfonts_eastasian.html
[6] http://eyegene.ophthy.med.umich.edu/unicode/fontguide/
[7] Kanaya, A. "Current status and issues of information technology standardization policies in Japan," in *Proceeding of AFSIT*. Available at http://www.cicc.or.jp/english/hyoujyunka/af10/10-06.html
[8] http://en.wikipedia.org/wiki/W%C4%81puro_r%C5%8Dmaji
[9] http://www.microsoft.com/globaldev/handson/user/IME_Paper.mspx
[10] http://www.suse.de/~mfabian/suse-cjk/KDE-input-style.html
[11] http://www-950.ibm.com/software/globalization/icu/demo/locales/en/?d=en&=ja_JP
[12] http://www.GNOME.org
[13] http://www.KDE.org
[14] http://www.slt.atr.jp/slc-e/
[15] http://en.wikipedia.org/wiki/ISO_2022

Khmer

Khmer is the official language of the Kingdom of Cambodia. It is spoken by about 13 million speakers, mostly residing in Cambodia, Vietnam, Laos, Thailand, China, France and the USA [1]. Khmer belongs to Mon-Khmer group of Austro-Asiatic languages (see Figure 1), and shares many features and vocabulary with Thai as a result of centuries of two-way borrowing. Khmer also has significant influence of Sanskrit, Pali, French, and Chinese languages [2].

- Austro-Asiatic
  - Mon-Khmer
    - Eastern Mon-Khmer
      - Khmer
        - KHMER, CENTRAL

Figure 1: Language Family Tree of Khmer [1]

Khmer alphabet is derived from Brahmi script and resembles Thai and Lao writing systems. The earliest known inscriptions in Khmer, found at Angkor Borei (in Takev Province south of Phnom Penh), dates back to 611 AD [2].

Character Set and Encoding

Unicode chart 1780-17FF [3] is the standard encoding for Khmer character set. This encoding is being increasingly used at the national level. However other ad hoc 8-bit encodings are also being used nationally. Among the few more popular encodings used are Limon, Khek and ABC. Encoding converters between Unicode and these fonts are available through PAN Localization Cambodian component [4] and Khmer OS Project [5].

Fonts and Rendering

As mentioned above, most fonts currently being used are based on ad hoc 8-bit encoding schemes. However, now increasing number of Unicode based fonts are available, which can work on both Microsoft and Linux systems, if rendering support is available. For a list see [5, 7].

Microsoft Platform

Microsoft Office 2003 with Service Pack 1 have now included support for rendering Khmer text, but Khmer fonts are not shipped. Microsoft has published some guidelines to develop Khmer fonts based on Unicode [6].

Linux Platform

Khmer OS has been working on developing rendering support for Khmer on Linux. Qt shipped with KDE 3.3, now supports complete rendering for Khmer in KDE. Patch for Pango has also been developed for support in GNOME [5].

Keyboard

No standard keyboard layout exists for Khmer. AZERTY keyboard layout is commonly used. This is shown in Figure 2 below.

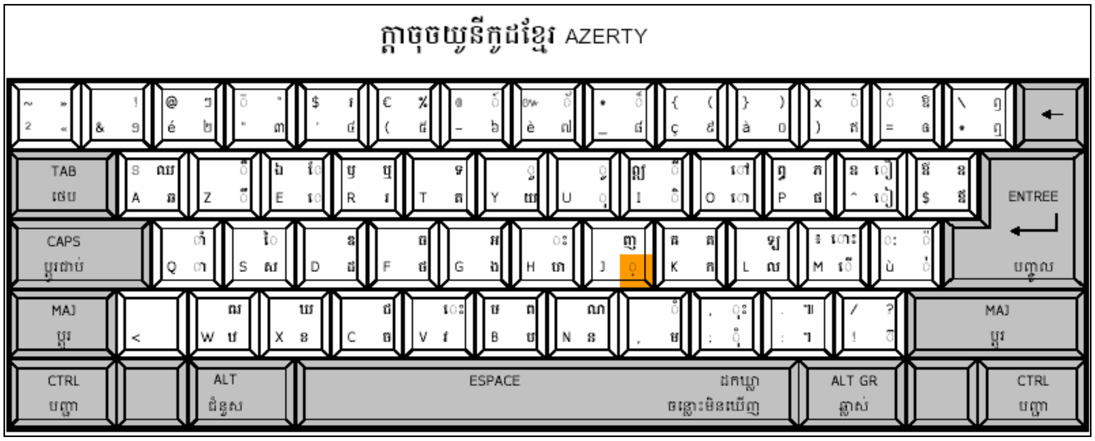
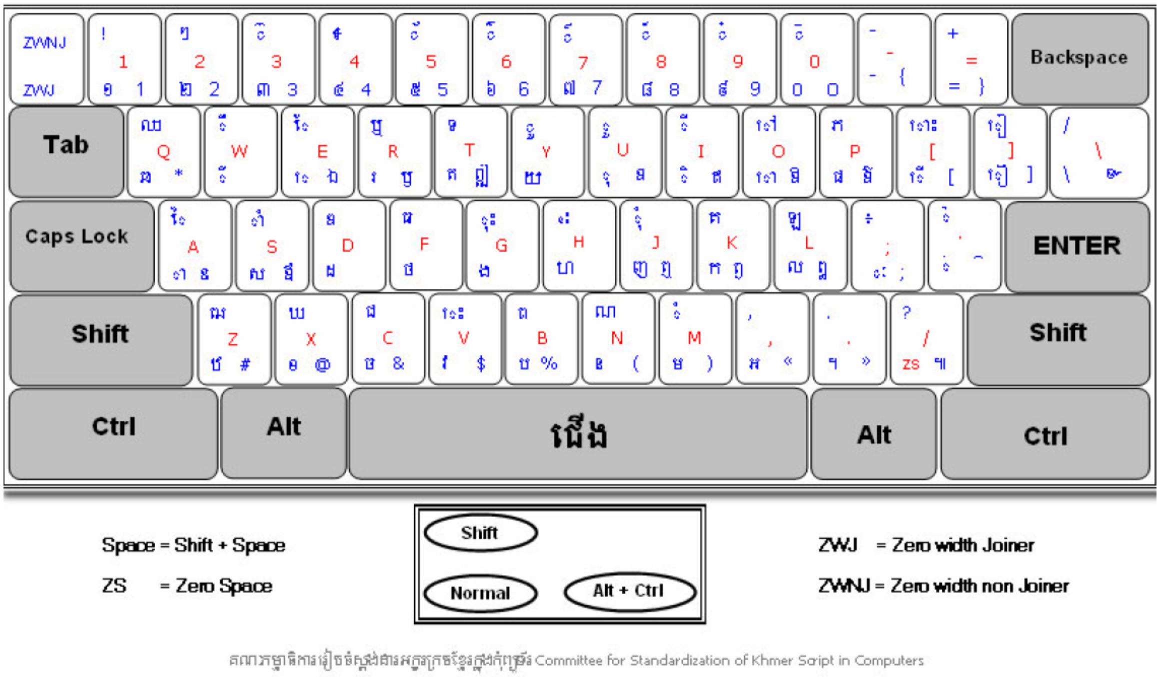
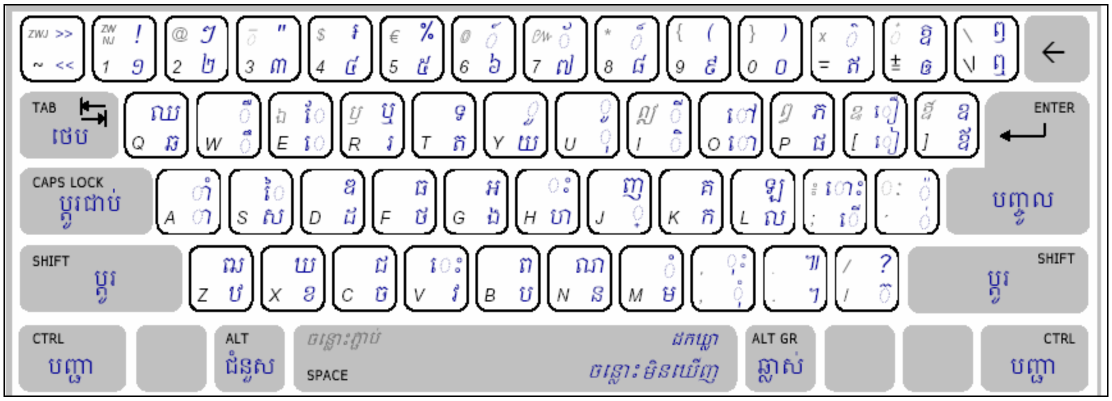


Figure 2: AZERTY French Khmer Keyboard [5]

Additional keyboard layouts are being proposed by various individuals and organizations, especially to cater to Unicode standard, e.g. by National Committee for Standardization Khmer Script in Computers (NCSKSC) and KhmerOS [5] (see Figure 3). Keyboard by NCSKSC has been submitted for standardization.



(a)



(b)

Figure 3: Unicode Based Khmer Keyboards by (a) NCSKSC and (b) KhmerOS [5]

Microsoft Platform

Microsoft does not provide built-in keyboard support for Khmer. However keyboard setups based on MSKLC have been developed (e.g. see Figure 3).

Linux Platform

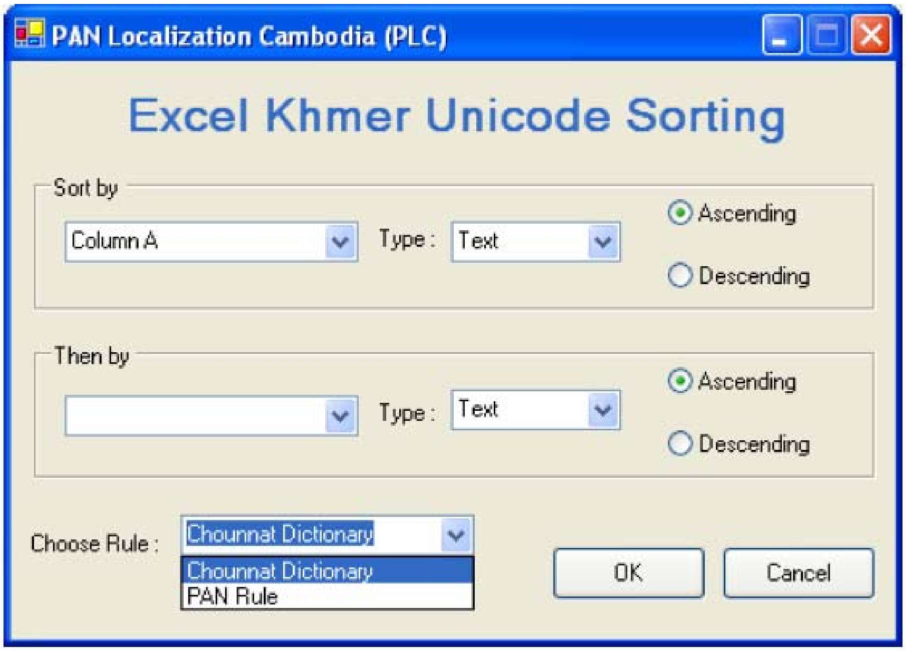
Keyboard driver for Khmer has also been developed. Both AZERTY and KhmerOS versions are available [5].

Collation

Standardization of a single linguistic Khmer sorting sequence is currently under review by national as well as international organizations. This collation is based on the Chuonnat dictionary, the only official Khmer dictionary. A different collation based on Headley's Khmer-English Dictionary is also possible [5].

Microsoft Platform

Microsoft does not support Khmer sorting according to Chuonnat dictionary. However, sorting utility has been developed by PAN Localization Cambodian component [4], which can sort data in Microsoft Office applications. New words, not in Chuonnat dictionary, are sorted through a phonetic mechanism. Screenshots of this application for Word and Excel are given in Figure 4 below.



PAN Localization Cambodia (PLC)

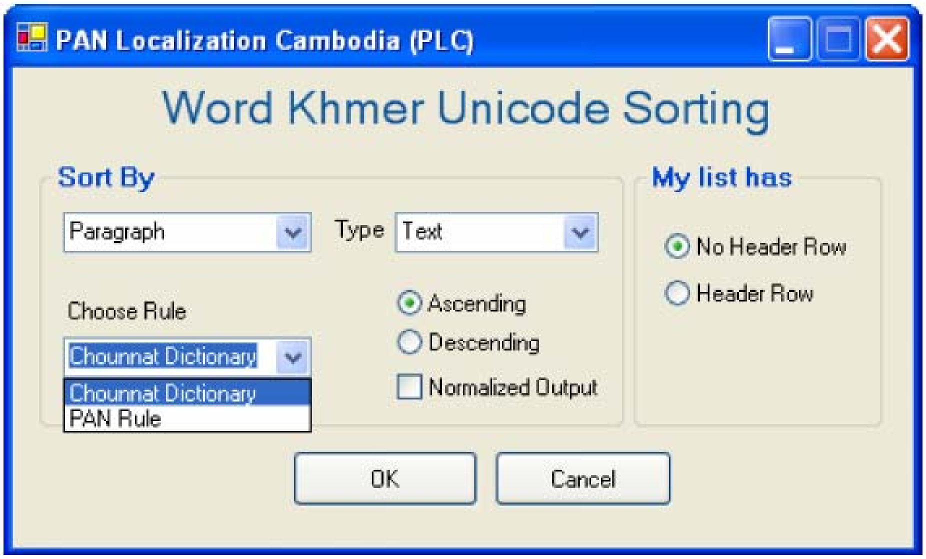
Excel Khmer Unicode Sorting

Sort by: Column A, Type : Text, ● Ascending ○ Descending

Then by: (empty), Type : Text, ● Ascending ○ Descending

Choose Rule : Chounnat Dictionary (options: Chounnat Dictionary, PAN Rule)

(a)



PAN Localization Cambodia (PLC)

Word Khmer Unicode Sorting

Sort By: Paragraph, Type Text

Choose Rule: Chounnat Dictionary (options: Chounnat Dictionary, PAN Rule)

● Ascending ○ Descending □ Normalized Output

My list has: ● No Header Row ○ Header Row

(b)

Figure 4: Khmer Sorting Utility for (a) Excel, and (b) Word for Microsoft Office 2003 by PAN Localization Project

Linux Platform

KhmerOS has also developed collation algorithms based on Headly's and Chuonnat dictionaries which can be compiled on Linux [5] and on other platforms.

Locale

Khmer locale is defined in CLDR 1.3 [8]. It has been incorporated in some Linux platforms [5]. However, there is still no support for Khmer locale in Microsoft.

Interface Terminology Translation

No standard terminology translations exist for Khmer. There is no support on Microsoft platform. However, significant work is being done for open source applications through KhmerOS initiative for Microsoft and Linux platforms. Localization of many open source applications has now been initiated. For example a completely localized beta version of Open Office 2.0 is available in Khmer language. Localized Firefox and Thunderbird, Internet and email clients based on Mozilla, are also available.

Status of Advanced Applications

Work is under progress on development of line breaking algorithm by both PAN Localization and KhmerOS projects. Both are lexically based. Algorithm being developed through PAN Localization is also statistical and based on Khmer corpus.

Khmer corpus and lexicon are also being developed by PAN Localization Cambodian component. The lexicon contains lexemes and additional information, e.g. parts-of-speech. In addition, work is also under progress within this project to develop Khmer a spell checker [4]. .

References

[1] http://www.ethnologue.com/show_language.asp?code=khm
[2] http://www.omniglot.com/writing/khmer.htm
[3] http://www.unicode.org/charts/PDF/U1780.pdf
[4] http://www.PANL10n.net
[5] http://www.khmerOS.info/khmerOS_download.html
[6] http://www.microsoft.com/typography/OpenType%20Dev/Khmer/intro.mspx
[7] http://www.travelphrases.info/gallery/fonts_Khmer.html
[8] http://www.unicode.org/cldr/version/1.3.html

Korean

Korean is the national language of Korea spoken by about 78 million people [1], mostly in North and South Korea. Scholars conflict on the development of Korean language. Few believe that Korean is related to Japanese while others disagree and argue its independent development.

Language Isolate
KOREAN

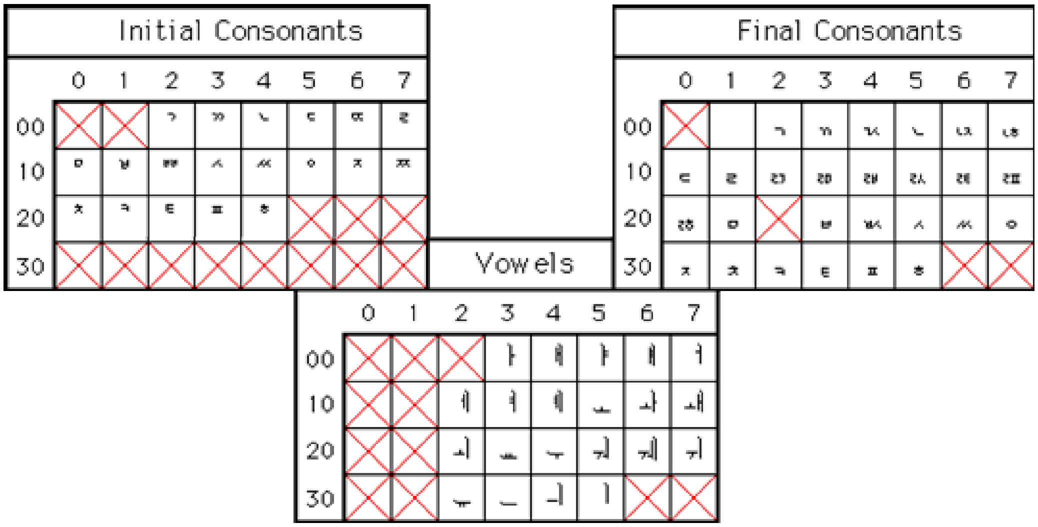
Figure 1: Language Family Tree for Korean [1]

Korean uses two writing systems known as Hangul and Hanja. Hangul is the syllabic writing system used in Korea. Hanja refers to the Chinese ideograph characters. Earlier, since about 5 AD, Korean has been written in Chinese script. The Korean alphabet started during the reign of King Sejong (1418-1450). The alphabet was originally called Hunmin jeongeum, or "The correct sounds for the instruction of the people", but has also been known as Eonmeun (vulgar script) and Gukmeun (national writing). The modern name, Hangeul, was given by a Korean linguist Ju Si-gyeong (1876-1914). The academic papers and official documents tend to be written in a mixture of Hangul and Hanja [2].

Character Set and Encoding

The primary character set encoding is KSC 5601 standard for Hangul and Hanja. This includes 17,100 characters [3]. This was augmented by KSC 5657. However, these standards have been superseded by KSC 5700, which contains the same set of characters as in Unicode. Other Korean standards include ANSI Z39.64 (also called as REAC), EACC, CCCII, CP949, EUC-KR, GB12052, ISO2022-KR, KS C 5636 and MOJIKYO [4]. Also see [6].

One of the encoding that emerged in 1992 is Johab standard. Johab is a 16 bit-code having first bit denoting the Johab encoding, the following 5 bits for initial consonant, next 5 for vowels and the last 5 for final consonant. Following figure shows some relative positions of the symbols. This standard has now been replaced by other standards discussed above [5].



Initial Consonants

	0	1	2	3	4	5	6	7
00	☒	☒	ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ
10	ㅁ	ㅂ	ㅃ	ㅅ	ㅆ	ㅇ	ㅈ	ㅉ
20	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ	☒	☒	☒
30	☒	☒	☒	☒	☒	☒	☒	☒

Final Consonants

	0	1	2	3	4	5	6	7
00	☒		ㄱ	ㄲ	ㄳ	ㄴ	ㄵ	ㄶ
10	ㄷ	ㄹ	ㄺ	ㄻ	ㄼ	ㄽ	ㄾ	ㄿ
20	ㅀ	ㅁ	☒	ㅂ	ㅄ	ㅅ	ㅆ	ㅇ
30	ㅈ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ	☒	☒

Vowels

	0	1	2	3	4	5	6	7
00	☒	☒	☒	ㅏ	ㅐ	ㅑ	ㅒ	ㅓ
10	☒	☒	ㅔ	ㅕ	ㅖ	ㅗ	ㅘ	ㅙ
20	☒	☒	ㅚ	ㅛ	ㅜ	ㅝ	ㅞ	ㅟ
30	☒	☒	ㅠ	ㅡ	ㅢ	ㅣ	☒	☒

Figure 2: Johab Encoding for Korean [5]

Fonts and Rendering

Many Korean fonts are available for most of the computing platforms, which are based on different encodings discussed in the previous section. Bating, Dotum, Gulim, Gungsuh and Arial Unicode are some fonts which work for Microsoft platform. Hanyang font is available for Linux with commercial license terms, Baekmuk font is supplied under BSD license, while Un-font is available under GPL license.

Keyboard and Input Method Engine

KSC 5715 is the standard keyboard layout [6]. However, there are other phonetic layouts also available (e.g. see [7]). Dubeolsik is also a common Hangul keyboard layout in use in South Korea [8]. In addition, various platforms also provide support for entering Korean, which require complex input methods in addition to the keyboard.

Microsoft Platform

Microsoft Windows supports the Korean standard and alternative keyboard layouts. The Korean IME enables users to input Korean (Hangul) and Chinese characters (Hanja). Hangul is entered by Jamos, which are 24 basic elements and combination of elements, on the standard 101 keyboard. By combining these Jamos, all 11,172 Hangul character combinations are produced [9]. Figure 3 shows MS keyboard layout for Korean language.

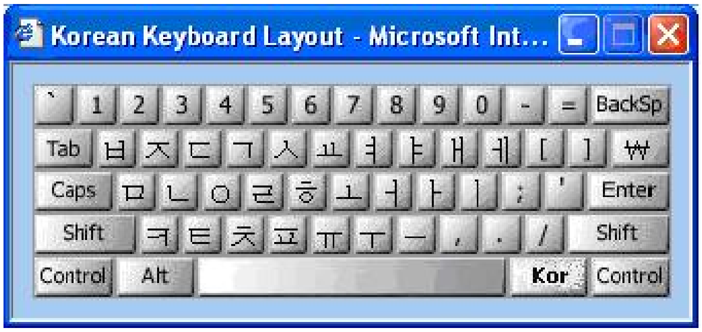
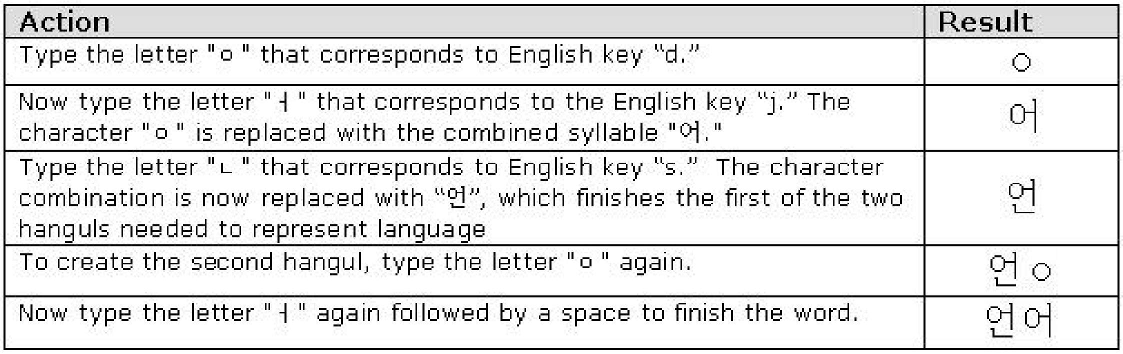

Korean Keyboard Layout - Microsoft Int...
` 1 2 3 4 5 6 7 8 9 0 - = BackSp
Tab ㅂ ㅈ ㄷ ㄱ ㅅ ㅛ ㅕ ㅑ ㅐ ㅔ [ ] ₩
Caps ㅁ ㄴ ㅇ ㄹ ㅎ ㅗ ㅓ ㅏ ㅣ ; ' Enter
Shift ㅋ ㅌ ㅊ ㅍ ㅠ ㅜ ㅡ , . / Shift
Control Alt Kor Control

Figure 3 Korean Keyboard Layout [9]

The following figure explains how the Korean characters are entered into an application.



Action	Result
Type the letter "ㅇ" that corresponds to English key "d."	ㅇ
Now type the letter "ㅓ" that corresponds to the English key "j." The character "ㅇ" is replaced with the combined syllable "어."	어
Type the letter "ㄴ" that corresponds to English key "s." The character combination is now replaced with "언", which finishes the first of the two hanguls needed to represent language	언
To create the second hangul, type the letter "ㅇ" again.	언 ㅇ
Now type the letter "ㅓ" again followed by a space to finish the word.	언어

Figure 4: Process of Input of Korean Characters through MS Korean Keyboard [9]

Once Hangul has been formed, the user can then press a Hanja key that will allow Hangul to be transformed into corresponding Chinese characters. Microsoft also provides IME Pad that allows the user to input Hangul via soft keyboard, shown in Figure 5. In addition, Microsoft also provides a handwriting recognition based interface to input Korean, also shown in Figure 5.

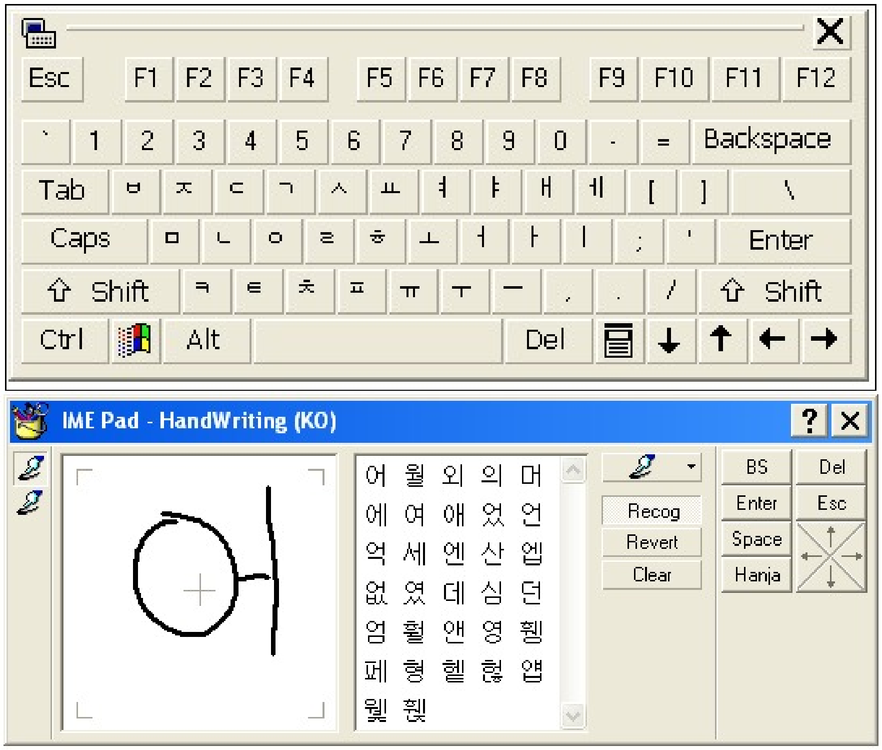

Esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12
` 1 2 3 4 5 6 7 8 9 0 - = Backspace
Tab ㅂ ㅈ ㄷ ㄱ ㅅ ㅛ ㅕ ㅑ ㅐ ㅔ [ ] \
Caps ㅁ ㄴ ㅇ ㄹ ㅎ ㅗ ㅓ ㅏ ㅣ ; ' Enter
Shift ㅋ ㅌ ㅊ ㅍ ㅠ ㅜ ㅡ , . / Shift
Ctrl Alt Del
IME Pad - HandWriting (KO)
어 월 외 의 머
에 여 애 었 언
억 세 엔 산 엡
없 였 데 심 던
엄 훨 앤 영 휑
페 형 헬 헗 앱
웰 휀
Recog
Revert
Clear
BS Del
Enter Esc
Space
Hanja

Figure 5: Soft Pad for Korean and Handwriting Recognition Based Input Method by Microsoft [7]

Linux Platform

Korean input methods in Linux support both the standard and alternative keyboards. The standard keyboard is widely used. "Ami" is an input method used for Korean [12].

Collation

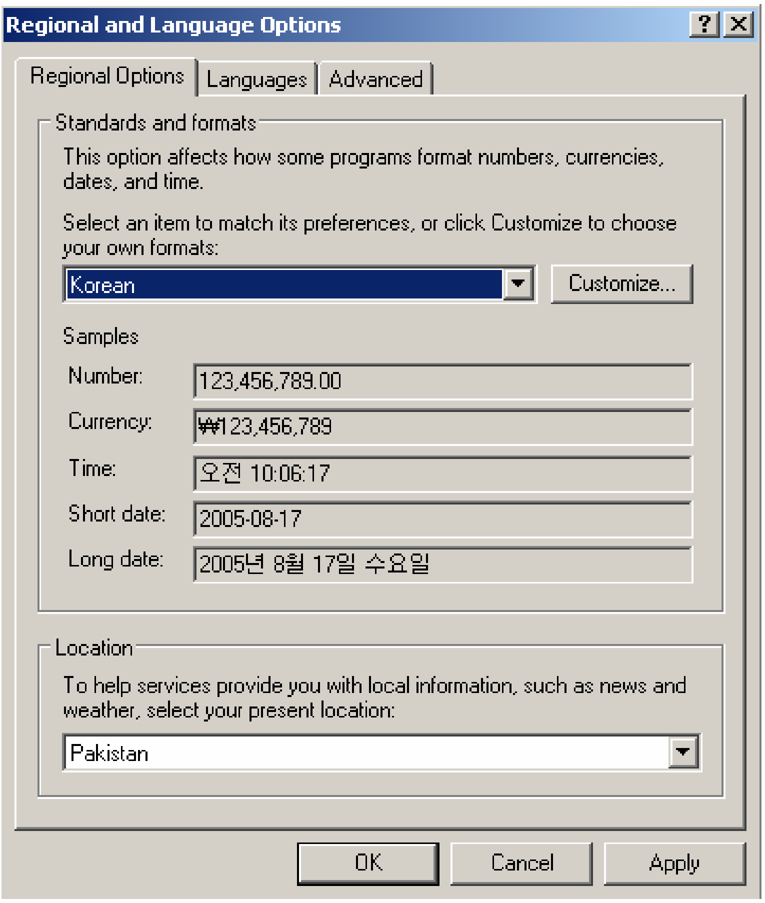
Korean collation is supported in all MS applications, Linux platforms, and also defined in Glibc. There is some work on defining a standard collation of Korean. See [10] for Korean collation methodology. Collation rules are also defined in the locale definition of Korean (see below).

Locale

Standard Locale for Korean has been defined in IBM ICU. The code is ko_KR (for South Korea). This locale definition constitutes localized settings for date, time, days, months, currency symbols, number format and collation [10]. CLDR 1.3 also includes a locale for Korean.

Microsoft Platform

Microsoft Windows XP provides support for Korean locale. If system locale is switched to Korean, changes may be monitored in the date, time, and currency symbols within all application of Microsoft Windows. The following figure shows Korean locale on MS platform.



Standards and formats

This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Korean

Samples

Number:	123,456,789.00
Currency:	₩123,456,789
Time:	오전 10:06:17
Short date:	2005-08-17
Long date:	2005년 8월 17일 수요일

Location

To help services provide you with local information, such as news and weather, select your present location:

Pakistan

Figure 6: Korean Locale on Microsoft Platform

Linux Platform

Korean locale is also available on Linux platforms, e.g. see [12].

Interface Terminology Translation

Most of famous Linux distributions such as RedHat Linux, Fedora Core, Debian GNU/Linux, Mandrake and Novell Linux (or Suse Linux) support Korean. In addition, there are also Linux distributions by Korean vendors, which include Hansoft Linux, Hancom Linux, and Wow Linux.

Microsoft Platform

A complete version of MS Windows is shipped in Korean. The Korean Windows has translated interface, menus and dialogue boxes.

Linux Platform

Terminology translation for localization of GNOME is managed by GNOME Korean user community [13]. 94% for GNOME 2.12 has been translated for Korean language [14] while 14% of KDE has been done [15]. Mozilla terminology is maintained by volunteers under the Hangul Mozilla project [16]. On the Linux platform the Open Office interface and help files have also been localized [17]. There is a significant push by Korean government to develop open source platforms and many other initiatives are under progress [18].

Status of Advanced Applications

Significant language resources are available for Korean language, including lexicons, corpora (e.g. [19]), spell checkers and grammar checkers. Korean grammar checker has been embedded in Microsoft products (e.g. Word) [20] and Hangul (Korean-made word processing program). Word breaking utilities, stemmers and morphological parsers are also available, e.g. [21].

Korean text-to-speech systems [22, 23], and speech recognition systems [24, 25] are also commercially available. OCR software and handwriting recognition systems for Korean are widely available for a variety of platforms [26]. There are also Korean machine translation systems available for multiple languages, e.g. [27, 28].

References

[1] http://www.ethnologue.com/14/show_language.asp?code=KKN
[2] http://www.omniglot.com/writing/korean.htm
[3] http://www.ascendercorp.com/cjk.html
[4] http://www.jbrowse.com/text/charsets.html
[5] http://homepages.cwi.nl/~dik/english/codes/stand.html
[6] Hyun-Il, K. "Trends and Plans for the IT Standardization in Korea," in the *Proceedings of 5th AFSIT,* Tokyo, Japan, 1991. Available at http://www.cicc.or.jp/english/hyoujyunka/af05/5-07.html
[7] http://www.freepatentsonline.com/6462678.html
[8] http://www.absoluteastronomy.com/encyclopedia/k/ke/keyboard_layout.htm
[9] http://www.microsoft.com/globaldev/handson/user/IME_Paper.mspx
[10] http://www.open-std.org/jtc1/sc22/wg20/docs/n1037-Hangul%20Collation%20Requirements.htm
[11] http://www-950.ibm.com/software/globalization/icu/demo/locales/?d_=en&_=ko
[12] http://www.freshports.org/korean/
[13] http://www.GNOME.or.kr/
[14] http://l10n-status.GNOME.org/GNOME-2.12/ko/index.html
[15] http://i18n.KDE.org/stats/gui/stable/ko/
[16] http://www.mozilla.or.kr/
[17] http://projects.openoffice.org/native-lang.html
[18] http://www.ima.umn.edu/~klee/kr.linux.html
[19] www.ldc.upenn.edu
[20] http://www.translation.net/microsoft_proof_2003.html

[21] http://www.teragram.com/oem/asian_lang.htm#korean
[22] http://www.microsoft.com/MSAGENT/downloads/user.asp
[23] http://www.voicesignal.com/news/press/release_12_03_03.html
[24] http://www.mobileburn.com/news.jsp?Id=912&source=BROWSER
[25] http://www.asahi-kasei.co.jp/vorero/en/news/
[26] http://www.worldlanguage.com/Products/Korean/OCR/Page1.htm
[27] http://world.altavista.com/
[28] http://www.worldlingo.com/en/products_services/worldlingo_translator.html

Lao

Lao is a Thai-Kadai language spoken by approximately 15 million people in Laos and Thailand [1]. It is one of the many Tai languages of the Southwestern branch. The origins of the Tai language can be traced back to Guangxi-Guizhou-Hunan region in southern China and bordering areas of northern Vietnam about 2000 years ago [2].

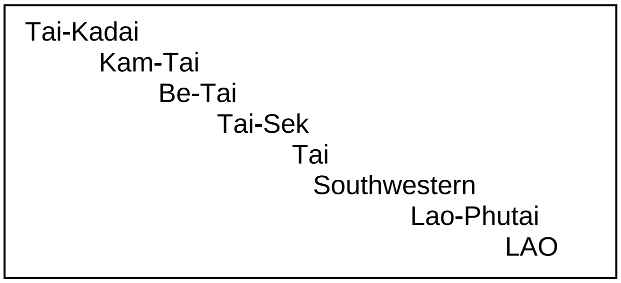


Figure 1: Language Family Tree for Lao [1]

Traditionally, Lao language and its literature have been written in two scripts, Lao and Tham. Lao script has been largely influenced by the ancient Thai script that was developed from the old Khmer script [2].

Character Set and Encoding

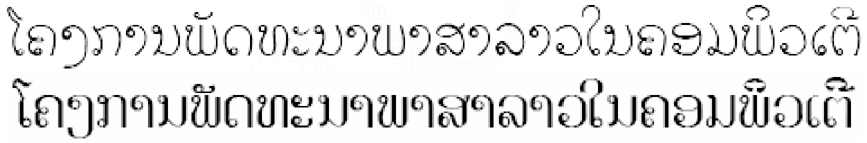
Lao computing has progressed slowly due to lack of Lao encoding standards [3]. Laos' Science Technology and Environment Agency (STEA) is currently working towards IT standardization. Due attention is being given to formulation and implementation of national standards and adoption of available international standards [3]. Unicode has also included Lao script from 0E80-0EFF but is not widely used. Current ad hoc standard is based on encoding developed Lao Script for Windows [4]. Earlier Lao encoding includes IBM ISO-8 Code Page 01033 and EBCDIC Code Page 01032 developed by STEA [3].

Fonts and Rendering

Multiple fonts are available for Lao based on ad hoc encodings [4] and based on Unicode standard (e.g. [5]), in TTF and OTF formats. Even though basic research in Lao font development has been conducted, advanced issues e.g. mark to mark positioning and hinting, are still unresolved. The most widely used font for Lao is Saysettha OT. It is developed by Lao Script for Windows [4] and includes proper hinting. STEA, through PAN Localization project, has also developed two Lao fonts [5].

Microsoft Platform

Windows XP does not provide in-built fonts that support Lao script. Locally some of the fonts have been developed that adequately satisfy Lao script requirements. The following figure presents the Lao font rendering on Microsoft platform.



ໂຄງການພັດທະນາພາສາລາວໃນຄອມພິວເຕີ

ໂຄງການພັດທະນາພາສາລາວໃນຄອມພິວເຕີ

Figure 2: Two Lao fonts Displayed on Microsoft Platform [5]

Linux Platform

Several Lao Open Type fonts are developed to be used on the Linux platform. Popularly used are Saysettha OT and Jason variety Lao Unicode fonts. These can also be rendered in Open Office and Mozilla.

Keyboard Standard

No keyboard layout has been standardized for Lao language. However there are few ad hoc standards that are popularly used for Lao text input. A commonly used Lao keyboard is based on the Lao typewriter layout. Figure 3 shows the keyboard layout based on the typewriter.

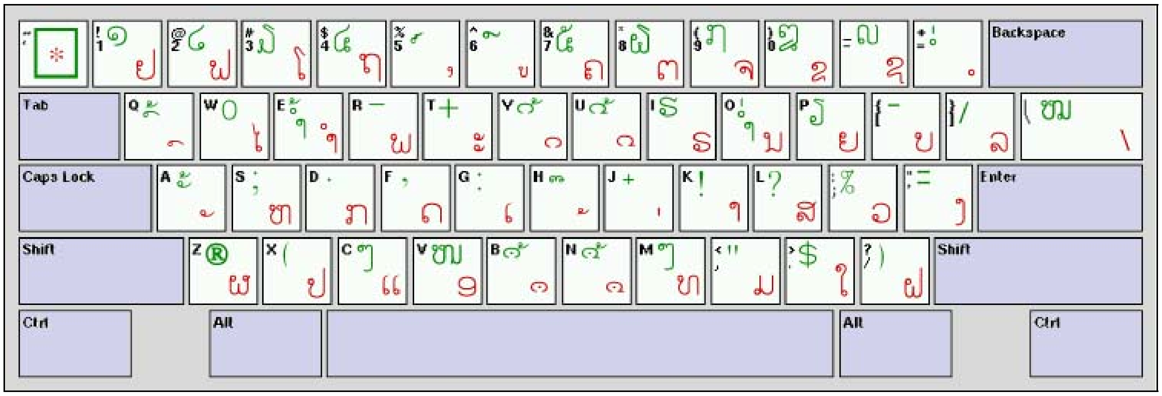


Figure 3: The Old Typewriter Based Keyboard Layout [6]

Lao Script for Windows is a shareware Lao solution for Windows. Its LSWin version 5.0 has a separate keyboard interface specifically designed for Lao input [4].

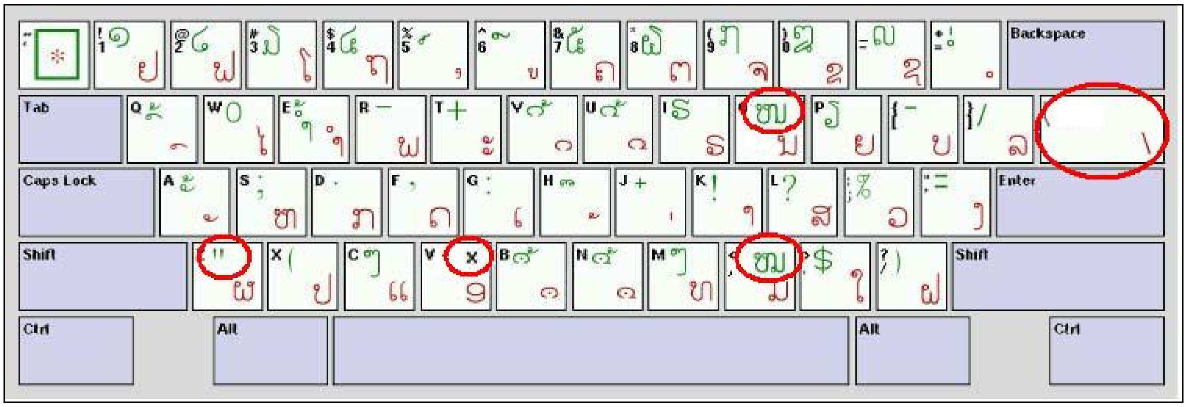
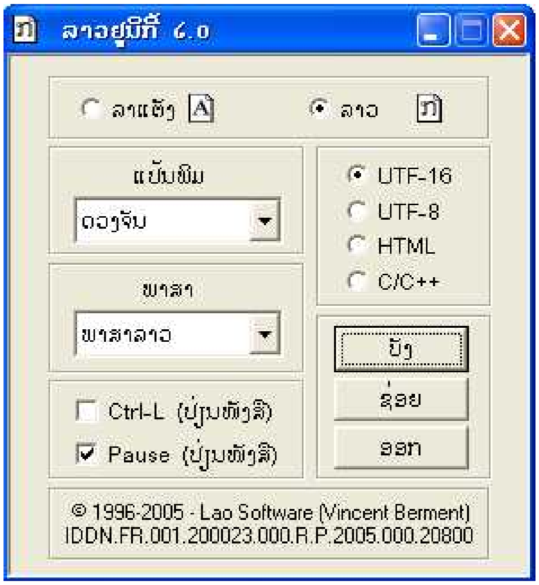


Figure 4: Keyboard Layout Based for LSWIN [4]

Lao Software [7] has developed Lao Unikey, the Lao keyboard drivers compatible with Unicode. Five popularly used Lao keyboards have been included in Lao Unikey. They are Duang Jan, Lao US, Lao France, Lao France New and Sida Thong. Figure 5 shows the Lao Unikey developed by Lao software. Other keyboards are also available, e.g. [6].



ລາວຢູນິກີ້ ໖.໐

○ ລາແຕັງ ● ລາວ

ແປ້ນພິມ: ດວງຈັນ

ພາສາ: ພາສາລາວ

● UTF-16 ○ UTF-8 ○ HTML ○ C/C++

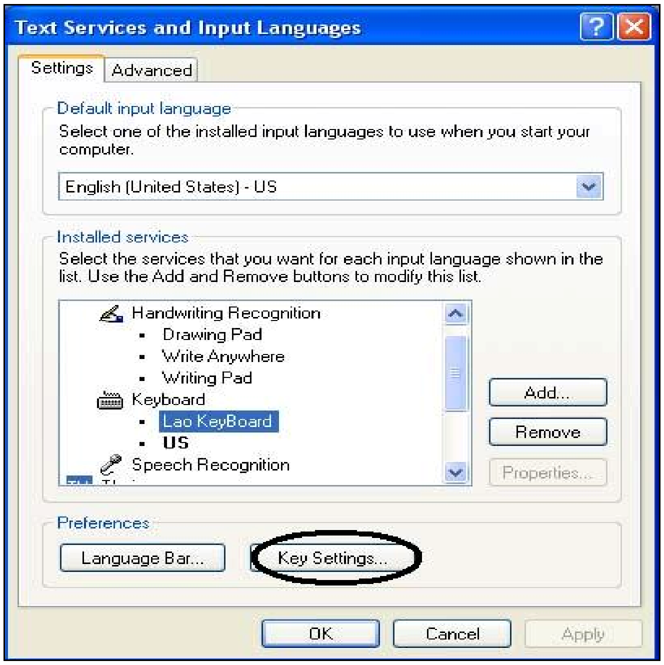
□ Ctrl-L (ປຸ່ມທັງສີ)

☑ Pause (ປຸ່ມທັງສີ)

© 1996-2005 - Lao Software (Vincent Berment)
IDDN.FR.001.200023.000.R.P.2005.000.20800

Figure 5: Lao Unikey [7]

Microsoft Windows does not provide keyboard for Lao language. However many Lao localization groups have developed keyboards for Lao. The following figure displays the facility to enable Lao keyboard support on Windows platform.



Text Services and Input Languages

Default input language

Select one of the installed input languages to use when you start your computer.

English (United States) - US

Installed services

Select the services that you want for each input language shown in the list. Use the Add and Remove buttons to modify this list.

- Handwriting Recognition
  - Drawing Pad
  - Write Anywhere
  - Writing Pad
- Keyboard
  - Lao KeyBoard
  - **US**
- Speech Recognition

Figure 6: Setting-up Lao Keyboard on MS Windows

Keyboards based on popularly used Lao ad hoc encodings are being developed for KDE and GNOME through LaoNux project [8].

Collation Sequence

There is no collation standard for Lao language. However two most popularly used schemes are based on lexical order in dictionaries. Projects are underway at National University of Laos and STEA, through PAN Localization project, to standardize the collation order of Lao.

Lao sorting utility has been developed by STEA based on lexicographic orders. It does syllable based sorting, after rule-based syllabification of input string. Sorting utility is also available through LSWIN and Lao Software projects. These orders are based on existing dictionaries.

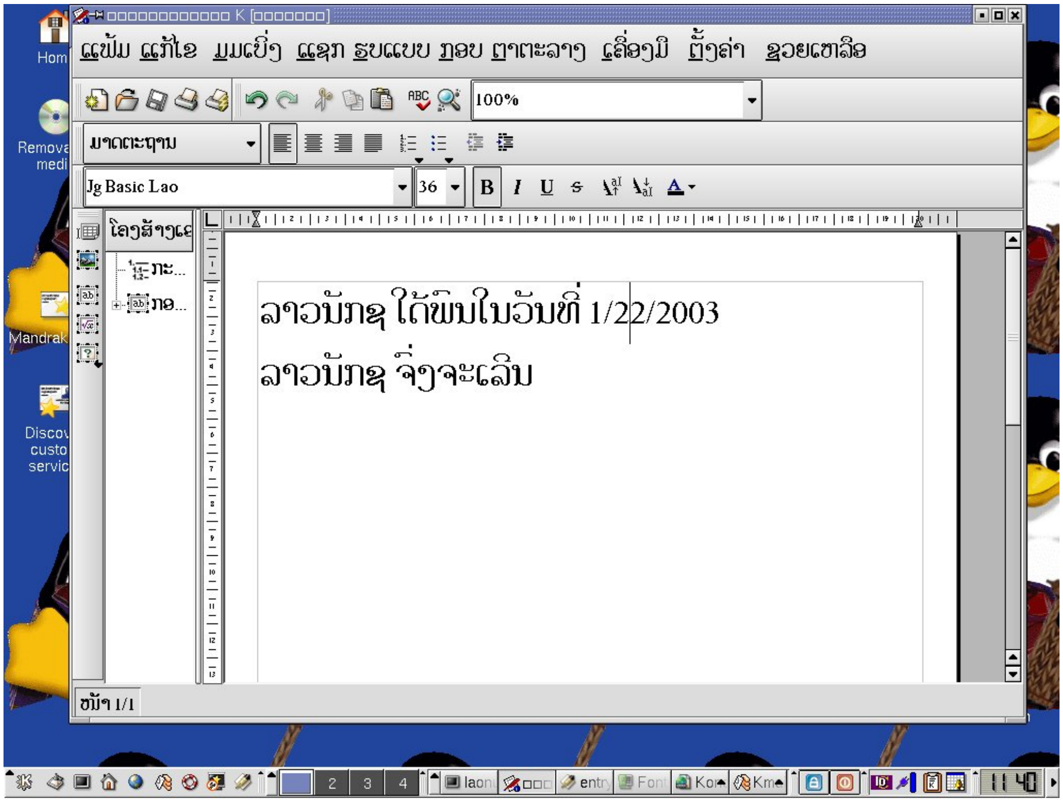
Locale

CLDR 1.3 includes locale definitions for Lao (lo_LA). It includes date, time, currency symbol and names of the days and months. In addition the National University of Laos (NUoL) is working on compiling NLS file for Lao language. Currently Lao locale is not included in any version of Microsoft Windows. Locale file is being developed under LaoNux project to enable Lao locale on Linux platform. Until now the date, time and currency formats have been defined. LaoNux is a KDE based Lao Linux distribution [8].

Interface Terminology

No terminology standard exists for Lao. Lao version does not exist for Microsoft software. A Lao Editor has been developed on Microsoft platform by STEA through PAN Localization project. The Lao Editor provides Lao and English interface. Help files have also been translated in Lao language.

The LaoNux project team is working to develop Lao technical glossary for localization of open source software. Currently there are no considerable outputs on the glossary translation of Linux applications. The LaoNux project is at its initial phases [8]. There is also some work on KDE translation of Lao, as shown in the figure below.


ແຟ້ມ ແກ້ໄຂ ມຸມເບິ່ງ ແຊກ ຮູບແບບ ກອບ ຕາຕະລາງ ເຄື່ອງມື ຕັ້ງຄ່າ ຊ່ວຍເຫລືອ
100%
ມາດຕະຖານ
Jg Basic Lao
36
ໜ້າ 1/1

ລາວນັກຊ ໄດ້ພົບໃນວັນທີ່ 1/22/2003

ລາວນັກຊ ຈົ່ງຈະເລີນ

Figure 7: Localized KDE environment in Lao [9]

Status of Advanced Applications

Some research is presently underway for the development of Lao language processing applications. STEA, through PAN Localization project, has already developed encoding convertors, Lao syllabification and sorting utility and a Lao lexicon, all incorporated within a Lao text editor. Lao lexicon is also available through Lao Software [7]. Lao Software has developed support for Lao online translation service. The Lao translation applications developed include, Lao- French-Lao word translation and transcription of texts utility from Thai to Lao [7].

Lao script does not have space between words. Thus, line breaking utilities are needed for basic word processing. LSWIN has developed a lexicon based utility. STEA has also developed a utility based on the syllable structure in Lao language [5]. Reordering is also done during line breaking algorithm when some of the vowels and tone marks are used in reverse order. Work is under progress on more advanced applications through PAN Localization project.

References

[1] http://www.ethnologue.com/show_language.asp?code=lao
[2] http://www.omniglot.com/writing/lao.htm
[3] http://www.undplao.org/governance/ProjectDocuments/ICT%20ProDoc.pdf
[4] http://www.laoscript.net/
[5] http://www.laol10n.info.la/Eng.htm
[6] http://www.laohub.com/static_html/lang/keyboard/laokey05v11.pdf
[7] http://www.laosoftware.com/index.php?Langue=en
[8] http://opensource.muanglao.com/laonux.htm
[9] http://i18n.KDE.org/screenshots/

Mongolian

Mongolian is an Altaic language spoken by approximately 2 million speakers in Mongolia [1]. Khalkha Mongolian (the most widely used dialect for Mongolian language) is the national language of Mongolia. It is also spoken in some parts of China, Afghanistan and Russia [1]. The figure given below shows the language family tree of Mongolian language.

- Altaic
  - Mongolian
    - Eastern
      - Oirat-Khalkha
        - Khalkha-Buriat
          - MONGOLIAN PROPER

Figure 1: Language Family Tree for Mongolian [1]

Traditional Mongolian alphabet was derived from Uighur alphabet in the 12th century, which in turn evolved from Sogdian alphabet, which ultimately came from Aramaic. Mongolian has also been written in Chinese, Arabic and Tibetan alphabet. Eventually, due to Soviet Union, Latin alphabet was adopted in 1931 and eventually changed to Cyrillic alphabet, which is still widely used. In 1941, Mongolian government abolished Mongolian alphabet. However, since 1994 there is move to revive the Mongolian alphabet [2, 3]. Traditionally, Mongolian script is written from top to bottom, left to right [2, 20].

Character Set and Encoding

Mongolian, written in the Mongolian script uses Unicode code block 1800-18AF as the standard encoding [4]. Mongolian written in Cyrillic script uses the Unicode Cyrillic code page 0400-04FF as the standard encoding.

For the Mongolian character set based on Cyrillic script adopted, Microsoft code page 866 and Windows code page 1251 are adopted as national standards [5]. Some guidelines to convert between these encodings are available at [9]. Other encodings include ISO 8859-5 and KO18 [14].

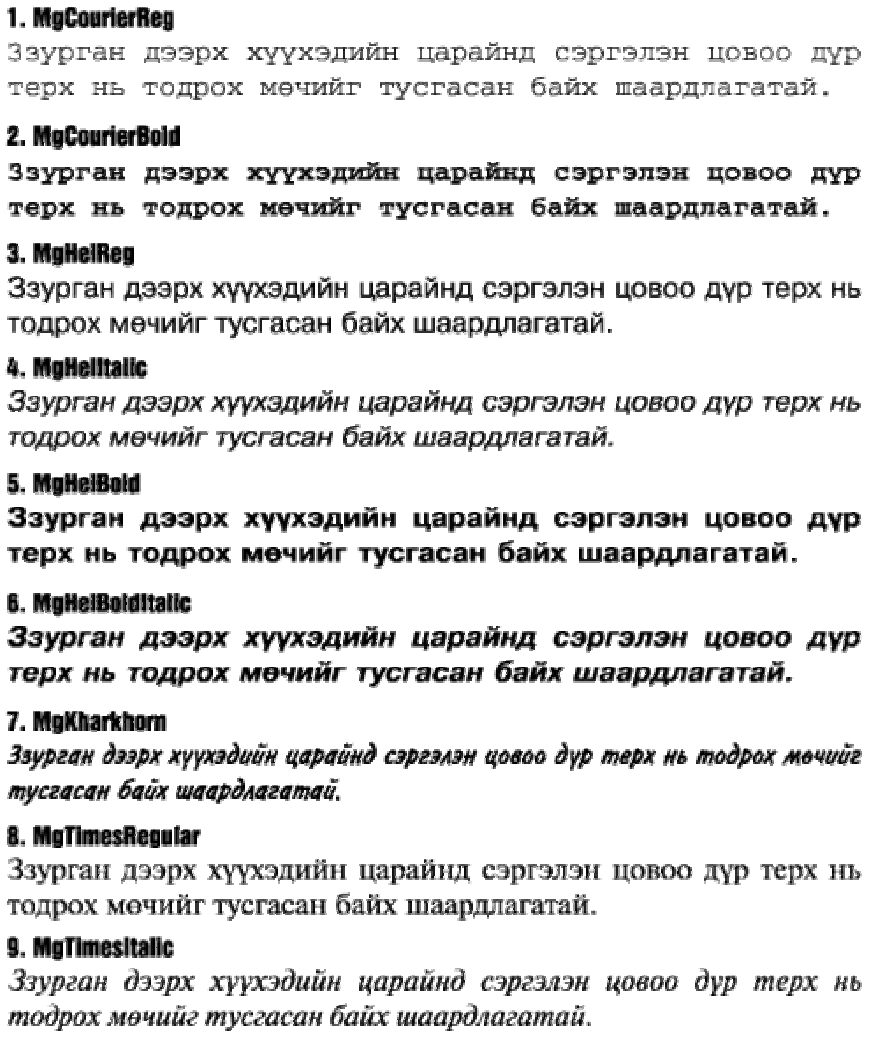
The Mongolian national standards development body, National Center for Standardization and Metrology (MNCSM) has recently adopted 31 ISO standards on IT terminology as their Mongolian national standards. For Mongolian script, the ISO/IEC 10646 (Unicode) character encoding standard has been adopted as the national standard [4].

Fonts and Rendering

Cyrillic script fonts are available through vendors. Mongolian script fonts are fewer and not as readily available. Both are rendered well on various platforms (see [7, 13] for a list of (mostly) Cyrillic fonts based on various encoding schemes).

Microsoft Platform

Cyrillic script is supported in the Microsoft platform. Many Mongolian language Cyrillic fonts have also been developed by other research groups. Figure 2 shows few of the Cyrillic script fonts for Mongolian on Microsoft platform.



1. MgCourierReg

Ззурган дээрх хүүхэдийн царайнд сэргэлэн цовоо дүр терх нь тодрох мөчийг тусгасан байх шаардлагатай.

2. MgCourierBold

Ззурган дээрх хүүхэдийн царайнд сэргэлэн цовоо дүр терх нь тодрох мөчийг тусгасан байх шаардлагатай.

3. MgHelReg

Ззурган дээрх хүүхэдийн царайнд сэргэлэн цовоо дүр терх нь тодрох мөчийг тусгасан байх шаардлагатай.

4. MgHelItalic

Ззурган дээрх хүүхэдийн царайнд сэргэлэн цовоо дүр терх нь тодрох мөчийг тусгасан байх шаардлагатай.

5. MgHelBold

Ззурган дээрх хүүхэдийн царайнд сэргэлэн цовоо дүр терх нь тодрох мөчийг тусгасан байх шаардлагатай.

6. MgHelBoldItalic

Ззурган дээрх хүүхэдийн царайнд сэргэлэн цовоо дүр терх нь тодрох мөчийг тусгасан байх шаардлагатай.

7. MgKharkhorn

Ззурган дээрх хүүхэдийн царайнд сэргэлэн цовоо дүр терх нь тодрох мөчийг тусгасан байх шаардлагатай.

8. MgTimesRegular

Ззурган дээрх хүүхэдийн царайнд сэргэлэн цовоо дүр терх нь тодрох мөчийг тусгасан байх шаардлагатай.

9. MgTimesItalic

Ззурган дээрх хүүхэдийн царайнд сэргэлэн цовоо дүр терх нь тодрох мөчийг тусгасан байх шаардлагатай.

Figure 2: Unicode Cyrillic Script Fonts for Mongolian [7]

Mongolian script fonts are presently not provided by Microsoft, but Microsoft does provide some guidelines to develop these fonts [8]. Mongolian script Unicode based fonts have been developed by other vendors. Figure 3 shows a sample of Mongolian font rendered on MS platform.

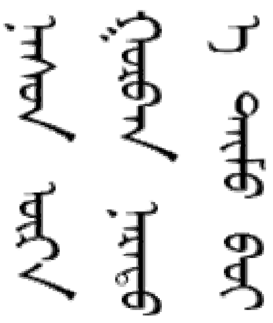


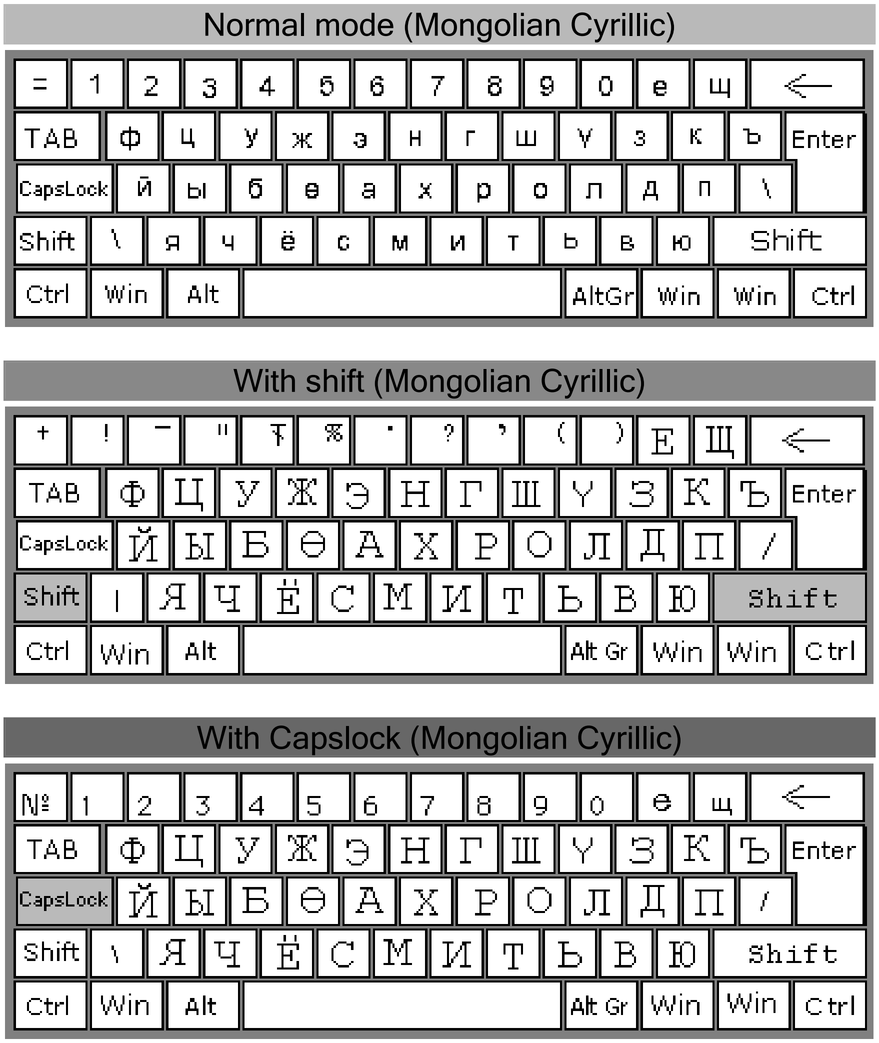
Figure 3: Unicode Mongolian Script Fonts [10]

Linux Platform

Mongolian Cyrillic script fonts are available and supported on Linux platform. However, there is no support for Mongolian script fonts. See [11, 12] for details.

Keyboard

Cyrillic keyboard standard is available. In addition, phonetic based and other layouts are also used, e.g. see [15, 16, 17]. These keyboards are available for all the different encoding used for Cyrillic. Keyboard standard for Mongolian script is not available. However, ad hoc layouts are available, e.g. the one shown in Figure 4.


Normal mode (Mongolian Cyrillic)
= 1 2 3 4 5 6 7 8 9 0 е щ ←
TAB ф ц у ж э н г ш ү з к ъ Enter
CapsLock й ы б ө а х р о л д п \
Shift \ я ч ё с м и т ь в ю Shift
Ctrl Win Alt AltGr Win Win Ctrl
With shift (Mongolian Cyrillic)
+ ! - " ₮ % . ? , ( ) Е Щ ←
TAB Ф Ц У Ж Э Н Г Ш Ү З К Ъ Enter
CapsLock Й Ы Б Ө А Х Р О Л Д П /
Shift | Я Ч Ё С М И Т Ь В Ю Shift
Ctrl Win Alt Alt Gr Win Win Ctrl
With Capslock (Mongolian Cyrillic)
№ 1 2 3 4 5 6 7 8 9 0 е щ ←
TAB Ф Ц У Ж Э Н Г Ш Ү З К Ъ Enter
CapsLock Й Ы Б Ө А Х Р О Л Д П /
Shift \ Я Ч Ё С М И Т Ь В Ю Shift
Ctrl Win Alt Alt Gr Win Win Ctrl

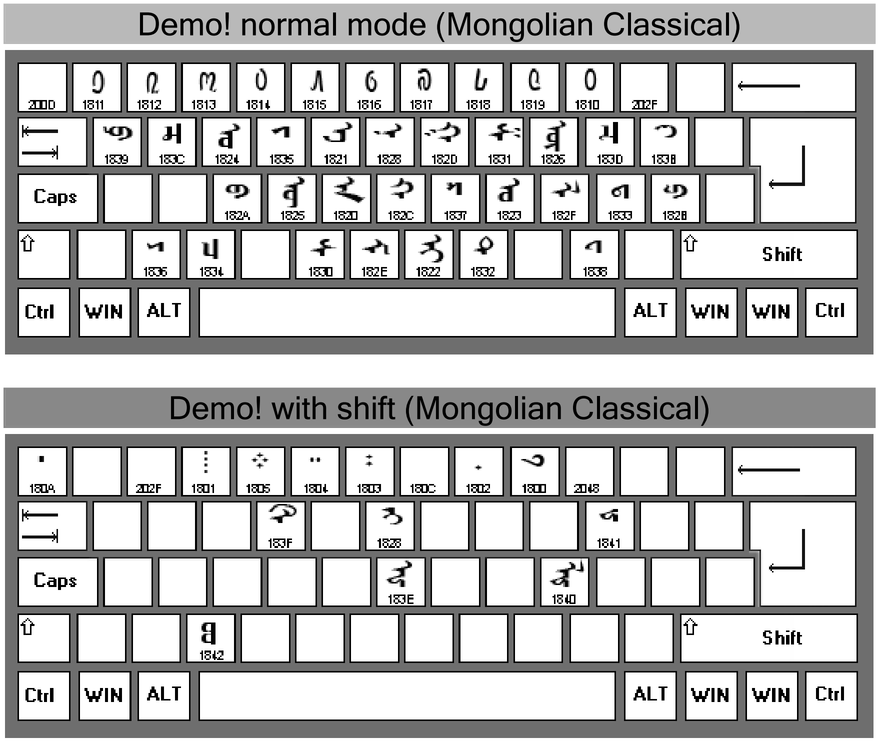
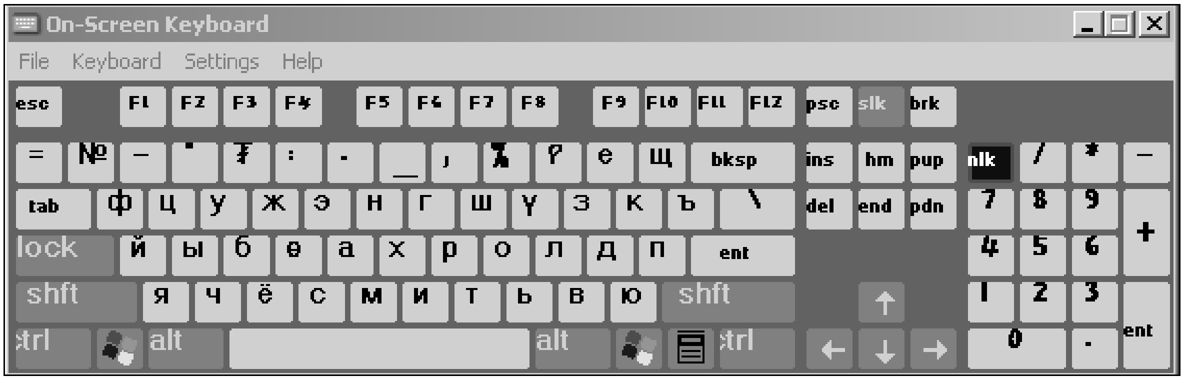

Demo! normal mode (Mongolian Classical)
200D 1811 1812 1813 1814 1815 1816 1817 1818 1819 1810 202F
1839 183C 1824 1835 1821 1828 182D 1831 1826 183D 1838
Caps 182A 1825 1820 182C 1837 1823 182F 1833 182B
1836 1834 1830 182E 1822 1832 1838 Shift
Ctrl WIN ALT ALT WIN WIN Ctrl
Demo! with shift (Mongolian Classical)
180A 202F 1801 1805 1804 1803 180C 1802 1800 2048
183F 1828 1841
Caps 183E 1840
1842 Shift
Ctrl WIN ALT ALT WIN WIN Ctrl

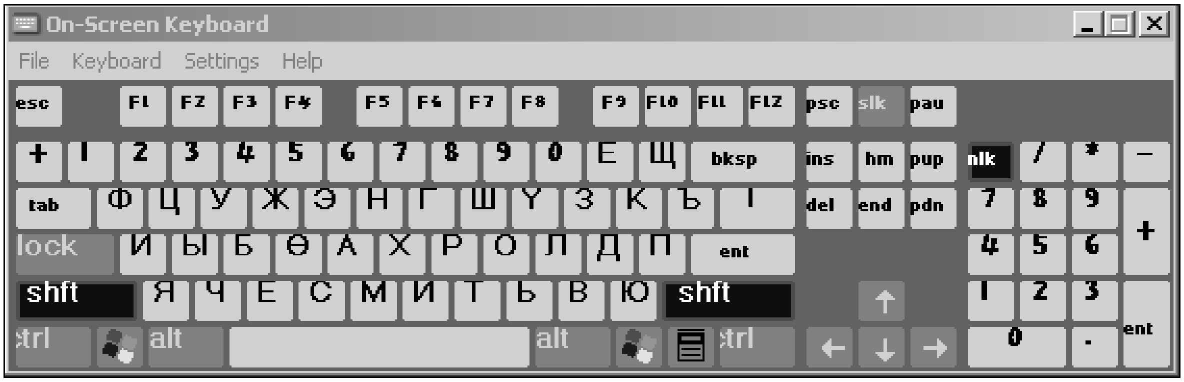
Figure 4: Cyrillic and Mongolian Script Keyboard Layouts for Mongolian [19]

Microsoft Platform

Microsoft Windows XP provides Cyrillic based in-built support for Mongolian keyboard. Once this keyboard layout is enabled it will facilitate Mongolian (Cyrillic) text input. Figure 5 below shows Mongolian keyboard layout (normal state and shift state) on Microsoft platform (also see [18]).


On-Screen Keyboard
File Keyboard Settings Help
esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 psc slk brk
= № – " ₮ : . _ , % ? е щ bksp ins hm pup nlk / * –
tab ф ц у ж э н г ш ү з к ъ \ del end pdn 7 8 9 +
lock й ы б ө а х р о л д п ent 4 5 6
shft я ч ё с м и т ь в ю shft 1 2 3 ent
ctrl alt alt ctrl 0 .

(a)


On-Screen Keyboard
File Keyboard Settings Help
esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 psc slk pau
+ I 2 3 4 5 6 7 8 9 0 Е Щ bksp ins hm pup nlk / * −
tab Ф Ц У Ж Э Н Г Ш Ү З К Ъ I del end pdn 7 8 9 +
lock И Ы Б Ө А Х Р О Л Д П ent 4 5 6
shft Я Ч Е С М И Т Ь В Ю shft 1 2 3
ctrl alt alt ctrl 0 . ent

(b)

Figure 5: (a) Normal and (b) Shift State for Mongolian (Cyrillic) Keyboard on Microsoft Platform

Linux Platform

There are multiple keyboards available for different encodings for Cyrillic on Linux platform (e.g. see [17]). However Mongolian script keyboards are not available.

Collation

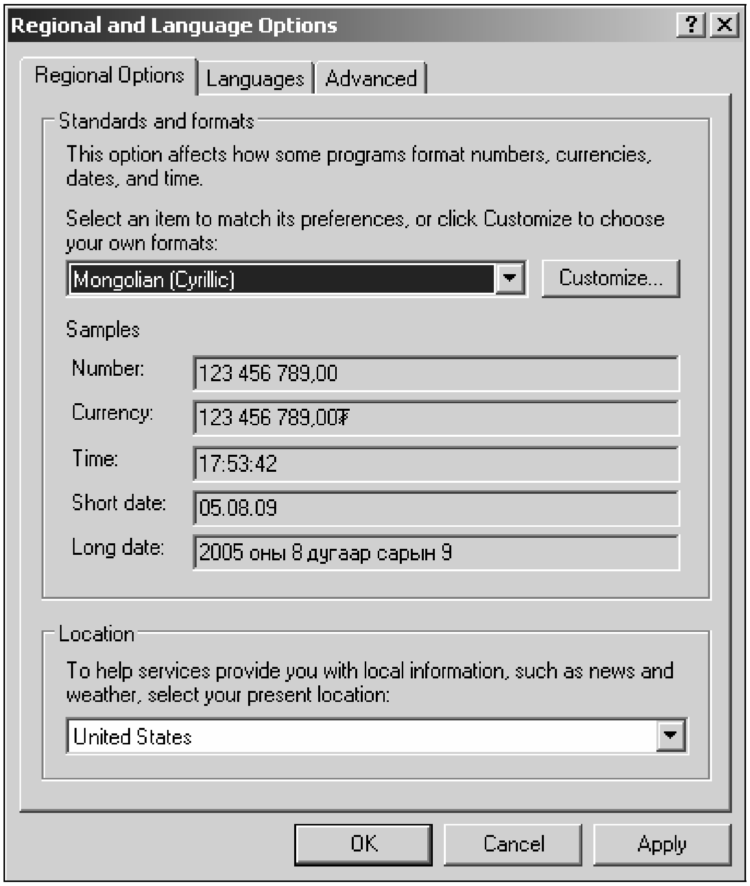
Collation for Cyrillic has been defined. However, there is no standard for Mongolian script. Cyrillic is supported by Microsoft and Linux platforms.

Locale

Locale for Mongolian has not been defined in IBM ICU [21]. CLDR 1.3 includes a locale for Mongolian, which is partially defined. The locale for Mongolian is mn_MN.

Microsoft Platform

Microsoft Windows XP provides support for Mongolian (Cyrillic) locale. If system locale is switched to Mongolian, changes may be monitored in the date, time, and currency symbol within all application of Microsoft Windows. Figure 6 shows an example of setting user locale to Mongolian for Microsoft Windows XP. Mongolian script based locale is not available.



Standards and formats

This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Mongolian (Cyrillic)

Samples

Number:	123 456 789,00
Currency:	123 456 789,00₮
Time:	17:53:42
Short date:	05.08.09
Long date:	2005 оны 8 дугаар сарын 9

Location

To help services provide you with local information, such as news and weather, select your present location:

United States

Figure 6: Mongolian (Cyrillic) Locale on Microsoft Platform

Linux Platform

Cyrillic based locale is also defined for Linux platform for Mongolian language, e.g. see [22].

Interface Terminology Translation

Mongolian interface is not available on Microsoft platform [23, 24]. Majority of localization work done on Linux platform has been done for Cyrillic. A complete Mongolian Linux distribution, Soyombo, with completely localized Mozilla has been developed by the OpenMN research Group for Cyrillic [12]. In addition the localization teams for Mongolian have also registered for translating GNOME 2.2 where about 99% of Developers.lib and 74% of Desktop strings have been translated [25]. Cyrillic glossary translation for KDE [26] and Open Office [27] in Mongolian is also under progress.

Status of Advanced Applications

There are a few Mongolian-English dictionaries available on the Internet. Some links are listed in [2]. Though most of them are in Cyrillic, some are also in Mongolian script [28]. However there is very limited work on Mongolian within Mongolia on other language technologies. Spell checkers are available [29]. Some work has been initiated on Mongolian text-to-speech system based on Cyrillic input [30]. Work is also being done on Cyrillic based OCR.

References

[1] http://www.ethnologue.com/show_language.asp?code=mvf
[2] http://www.omniglot.com/writing/mongolian.htm
[3] http://en.wikipedia.org/wiki/Mongolian_language
[4] http://www.unicode.org/charts/PDF/U1800.pdf
[5] Bayarmagnai, N. "Current Status of Information Technology Standardization and its Issues in Mongolia," in Proceedings of 13th AFIT, Myanmar, 1999. Available at http://www.cicc.or.jp/english/hyoujyunka/af13/13-11.html
[6] http://www.paratype.ru/default.asp?page=/library/languag?langCode=51
[7] http://www.salika.co.jp/emgfont.html
[8] http://www.microsoft.com/typography/otfntdev/mongolot/default.htm
[9] http://asuult.net/badaa/docs.php?p=trans_table
[10] http://www.travelphrases.info/gallery/fonts_Mongolian.html
[11] http://www.mnbsd.org/
[12] http://openmn.sourceforge.net/
[13] http://cgm.cs.mcgill.ca/~luc/mongolian.html
[14] http://ourworld.compuserve.com/homepages/PaulGor/fonts_e.htm
[15] http://ourworld.compuserve.com/homepages/PaulGor/screen_e.htm
[16] http://www.amherst.edu/it/software/languages/cyrillic/keyboard98.html
[17] http://asuult.net/badaa/pros.php?p=offline
[18] http://www.microsoft.com/globaldev/reference/keyboards.mspx
[19] http://asuult.net/badaa/docs.php?p=mongol_layout
[20] http://www.unicode.org/versions/Unicode4.0.0/ch12.pdf
[21] http://www-950.ibm.com/software/globalization/icu/demo/locales
[22] http://lists.debian.org/debian-devel/2003/02/msg01965.html
[23] http://www.microsoft.com/globaldev/DrIntl/faqs/lipfaq.mspx#EWB
[24] http://www.microsoft.com/office/editions/prodinfo/language/availability.mspx
[25] http://l10n-status.GNOME.org/GNOME-2.8/mn/index.html
[26] http://l10n.openoffice.org/languages.html#content
[27] http://i18n.KDE.org/teams/index.php?a=i&t=mn
[28] http://laurencio.webz.cz/mongolxel/classical/
[29] http://www.magicnet.mn/modules.php?name=News&file=print&sid=177
[30] http://www.apdip.net/projects/ictrnd/2005/l13-mn/

Nepali

Nepali is an Indo Aryan language spoken by about 17 million people in Nepal, Bhutan and some parts of India, and is the national and official language of Nepal [1]. Figure 2 shows the family tree for Nepali language.

- Indo-European
  - Indo-Iranian
    - Indo-Aryan
      - Northern zone
        - Eastern Pahari
          - NEPALI

Figure 1: Language Family Tree of Nepali [1]

Bhujimol script was earlier used to write Nepali language. Now Nepali is written using Devanagari script [2,3].

Character Set and Encoding

Nepali uses the internationally standardized Devanagari block 0900-097F of Unicode. This standard is gaining popularity in Nepal. Unicode has adopted encoding characteristics from ISCII standard. However, there are still vendor specific encodings being used. Two of the other commonly used encoding schemes, Sabdatara and Anapurna are given in Figure 2 [4].

Another character set encoding standard was developed by the Nepali Fonts Standardization Committee in 1998 [4], but is not frequently used.

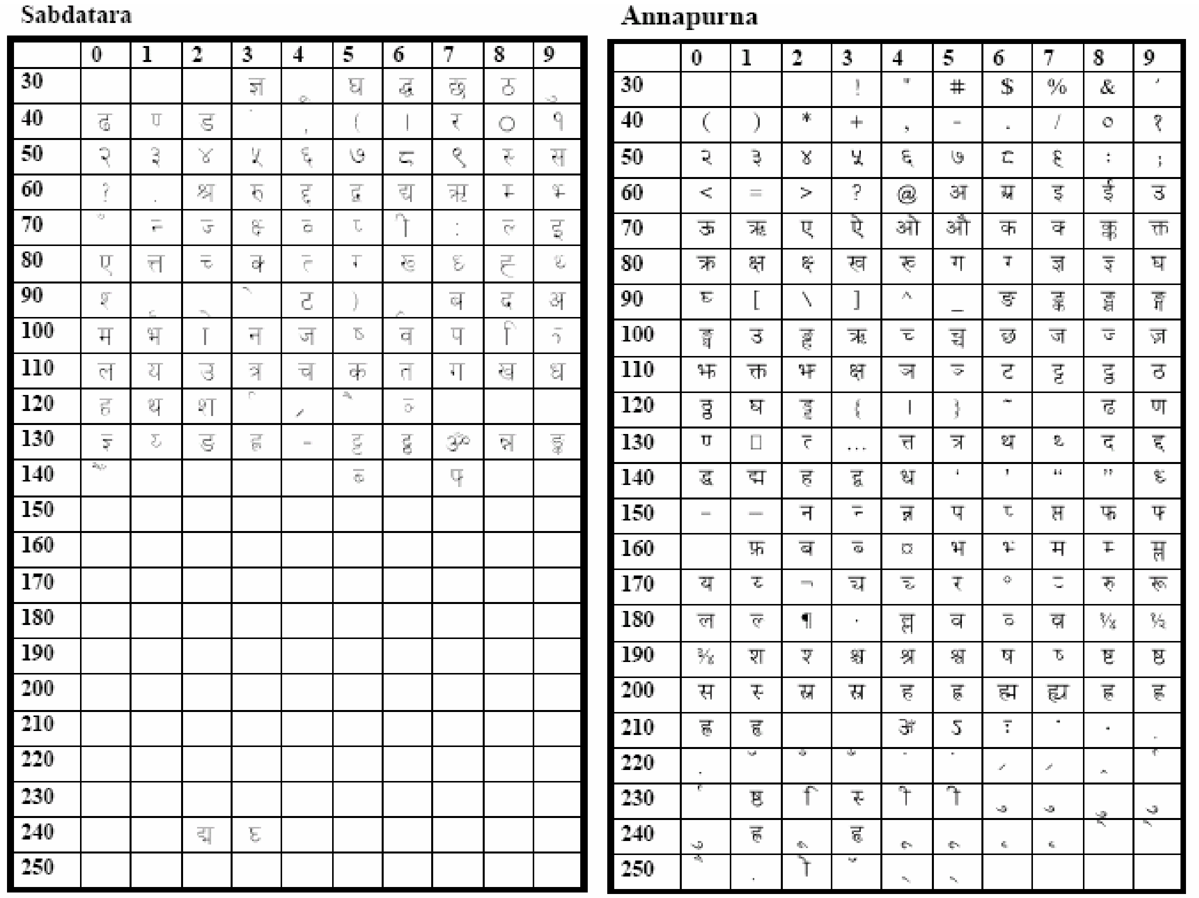
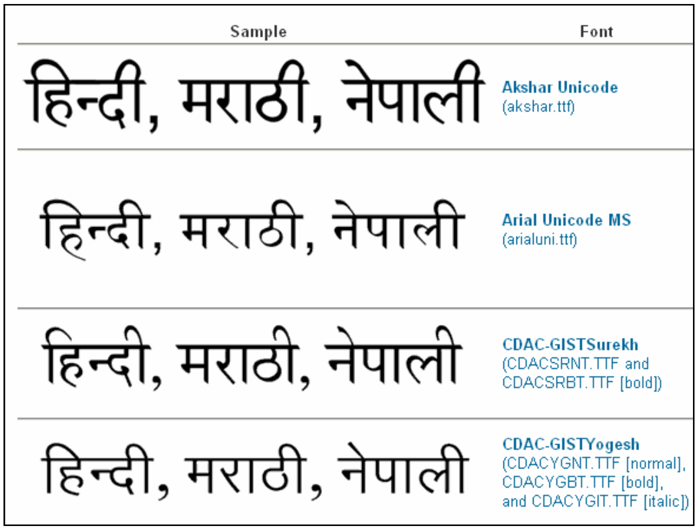

Sabdatara
Annapurna

Figure 2: Popularly Used Nepali Encodings [4]

Fonts and Rendering

Microsoft provides support for Devanagari script in Arial and Mangal fonts, which are shipped with Windows and Office. These fonts can be used for Nepali. Many other Nepali Unicode fonts have also developed by other groups. Some of these fonts are Gauri, Himali, Fontasy Himali, Kanchan, Kantipur, MtEverest, Nepali, Kalimati, and Kanjiwari [5, 6, 7]. Figure 3 shows results of rendering Nepali text using Devanagari fonts on MS platform. Devanagari script is also supported on Linux, through work done by India Linux [8] and Madan Puraskar Pustakalaya (through PAN Localization Nepali component) [7,9].

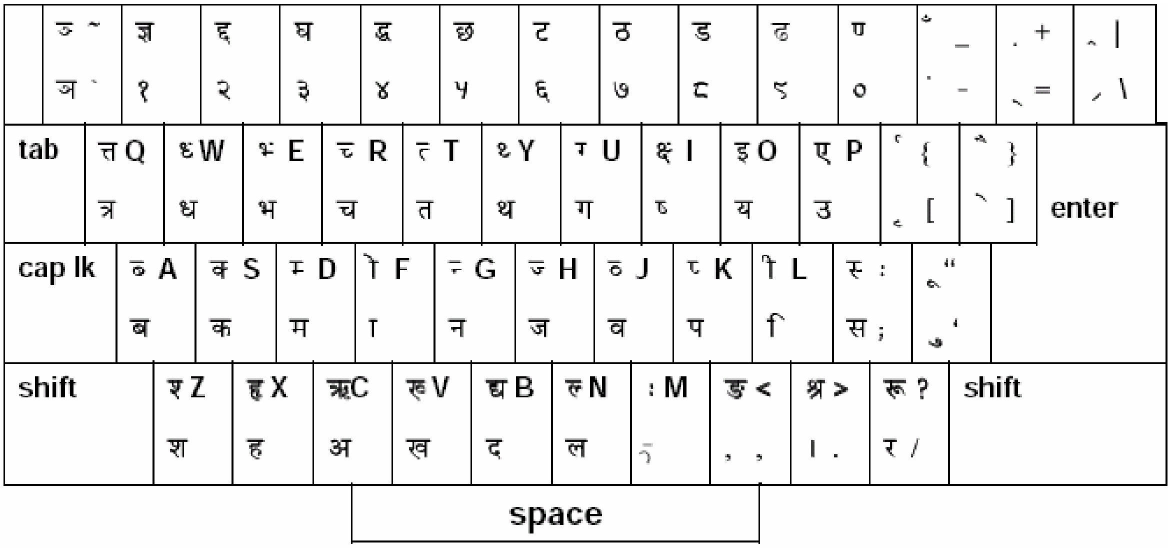


Sample	Font
हिन्दी, मराठी, नेपाली	Akshar Unicode (akshar.ttf)
हिन्दी, मराठी, नेपाली	Arial Unicode MS (arialuni.ttf)
हिन्दी, मराठी, नेपाली	CDAC-GISTSurekh (CDACSRNT.TTF and CDACSRBT.TTF [bold])
हिन्दी, मराठी, नेपाली	CDAC-GISTYogesh (CDACYGNT.TTF [normal], CDACYGBT.TTF [bold], and CDACYGIT.TTF [italic])

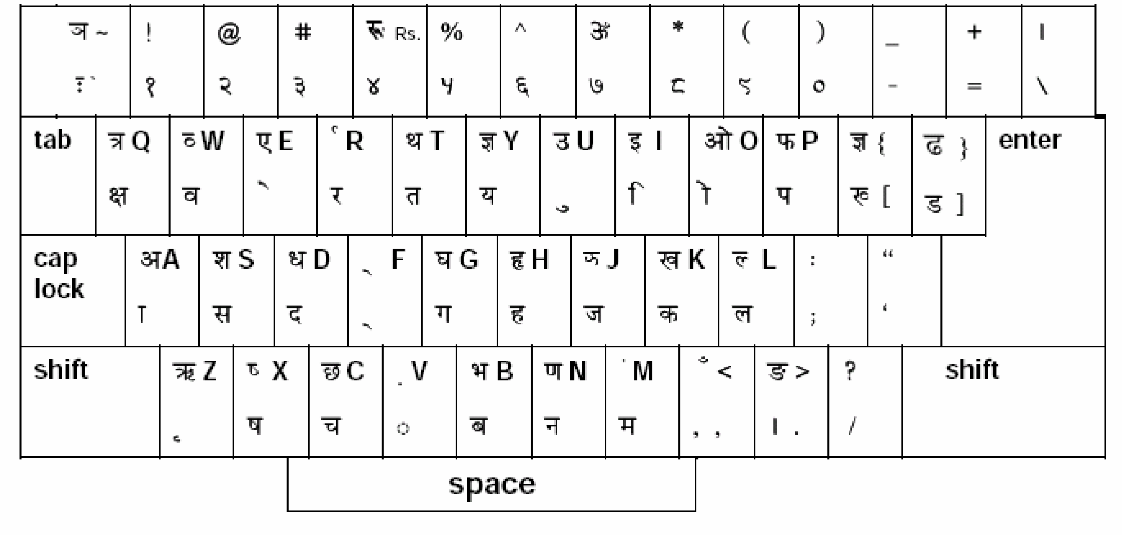
Figure 3: Devanagari Fonts for Nepali [6]

Keyboard

Keyboard layout for Nepali has not been standardized yet. Commonly used Nepali keyboard layouts, Remington and phonetic, are shown in Figure 4 [4].



(a)



(b)

Figure 4: (a) Remington, and (b) Phonetic Keyboard Layouts for Nepali

These keyboard layouts are also available for Linux platform [7], and may also be created on Microsoft platform through its MSKLC tool [10]. Microsoft has also released Nepali LIP, through which a Nepali on-screen Keyboard is provided.

Collation

Two different collation sequences are followed in Nepal. One which treats three conjoined characters a sequence of original characters and other which treats them as new characters. The former is taught in schools and used in phone books, and the latter has been used by Royal Nepal Academy to print its dictionary, Brihat Shabdakosh. Recently, after debate through Nepali language in IT committee, former was standardized nationally [11].

Microsoft Platform

Microsoft does not support Nepali sorting. Krama, a sorting utility developed by Madan Puraskar Pustakalya sorts Nepali strings. Sorting through this utility can be customized through various sort options provides in the utility [7].

Linux Platform

Support for Nepali collation has also been developed for Linux platform.

Locale

Nepali locale (ne_NP) has not been standardized, however some work on it has started [12]. Microsoft has released its Nepali LIP, which has Nepali locale data. Nepali Linux by MPP also has Nepali locale defined.

Interface Terminology Translation

Nepali glossary has been translated and standardized by Nepali language in IT committee. It has also been implemented to develop Nepali Linux through PAN Localization project.

Microsoft Platform

This translation has been used to produce Microsoft LIP for Nepali.

Linux Platform

On the Linux platform 84.85% of GNOME 2.12 has been done by the Nepali team [13] through PAN Localization project [9]. A linux distribution is available including localized Open Office, Nepali GNOME desktop and Mozilla browser.

Status of Advanced Applications

Madan Puraskar Pustakalya is currently developing support for advanced Nepali applications. Currently the MPP team has developed an encoding conversion utility Rupanter that converts non-Unicode Nepali text in to the Nepali Unicode text. Non-Unicode text might be in ad hoc True Type font encodings for Preeti, Kantipur etc. MPP has also developed a prototype version of Nepali spell checker, a Nepali 800 word dictionary, and a Nepali thesaurus of about 800 words. Work is also underway on English-Nepali Machine Translation project [7].

References

[1] http://www.ethnologue.com
[2] http://www.omniglot.com
[3] http://en.wikipedia.org/wiki/Nepali_language
[4] " Nepali Font Standards. " http://www.cicc.or.jp/english/hyoujyunka/mlit3/7-7-2.pdf, 1998.
[5] http://www.nepalhomepage.com/reference/fonts/
[6] http://salrc.uchicago.edu/resources/fonts/devanagarifonts.html
[7] http://www.mpp.org.np/
[8] http://www.IndLinux.org
[9] http://www.PANL10n.net
[10] http://www.mpp.org.np/detail_guide/winxp.htm
[11] Tuladhar, A. "Report on Activities of Standardization of Nepali In Computers. " http://www.unlimit.com/nepali/reports/malaysia.doc
[12] http://www.nepalinux.org/ldf/ne_NP
[13] http://l10n-status.GNOME.org/GNOME-2.12/index.html

Pashto

Pashto is an Indo-Iranian language spoken by about 25 million people in Afghanistan, India, Iran, Pakistan, Tajikistan, the UAE and the UK. There are three main varieties of Pashto: Northern Pashto and Central Pashto (spoken mainly in Pakistan), and Southern Pashto (spoken mainly in Afghanistan). Pashto was made the national language of Afghanistan by royal decree in 1936. Today both Dari and Pashto are official languages in that region. Pashto is also a state language of Pakistan.

Indo-European
Indo-Iranian
Iranian
Eastern
South Eastern
PASHTO

Figure 1: Language Family Tree for Pashto [1]

Pashto first appeared in writing during the 16th century in the form of an account of Sheikh Mali's conquest of Swat. It is written with modified Arabic script [2].

Character Set and Encoding

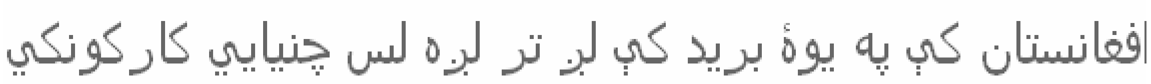
Unicode Arabic script code block 0600-06FF is used for Pashto text encoding. Most of contemporary Pashto is supported. However, older version of script developed by Khushal Khan Khattak is still not supported. A few other ad hoc encodings have also been in use. However, most of the recent work is being done using Unicode.

Fonts and Rendering

Pashto fonts are available in Open Type format (e.g. [3]).

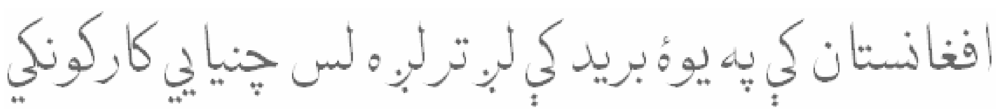
Microsoft Platform

Tahoma and Sans Serif fonts provide complete support for Pashto character set. These are built-in fonts shipped with XP. Some of the Pashto fonts have been developed by Pashto localization vendors. Figure 2 shows Microsoft Tahoma, Pashto Breshna and Nafees Pakistani Naskh on Microsoft platform.



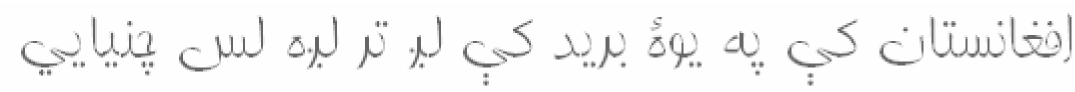
افغانستان کې په يوۀ بريد کې لږ تر لږه لس چنيايي کارکونکي

(a)



افغانستان کې په يوۀ بريد کې لږ تر لږه لس چنيايي کارکونکي

(b)



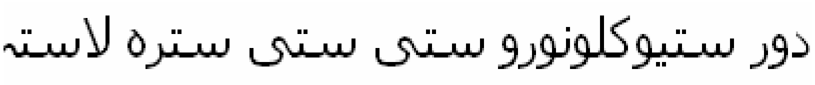
افغانستان کې په یوۀ برید کې لږ تر لږه لس چنیایي

(c)

Figure 2: (a) Tahoma, (b) Nafees Pakistani Naskh and (c) Pashto Breshna [8] on MS Office 2003

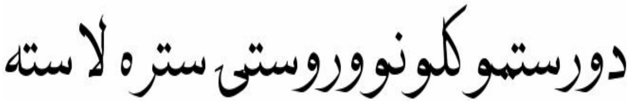
Linux Platform

Linux supports rendering of four shaped fonts for Arabic script. Pashto written through these fonts can be adequately rendered on Linux platform. The following figures show rendering Pashto text in various applications running in Linux platform. However there are still some rendering issues with some characters. Open Type Pashto fonts cannot be used on KDE. The platform only supports True Type Pashto fonts.



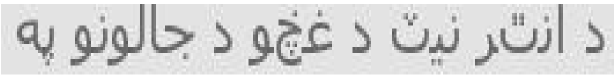
دور ستیوکلونورو ستی ستره لاسته

(a)



دورستموکلونوورروستی ستره لاسته

(b)



د انټرنیټ د غځو د جالونو په

(c)

Figure 3: (a) Four-Shaped Pashto Font on Open Office, (b) Nafees Naskh in GNOME, and (c) Pashto TTF Naskh on KDE

Keyboard

No standard keyboard layout exists for Pashto language. There are keyboard layouts available based on Arabic or Persian layouts or phonetic layouts based on English QWERTY. These keyboard layouts have been developed by local vendors (e.g. [5, 6]). A recent study has been conducted by UNDP which recommends specifications [4]. Work is currently in progress to standardize the keyboard (and other localization related issues) through Afghanistan Localization Program [7]. The keyboard layouts in following figures are recommended by UNDP in the draft locale requirements for Afghanistan, and those developed by Khapla Pashto Software and Liwal Software.

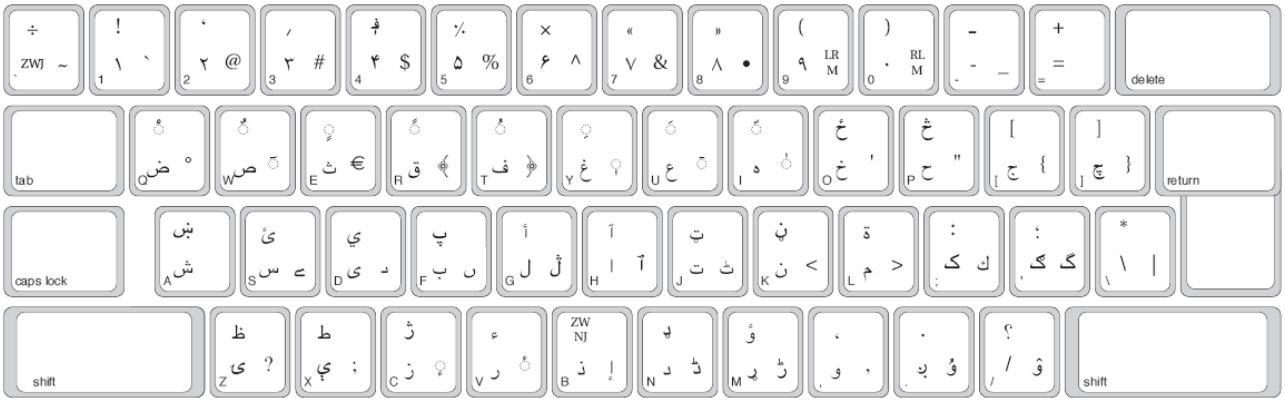


Figure 4: UNDP Recommended Pashto Keyboard (see [4] for clearer Image)

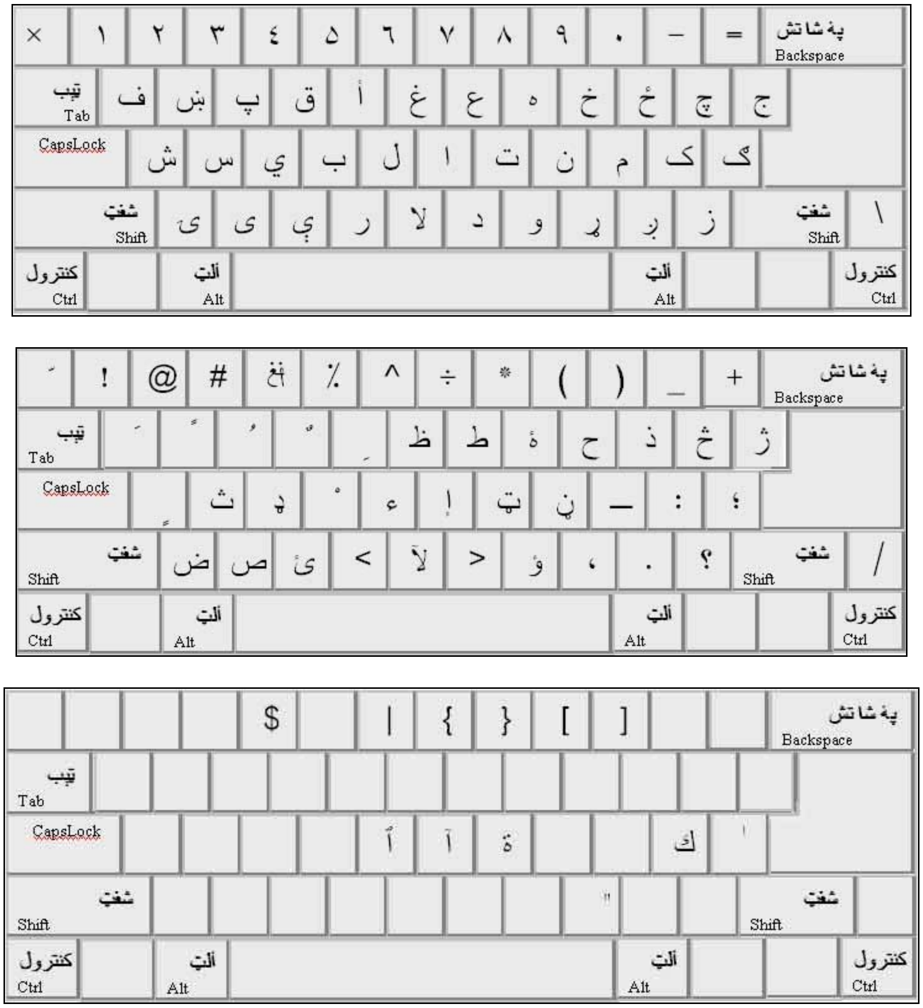


Figure 5: Pashto Keyboard by Khpala Software [5]

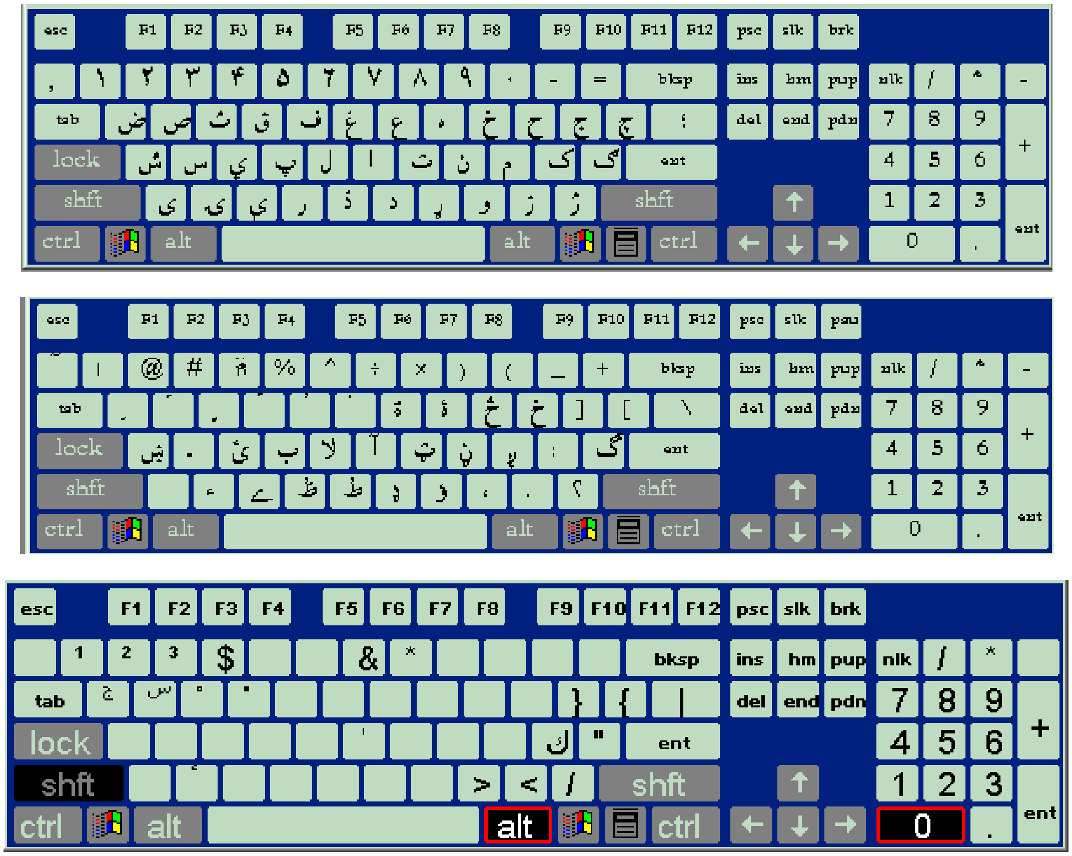


Figure 6: Pashto Keyboard by Liwal Software [6]

Microsoft or Linux platforms do not provide built-in keyboard layout for Pashto.

Collation

UNDP has given recommendations on collation order of Pashto text [4]. However, it has not been standardized. Work is currently in progress through PAN Localization project to standardize Pashto sequence for Afghanistan [7].

Microsoft and Linux platforms do not support Pashto collation.

Locale

A draft for Pashto locale requirements for Afghanistan has been developed by UNDP [8]. Incomplete versions of locale are available through IBM ICU and CLDR 1.3. The locale is also being standardized by Afghanistan (pa_AF). No such work has been initiated in Pakistan (pa_PK).

There is no locale support in Microsoft or Linux platforms. However, work is in progress to develop Microsoft Language Interface Pack for Pashto in Afghanistan [7].

Interface Terminology Translation

No standard terminology exists for Pashto. Work is in progress through Afghanistan Localization program to develop this terminology [7].

Status of Advanced Applications

Few language processing resources are available for Pashto. Some initial work has been done on Pashto lexica (e.g. [9]). Work has also been done on Arabic script based OCR by vendors, which can potentially be extended for Pashto. Pashto TTS and ASR systems are also being developed. Pashto-English machine translation systems are available through private vendors, e.g. [10, 11]. Most of this work is being done outside Afghanistan and Pakistan.

References

[1] www.ethnologue.com
[2] www.omniglot.com / www.ethnologue.com
[3] http://www.crulp.org/nafeesPakistaniNaskh.html
[4] http://www.evertype.com/standards/af/af-locales.pdf
[5] http://www.khpalapashtu.com/sitee/pashtusw/paskeyb.htm
[6] http://www.liwal.com/windows/pashto/keyboard.htm
[7] http://www.moc.gov.af/Projects/localization.asp
[8] http://www.evertype.com/standards/af/
[9] http://www.yorku.ca/twainweb/troberts/pashto/lexicon.html
[10] http://linguistical.com/shop/index.cgi?ID=143217967&PID=IT492&code=13
[11] http://www.aramedia.com/

Sinhala

Sinhala (also called Sinhalese) is an Indo-Aryan language spoken in Sri Lanka by about 13 million people [1]. A considerable number of speakers reside in Singapore, Thailand, Canada and the United Arab Emirates as well [1]. Sinhala is the national language of Sri Lanka. Figure 1 shows the family tree of Sinhala.

```
Indo-European
        Indo-Iranian
                Indo-Aryan
                        Sinhala-Maldivian
                                SINHALA [SNH]
```

Figure 1: Language Family Tree for Sinhala [1]

Sinhala is written using the Sinhala script. The script is believed to be an off-shoot of Brahmi script [2]. The structure of Sinhala script is similar to that of Sanskrit while it has imported few Tamil characters as well [3].

Character Set and Encoding

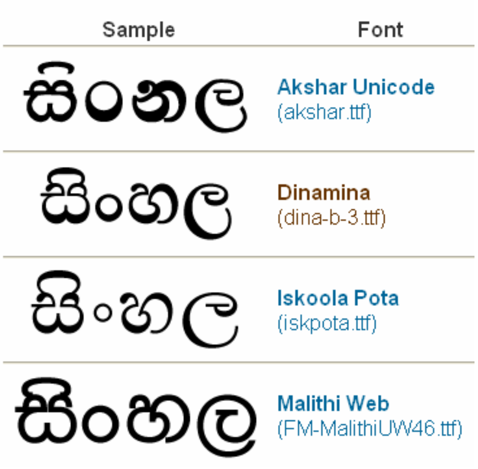
SLS1134, developed in 1997 by the Sri Lanka Standards Institute (SLSI), has been the national standard for Sinhala. Unicode code block from 0D80-0DFF is the international standard encoding scheme for Sinhala character set [10], also being adopted at national level but is still not widely used. Prior to the inclusion of Sinhala code page in Unicode in 2003, various character encodings had been developed by local vendors. As reported by University of Colombo School of Computing (UCSC), Sri Lanka, these are still popularly used encoding schemes, and include Thibus, Kaputa, AQNCL, Microimage, Phonetic and SLS1134.

Fonts and Rendering

True Type and Open Type fonts are available for Sinhala. These include free fonts and those developed by various vendors. However, they use different encoding schemes. These include Kaputadotcom, Mi_Dasun Tall (which have Open Type support as well), Dil-Silumina, FM-Basuru, FS-Dilu, Amallee Palin, thibus and Sarasavi. Most of these fonts use phonetic encoding, and Sarasavi Sinhal font uses the national standard SLS1134 encoding. Unicode based fonts are also available [4, 5].

Microsoft Platform

With the inclusion of Sinhala characters in Unicode, Sinhala text can now be rendered on Microsoft platform. Iskola Pota is packaged with the beta version of Sinhala kit for Windows XP which is available for download (through ICT Agency of Sri Lanka) [4, 6]. Outputs of a few fonts for Microsoft platform are given in Figure 2.

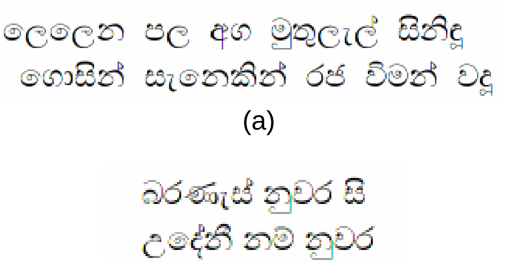


Sample	Font
සිංහල	Akshar Unicode (akshar.ttf)
සිංහල	Dinamina (dina-b-3.ttf)
සිංහල	Iskoola Pota (iskpota.ttf)
සිංහල	Malithi Web (FM-MalithiUW46.ttf)

Figure 2: Sinhala Fonts on Microsoft Platform [5]

Linux Platform

Sinhala GNU/Linux 0.1 contains Sinhala language patch for Pango having Sinhala Open Type font support. The figure below shows Sinhala fonts rendering in G-edit and Konquorer.



ලෙලෙන පල අග මුතුලැල් සිනිඳු
ගොසින් සැනෙකින් රජ විමන් වදූ

(a)

බරණැස් නුවර සි
උදේනි නම නුවර

(b)

Figure 3: Font Rendering in (a) GNOME and (b) KDE [7]

Keyboard

The standard keyboard layout is based on Wijesekara layout. It was first standardized in 1996 and minor modifications were made in 2001. In 2004 it was further modified and now it is used as the national standard for Sinhala keyboard layout (SLS1134:2004). The following figure shows this keyboard layout (see [9] for a more comprehensive overview of this and other standardization for Sinhala and [11] for the evolution of Sinhala keyboard). Another popular keyboard is based on phonetic mapping of Sinahala characters on QWERTY keyboard of English.

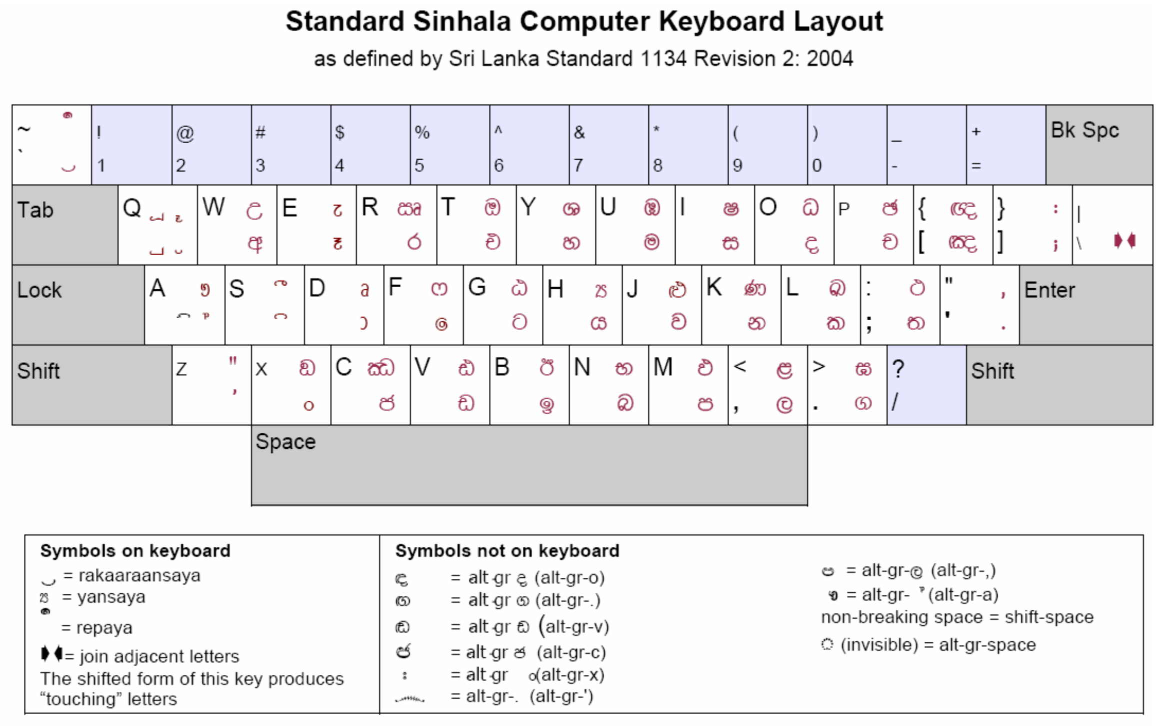


Figure 4: Standard Keyboard Layout [8]

Microsoft Platform

Microsoft does not provide support for Sinhala text input. However, a built-in Sinhala keyboard layout is installed with the beta version of the Sinhala Language Kit [4]. Once the kit is installed, the Sinhala keyboard is available.

Linux Platform

A phonetic-based keyboard layout is shipped with Sinhala Linux distribution.

Collation

To date, no standard collation order exists for Sinhala character set. However, currently a standard is being developed at University of Colombo School of Computing (UCSC) for Information Communication Technologies Agency (ICTA) which will be standardized through SLSI.

As no standard collation exists, there is no collation implemented on Microsoft or Linux platforms.

Locale

Sinhala locale has neither been defined in the IBM ICU nor in the CLDR 1.3 release. Currently Sinhala locale definition for Microsoft is being finalized by ICTA [6]. Locale definition for Sinhala and Sinhala interface terminology translation is also being done in parallel by a Lanka Linux User Group (LkLUG) [7].

Microsoft does not provide a locale for Sinhala. On the Linux platform the Lanka Linux User Group (LkLUG) has included limited support for a Sinhala locale in Glibc in the Sinhala Linux distribution GNU/Linux 0.1 [7].

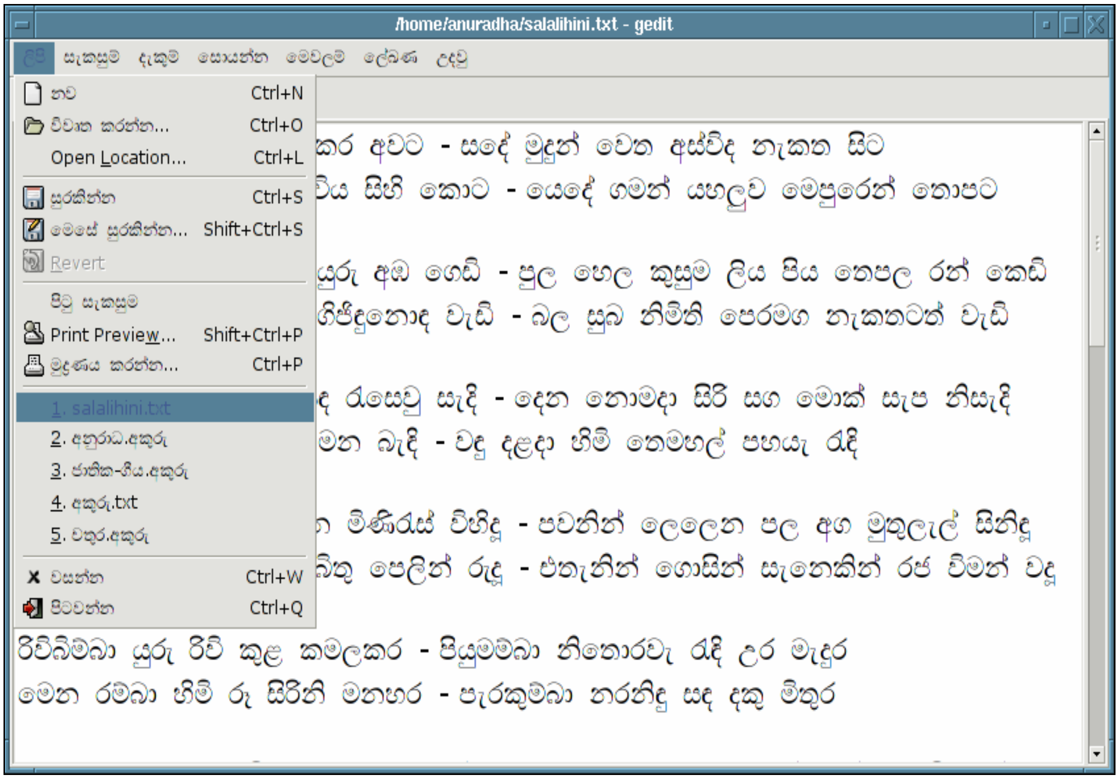
Interface Terminology Translation

Microsoft Platform

Microsoft developed a Sinhala version [12] at an accelerated pace to help support Tsunami relief in 2005 in collaboration with ICTA, Sri Lanka. A Sinhala LIP for Windows XP and a Sinhala kit for Microsoft Office 2003 is available for download [4]. However, more mature version of interface translation is still under development. ICTA has completed initial phase in translating interface terminology. The LkLUG has done more work on this than is available on Microsoft platform at present. They had a small project to translate the 200 most common UI terms. There is some on-going work on terminology translation (of some 2500 words) at UCSC which will be incorporated in future. This work will also be adopted by ICTA, as reported by University of Colombo School of Computing.

Linux Platform

A complete Sinhala Linux distribution GNU/Linux 0.1 has been developed and released by LkLUG. This distribution includes Sinhala fonts, phonetic keyboard, locale and a language patch for Pango. It is partially localized, as shown in Figure 5.



කර අවට - සදේ මුදුන් වෙත අස්විද නැකත සිට
විය සිහි කොට - යෙදේ ගමන් යහලුව මෙපුරෙන් තොපට

යුරු අඹ ගෙඩි - පුල හෙල කුසුම ලිය පිය තෙපල රන් කෙඩි
ගිජිඳුනොද වැඩි - බල සුබ නිමිති පෙරමග නැකතටත් වැඩි

ද රැසෙවු සැදි - දෙන නොමදා සිරි සග මොක් සැප නිසැදි
මන බැදි - වඳු දළදා හිමි තෙමහල් පහයැ රැදි

ා මිණිරැස් විහිදූ - පවනින් ලෙලෙන පල අග මුතුලැල් සිනිඳු
බිතු පෙලින් රුදූ - එතැනින් ගොසින් සැනෙකින් රජ විමන් වදූ

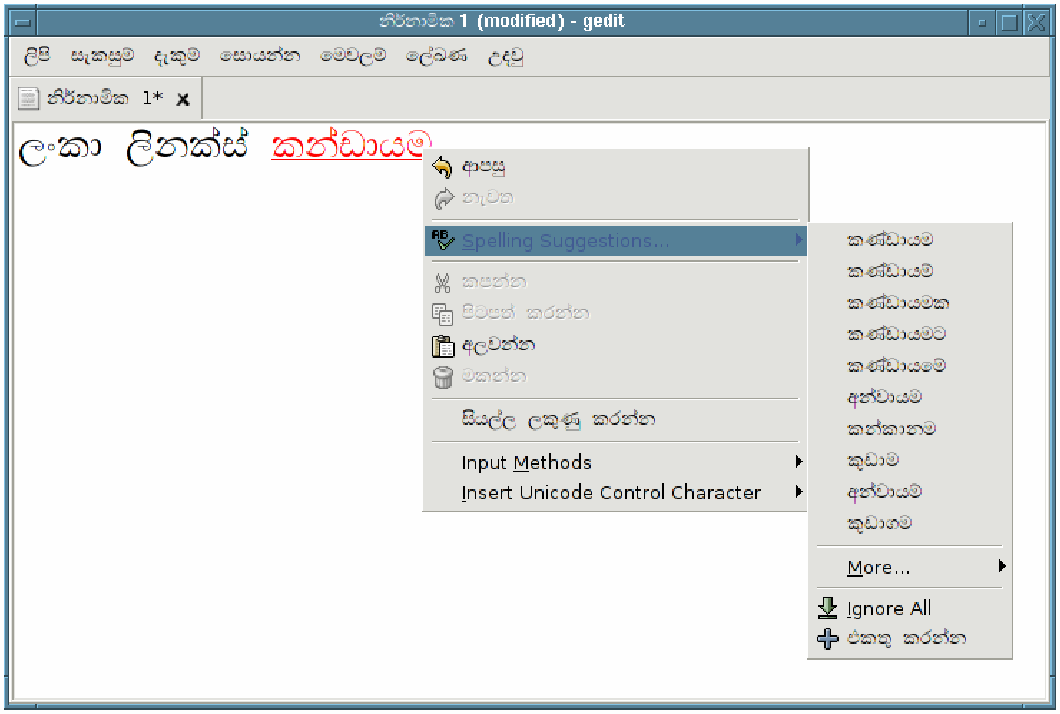
රිවිබිම්බා යුරු රිවි කුළ කමලකර - පියුමම්බා නිතොරවැ රැදි උර මැදුර
මෙන රම්බා හිමි රු සිරිනි මනහර - පැරකුම්බා නරනිඳු සඳ දකු මිතුර

Figure 5: GUI for G-Edit [7]

Translation teams for both KDE and GNOME have been registered on their official websites. Some work on the glossary translation of open source applications is currently underway. Language teams for Sinhala have registered on the official websites of KDE and Mozilla. Glossary translation of GNOME and Open Office has not been initiated yet.

Status of Advanced Applications

As reported by the University of Colombo School Of Computing, work is in progress to develop local language applications in Sinhala. The language processing teams have developed experimental versions of Sinhala Lexicon, corpus and spell checker. Spell checker has also been developed by LkLUG, incorporated in Linux distribution, as shown in Figure 6.



ලංකා ලිනක්ස් කන්ඩායම

Figure 6: GUI for G-Edit [13]

Through PAN Localization project, UCSC has developed and already released alpha versions of Sinhala OCR and text-to-speech system [14]. The work is underway for maturing these systems.

References

[1] http://www.ethnologue.com/show_language.asp?code=sin
[2] http://www.omniglot.com/writing/sinhala.htm
[3] http://acharya.iitm.ac.in/cgi-bin/script_disp.pl?sinhala
[4] http://www.fonts.lk/index.html
[5] http://salrc.uchicago.edu/resources/fonts/sinhalafonts.html
[6] http://www.icta.lk/
[7] http://sinhala.linux.lk/
[8] http://www.fonts.lk/doc/sin-kbd-layout5.pdf
[9] Dias, G. and Goonitilleke, A., “Development of Standards for Sinhala Computing.” 1st Regional Conference on ICT and E-Paradigms, Srilanka. http://www.fonts.lk/doc/sinhala%20standards.pdf. 2004.
[10] Dias, G. “Representation of Sinhala in Unicode.” http://www.fonts.lk/doc/Representation%20of%20Sinhala%20in%20Unicode.pdf
[11] http://www.fonts.lk/doc/workshop-sinhalakeyboards-anura.ppt
[12] http://www.dailynews.lk/2005/01/29/bus02.html
[13] http://sinhala.linux.lk/pub/0.2/screenshots/spell.png
[14] http://www.PANL10n.net

Telugu

Telugu is a Dravidian language spoken in southern Indian states, and is the official language of Andhra Pradesh. It is considered to be the the second most widely spoken language in India after Hindi. There are 75 million first language and about 5 million second language speakers of Telugu [1]. Figure 1 shows the language tree for Telugu.

Dravidian
South-Central
TELUGU

Figure 1: Language Family Tree of Telugu [1]

Telugu script derives from Brahmi script, and is similar to Kannada script. The earliest known inscriptions in the Telugu language date back from the 6th century AD. Telugu has been written in an old style. However, the writing system was modernized in second half of 20th century to agree more with the spoken language [2].

Character Set and Encoding

ISCII includes encoding for Telugu characters, with code page identifier 57006. Telugu script has also been incorporated in Unicode from 0C00-0C7F [3].

There are also other ad hoc encodings which are in use. Encoding converters between ISCII and these Telugu ad hoc encodings have also been developed [17].

Fonts and Rendering

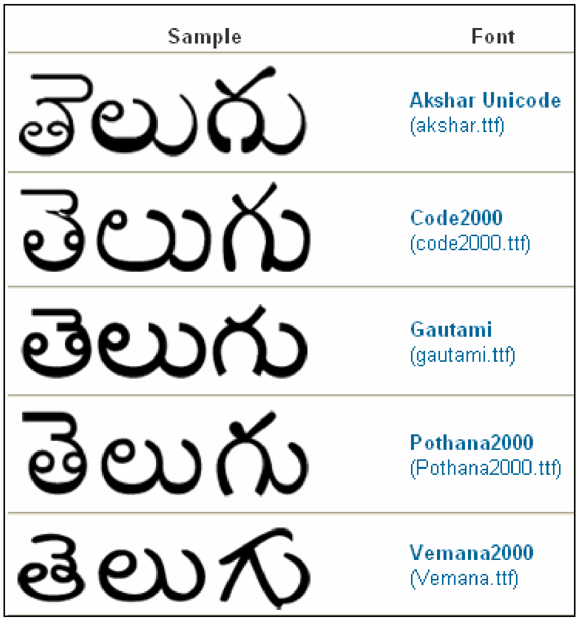
Many Open Type fonts are available, which can be used for rendering Telugu script, e.g. Akshar Unicode, Code 2000, Pothana 2000 and Vemana 2000 [4].

Microsoft Platform

Microsoft supports Telugu rendering, and provides font support for this script, e.g. in Arial Unicode MS font. Figure 2 shows results of rendering some of these fonts on Microsoft platform [4]. Microsoft also provides a guide to develop Open Type fonts for Telugu [8].

Linux Platform

Telugu Open Type fonts are not rendered on Linux platform in regular distributions. However, rendering support has been developed in Pango for GNOME and Firefox, and Qt engine for KDE. Upgraded versions are available at [5].

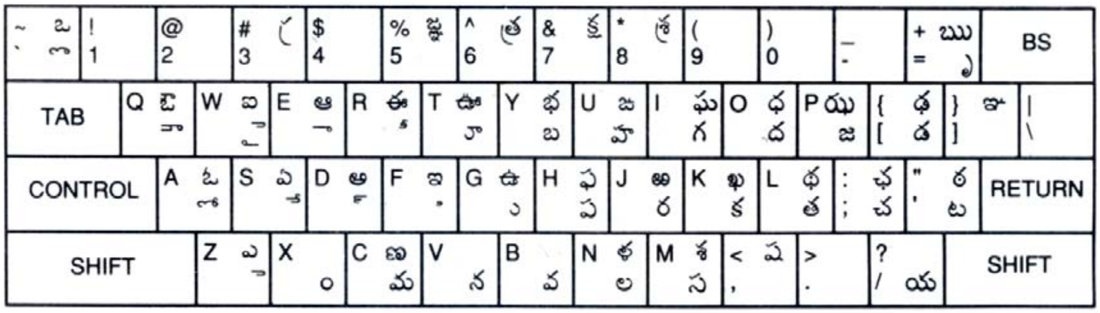


Sample	Font
తెలుగు	Akshar Unicode (akshar.ttf)
తెలుగు	Code2000 (code2000.ttf)
తెలుగు	Gautami (gautami.ttf)
తెలుగు	Pothana2000 (Pothana2000.ttf)
తెలుగు	Vemana2000 (Vemana.ttf)

Figure 2: Telugu Unicode fonts [4]

Keyboard

Three different keyboard layouts are used by Telugu, Inscript, RTS and WX, of which first two are very popular. RTS and WX are phonetic based [5]. Inscript is standardized through Government of India [6, 7] and is shown in Figure 3.



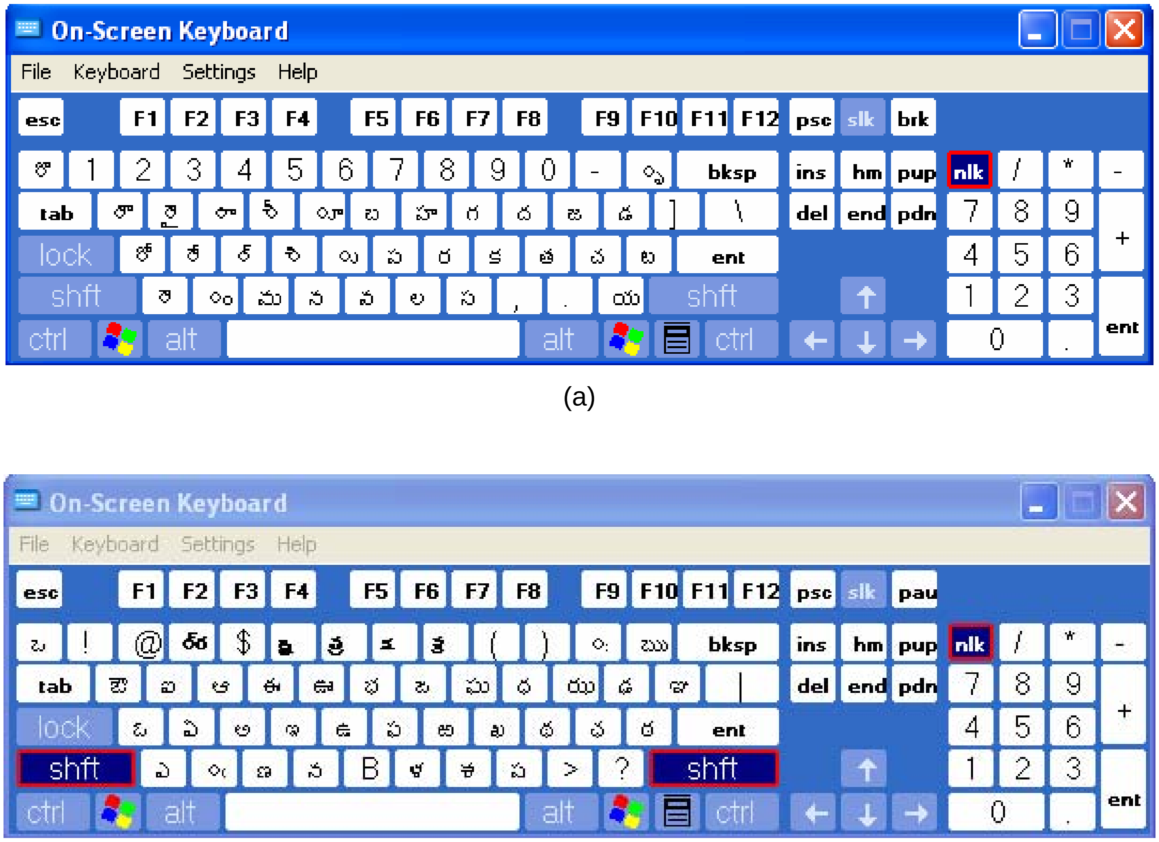
INSCRIPT OVERLAY FOR TELUGU

Notes : – The Macro-keys in the top-row generate :
(Rakar) ్ర = ్ర
జ్ఞ = జ్ఞ   త్ర = త్ర   క్ష = క్ష   శ్ర = శ్ర

Figure 3: Inscript Telugu Keyboard Layout [7]

Microsoft Platform

Microsoft provides a Telugu keyboard, enabling support in all Microsoft applications. Figure 4 shows the keyboard layout provided by Microsoft.


On-Screen Keyboard
File Keyboard Settings Help
esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 psc slk brk
1 2 3 4 5 6 7 8 9 0 - bksp ins hm pup nlk / * -
tab] \ del end pdn 7 8 9 +
lock ent 4 5 6
shft , . shft 1 2 3
ctrl alt alt ctrl 0 . ent
On-Screen Keyboard
File Keyboard Settings Help
esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 psc slk pau
! @ $ ( ) bksp ins hm pup nlk / * -
tab | del end pdn 7 8 9 +
lock ent 4 5 6
shft B > ? shft 1 2 3
ctrl alt alt ctrl 0 . ent

(a)

(b)

Figure 4: Microsoft On-Screen Keyboard for Telugu: (a) Normal, and (b) Shift State

Linux Platform

All keyboard layouts mentioned above (RTS, Inscript and WX) are available on Linux platform. The support is not built-in and has to be installed [5].

Collation

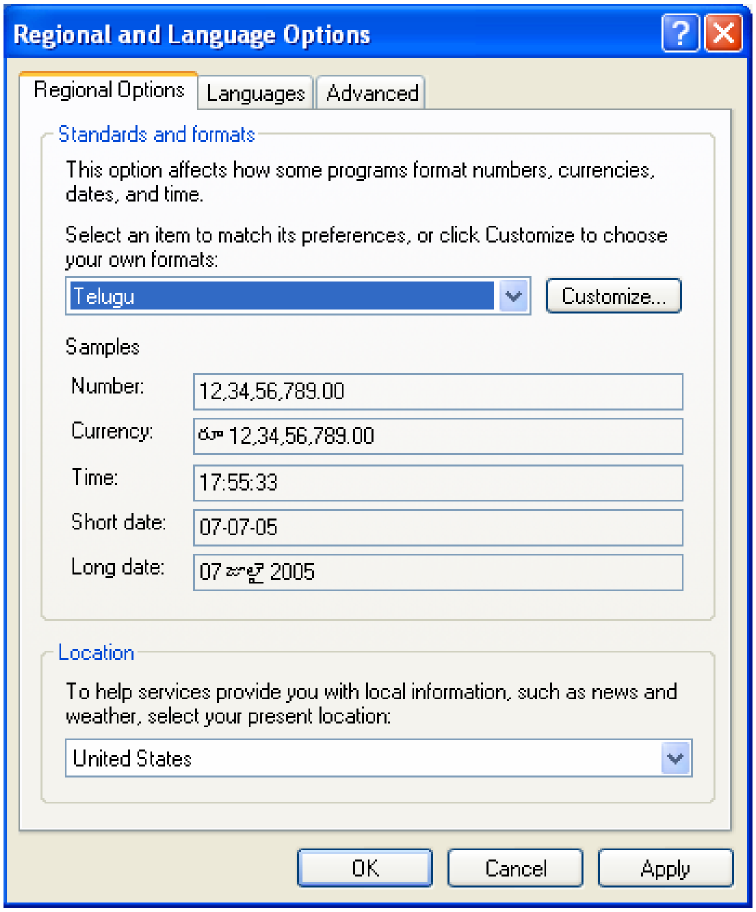
Telugu locale is defined but not standardized. Basic support is available on Microsoft and Linux platforms e.g. see [10].

Locale

Telugu locale (te_IN) is partially defined and is standardized through posting at IBM ICU and CLDR 1. Further information is available at [5].

Microsoft Platform

Microsoft provides a complete LIP for Telugu. Figure 5 shows a snapshot of the locale settings for Telugu.



Standards and formats

This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Telugu

Samples

Number: 12,34,56,789.00

Currency: రూ 12,34,56,789.00

Time: 17:55:33

Short date: 07-07-05

Long date: 07 జూలై 2005

Location

To help services provide you with local information, such as news and weather, select your present location:

United States

Figure 5: Telugu Locale on Windows XP

Linux Platform

Support for Telugu locale is available in Glib C. The Telugu locale for India has also been defined officially in Red Hat Linux [9]. It is also supported on other open source platforms [10].

Interface Terminology Translation

Interface terminology is not standardized, but is available on many platforms.

Microsoft Platform

Microsoft supports Telugu LIP, which contains Telugu interface for its applications [11].

Linux Platform

A “Native Language Project” has been initiated for localizing Open Office version 2.0. Its team has just started therefore no output from this group has been committed [12]. The India Linux project also includes a language team for localizing Linux in Telugu [14]. However no output of the team has been published. GNOME translation has started for Telugu and about 85% translations of glossary have been accomplished [15]. This team is working independently and has not registered at the www.GNOME.org for translating GNOME 2.8 GUI messages to local languages [15]. A Telugu localization team has also registered on the Mozilla Localization Project (MLP). This team is working on localizing Mozilla version 1.2.1. No outputs of this project are posted [13].

Status of Advanced Applications

Telugu support has been added into the Unicode text editor Yudit [10]. Work has also been done for Telugu Lexicon, open source spell checker Aspell [10] and prototypical work on optical character recognition system. Significant work on Telugu text-to-speech, lexicon and Telugu-Hindi machine translation is also under progress, with working systems already released [16, 17].

References

[1] http://www.ethnologue.com/show_language.asp?code=tel
[2] http://www.omniglot.com/writing/telugu.htm
[3]http://www.unicode.org/charts/PDF/U0C00.pdf
[4] http://salrc.uchicago.edu/resources/fonts/telugufonts.html
[5] http://telugu.sarovar.org/wiki/
[6] http://tdil.mit.gov.in/keyoverlay.htm
[7] http://tdil.mit.gov.in/isciichart.pdf
[8] http://www.microsoft.com/typography/OpenType%20Dev/telugu/intro.mspx
[9] http://www.indlinux.org/downloads/locale/Locales/te_IN
[10] http://telugu.sarovar.org/wiki/index.php/Status
[11] http://www.bhashaindia.com/downloadsV2/Category.aspx?ID=8
[12] http://te.openoffice.org
[13] http://www.mozilla.org/projects/l10n/mlp_status.html
[14] http://indlinux.org/lang/
[15] http://l10n-status.GNOME.org/GNOME-2.8/index.html
[16] http://ltrc.iiit.net/showfile.php?filename=research/
[17] http://ltrc.iiit.net/showfile.php?filename=downloads/

Thai

Thai belongs to the Tai language family of the Kadai or Kam-Tai family, latter arguably regarded, along with Austronesian, as a branch of Austro-Tai [1]. About 20 million people speak Thai in Thailand, where it also the official language of the country. Figure 1 shows the language tree for Thai [1].

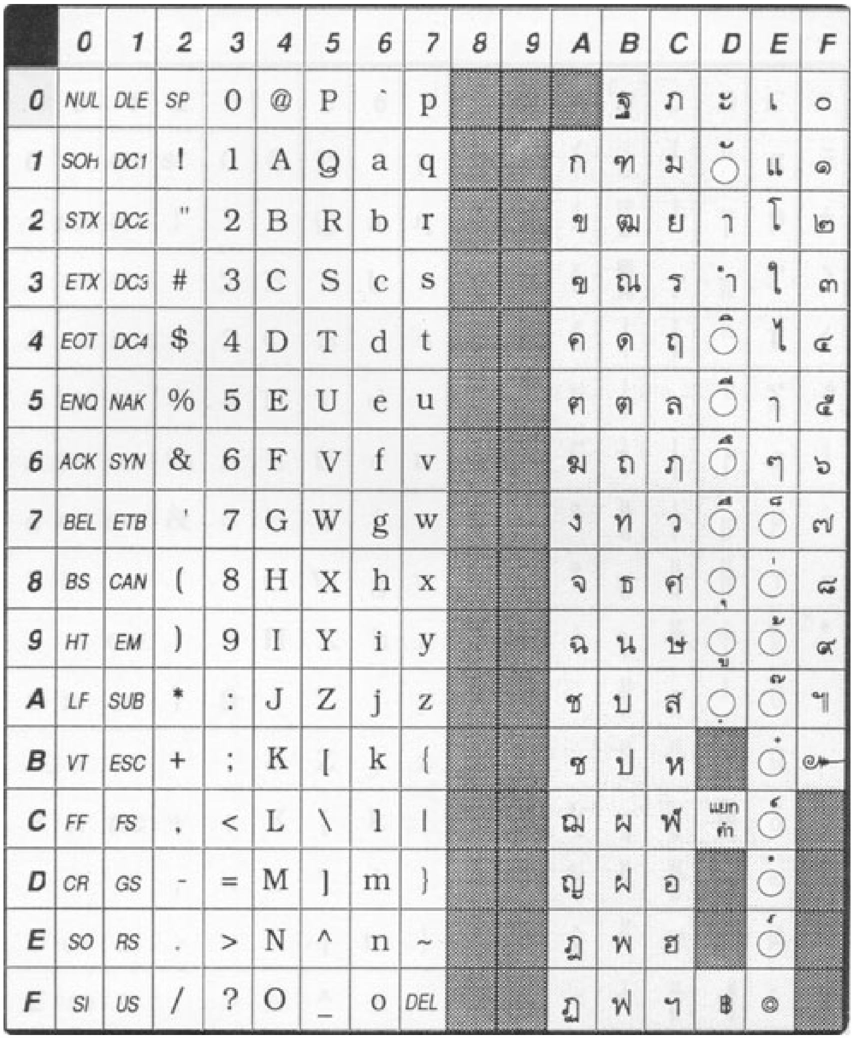
- Tai-Kadai
  - Kam-Tai
    - Be-Tai
      - Tai-Sek
        - Tai
          - Southwestern
            - East Central
              - Chiang Saeng
                - THAI

Figure 1: Language Family Tree for Thai

Thai is written using the Thai script, which was derived from Brahmi scripts around 12th century AD. Thai script is known to have been influenced by Khmer script [2].

Character Set and Encoding

Thailand Industrial Standards Institute (TISI) develops standards in response to the government policy [3]. TISI has approved many standards for local language computing, e.g. for characters set, keyboard, encoding, etc. Before Unicode more than a score different Thai code pages, defined by local vendors, were in use. This led to the lack of interoperability between applications. To counter the problem TISI introduced a national encoding standard TIS 620-2529/1986, later upgraded in TIS 620-2533/1990 [4, 5]. TIS 620-2533/1990 was used as a basis for IBM code page 874 (cp-874), Microsoft code page 874 (windows-874) and Apple Thai (MacThai) [4]. Eventually, Unicode character set was introduced, which is now widely used. The encoding for Thai ranges from 0E00-0E7F in Unicode. Figure 2 shows Thai code chart for TIS 620-2533 [6]. A comparison of TIS 620 and Unicode is presented in [7]. In 1999, the international standard ISO/IEC 8859-11 Latin/Thai characters was also reactivated but not accepted [8]. Also see [8] for a comprehensive coverage of historical development of standards and [14] for a complete list of standards. Microsoft has been using another encoding standard for Thai [31].



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	DLE	SP	0	@	P	`	p				ฐ	ภ	ะ	เ	๐
1	SOH	DC1	!	1	A	Q	a	q			ก	ฑ	ม	◌ั	แ	๑
2	STX	DC2	"	2	B	R	b	r			ข	ฒ	ย	า	โ	๒
3	ETX	DC3	#	3	C	S	c	s			ฃ	ณ	ร	◌ำ	ใ	๓
4	EOT	DC4	$	4	D	T	d	t			ค	ด	ฤ	◌ิ	ไ	๔
5	ENQ	NAK	%	5	E	U	e	u			ฅ	ต	ล	◌ี	ๅ	๕
6	ACK	SYN	&	6	F	V	f	v			ฆ	ถ	ฦ	◌ึ	ๆ	๖
7	BEL	ETB	'	7	G	W	g	w			ง	ท	ว	◌ื	◌็	๗
8	BS	CAN	(	8	H	X	h	x			จ	ธ	ศ	◌ุ	◌่	๘
9	HT	EM	)	9	I	Y	i	y			ฉ	น	ษ	◌ู	◌้	๙
A	LF	SUB	*	:	J	Z	j	z			ช	บ	ส	◌ฺ	◌๊	๚
B	VT	ESC	+	;	K	[	k	{			ซ	ป	ห		◌๋	๛
C	FF	FS	,	<	L	\	l	|			ฌ	ผ	ฬ	แยก คำ	◌์	
D	CR	GS	-	=	M	]	m	}			ญ	ฝ	อ		◌ํ	
E	SO	RS	.	>	N	^	n	~			ฎ	พ	ฮ		◌๎	
F	SI	US	/	?	O	_	o	DEL			ฏ	ฟ	ฯ	฿	๏	

Figure 2: Thai Character Code Chart TIS 620-2533 [7]

Fonts and Rendering

Microsoft Platform

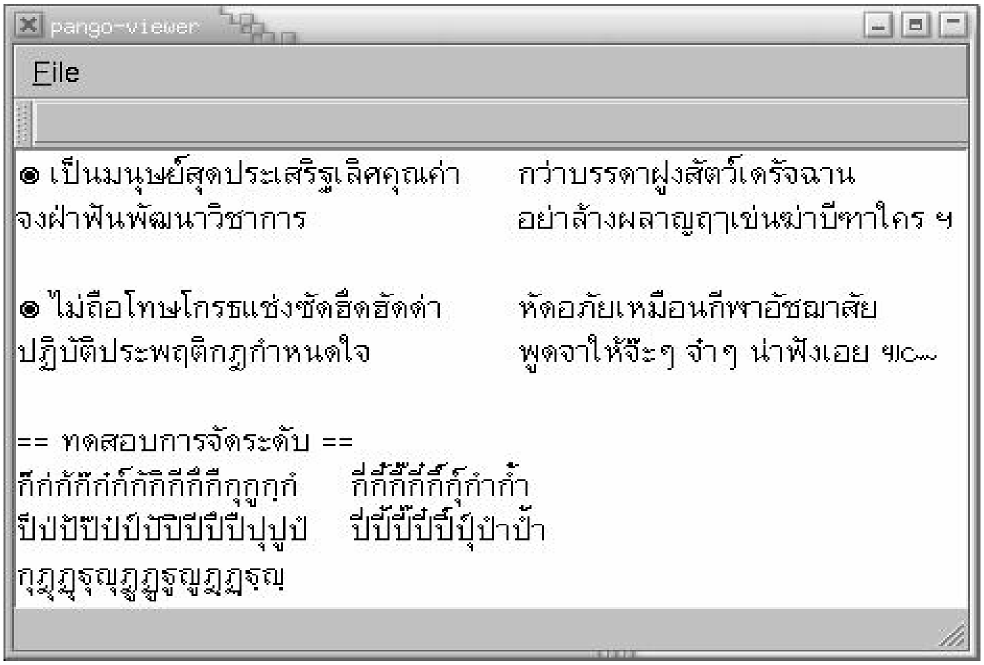
Microsoft provides rendering support for Thai, ships Thai fonts with its Thai version of Windows and Office [11], and also provides guidelines for Thai font development [12]. Figure 3 shows rendering results of some Thai fonts [9]. Many more Thai fonts are available through other organizations (e.g. [10]).

สวัสดี	Angsana
สวัสดี	Courier Mono Thai
สวัสดี	Courier Proportional Thai

Figure 3: Thai Fonts on Microsoft Platform [9]

Linux Platform

Thai rendering is also supported on Linux and open source platforms, e.g. in Pango, the rendering engine of GNOME, as shown in Figure 4.



๏ เป็นมนุษย์สุดประเสริฐเลิศคุณค่า กว่าบรรดาฝูงสัตว์เดรัจฉาน
จงฝ่าฟันพัฒนาวิชาการ อย่าล้างผลาญฤๅเข่นฆ่าบีฑาใคร ฯ

๏ ไม่ถือโทษโกรธแช่งซัดฮึดฮัดด่า หัดอภัยเหมือนกีฬาอัชฌาสัย
ปฏิบัติประพฤติกฎกำหนดใจ พูดจาให้จ๊ะๆ จ๋าๆ น่าฟังเอย ฯ๛

== ทดสอบการจัดระดับ ==
ก็ก่ก้ก๊ก๋ก์กักิกีกึกืกุกูกฺกํ กี่กี้กี๊กี๋กี์กุ์กำก้ำ
ป็ป่ป้ป๊ป๋ป์ปัปิปีปึปืปุปูปฺปํ ปี่ปี้ปี๊ปี๋ปี์ปุ์ปำป้ำ
กุฎุฏุฐุญุฎูฏูฐูญูฎฺฏฺฐฺญฺ

Figure 4: Thai Font Rendering on Pango

The following Thai fonts are supported on Linux. The encoding standard that has been used for the specific fonts is also given along with each font specification.

TIS-620 BDF fonts
Manop
Phaisarn
Yenbut
NECTEC
Type1 fonts
DearBook
Omega /NECTEC
ISO10646 BDF fonts
XFree86
TrueType fonts
Omega/NECTEC

Figure 5: Thai Fonts and Encodings

Keyboard

TIS 820-2531 was the initial national keyboard standard [13], later modified to TIS 820-2538. Both these layouts are based on Ketmanee layout [8].

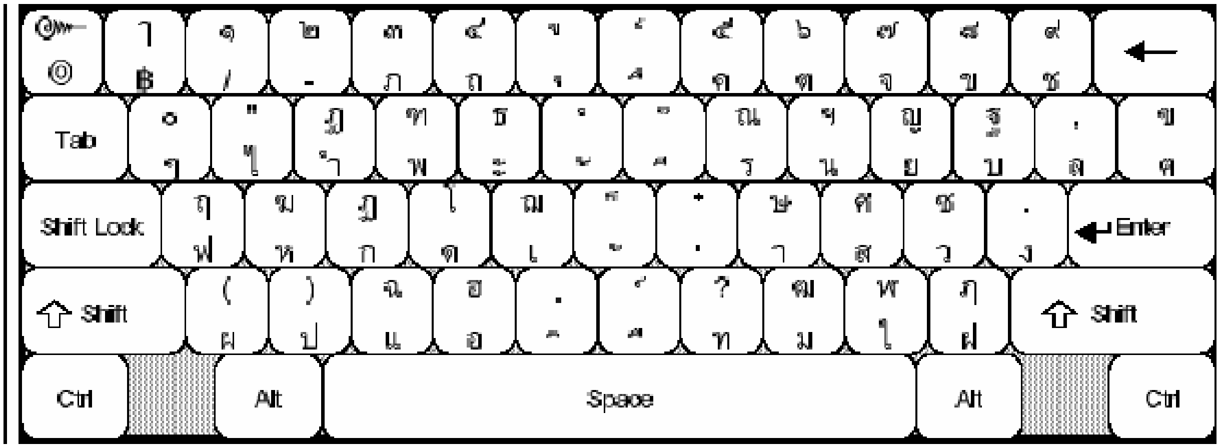
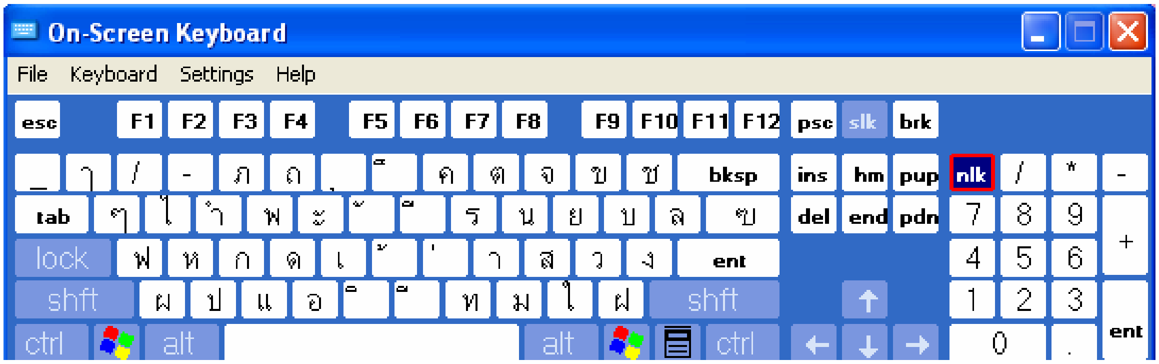


Figure 6: Standard Keyboard Layout TIS 820-2538 [8]

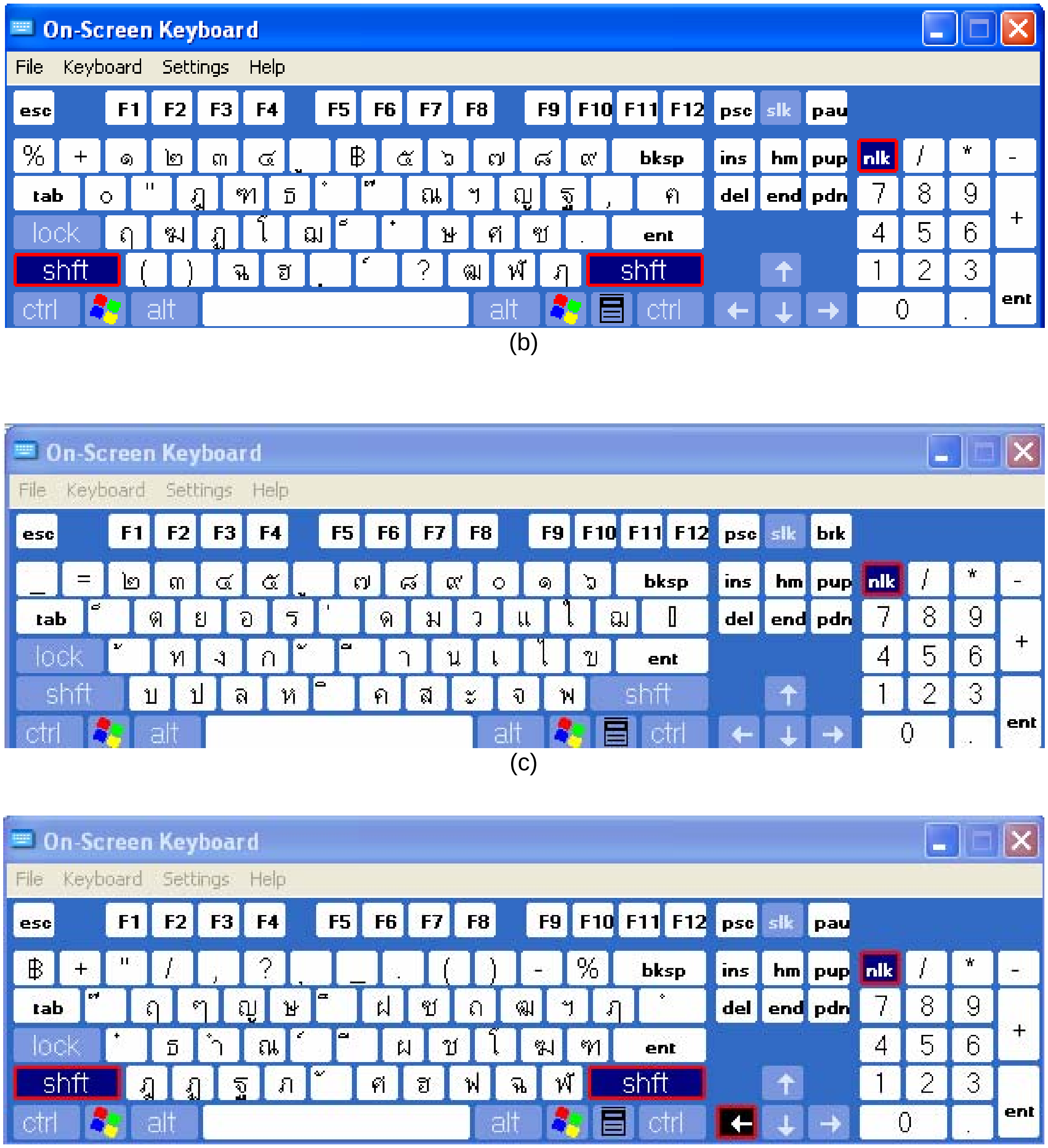
Thai language requires a complex input method, to use the keyboard and give adequate output. Though there is no official standard, there is now a single ad hoc solution supported by most vendors, e.g. Microsoft and Thai Language Environment (TLE) on Solaris. Details of this proposed standard, WTT 2.0, are given in [8].

Microsoft Platform

Microsoft Windows provides built-in support for Thai keyboard. Two different Thai keyboard layouts, Thai Ketmanee and Thai Pattachote, are available on Windows XP, shown in Figure 7.



(a)



(b)

(c)

(d)

Figure 7: Microsoft On-Screen Keyboard Layout for (a) Ketmanee (Normal Version), (b) Ketmanee (Shift Version), (c) Pattachote (Normal Version), and (d) Pattachote (Shift Version)

Linux Platform

Thai keyboard is available in Linux Red Hat version 9 and many other Thai Linux distributions. Ketmanee keymap has been supported as the standard keyboard layout in Thai Linux distributions.

Collation

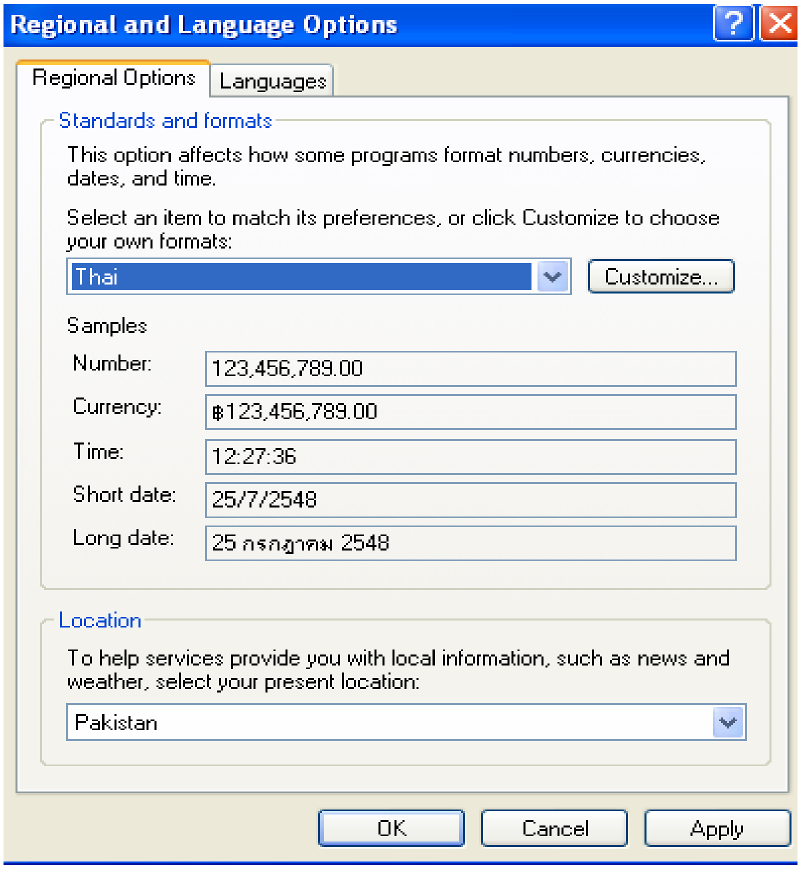
Thai collation standard is defined, and is based on Thai Royal Institute Dictionary 2525 B.E. edition, the official Thai dictionary. Thai encoding standards TIS 620 is based on this dictionary. Thai collation rules based on this collation order are also defined [8, 15, 16]. Thai collation is available on Linux platform, and also works on Microsoft platform.

Locale

Work started in 1990's for definition of Thai locale, which is now available and supported in Microsoft and Linux platforms [see 17]. Thai locale (th_TH) is completely supported in the IBM ICU locale data repository [18]. Both the Gregorian calendar and the Thai official Buddhist calendar are supported within the locale definition. In addition, the date format (long, medium, short), number formats, currency symbol, and weight and measures specific to the Thai conventions have been defined in the IBM ICU locale [18]. Thai locale is also available within CLDR 1.3. Also see [28, 29, 30] for further details.

Microsoft Platform

Microsoft provides support for Thai locale in Windows XP editions. If the system locale is switched to Thai, changes in date, time, and currency symbol of all application of Microsoft are displayed. This is shown in Figure 8.



This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Thai

Samples

Number: 123,456,789.00

Currency: ฿123,456,789.00

Time: 12:27:36

Short date: 25/7/2548

Long date: 25 กรกฎาคม 2548

To help services provide you with local information, such as news and weather, select your present location:

Pakistan

Figure 8: Thai Locale Settings on Microsoft Windows

Linux Platform

Thai locale is available on multiple platforms within Linux, including definition in Open Office and GNOME platforms [4, 19, 20].

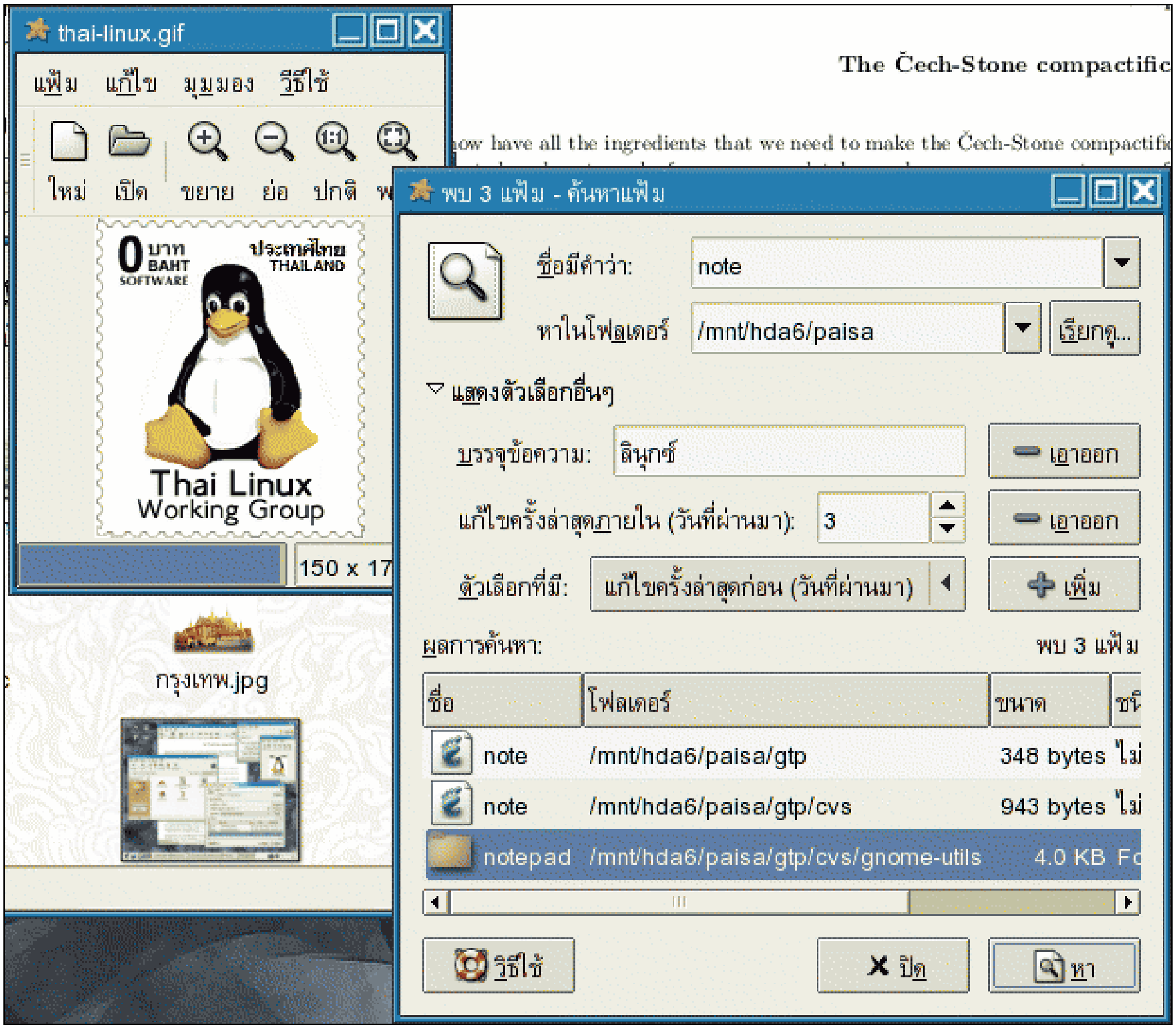
Interface Terminology Translation

Microsoft Platform

Localized versions of Microsoft Windows XP and Microsoft Office are available [11]. Proofing tools for the MUI pack of Office 2003 are also available to facilitate advanced desktop processing in Microsoft platform.

Linux Platform

Multiple Linux distributions are available. These are Kaiwal Linux, Linux School Internet Server (Linux SIS) with Thai Language Extension (Linux-TLE), and Burapha Linux [4]. In addition, the latest version of GNOME 2.12 and Firefox have been released, and work is also in progress on KDE [20]. Thai version of Open Office is also available [19]. Figure 9 below shows Thai Linux developed by the Thai Linux Working Group (TLWG).



The Čech-Stone compactific

ow have all the ingredients that we need to make the Čech-Stone compactifi

ชื่อมีคำว่า: note

หาในโฟลเดอร์ /mnt/hda6/paisa

บรรจุข้อความ: ลินุกซ์

แก้ไขครั้งล่าสุดภายใน (วันที่ผ่านมา): 3

ตัวเลือกที่มี: แก้ไขครั้งล่าสุดก่อน (วันที่ผ่านมา)

ผลการค้นหา: พบ 3 แฟ้ม

ชื่อ	โฟลเดอร์	ขนาด	ชนิ
note	/mnt/hda6/paisa/gtp	348 bytes	ไม่
note	/mnt/hda6/paisa/gtp/cvs	943 bytes	ไม่
notepad	/mnt/hda6/paisa/gtp/cvs/gnome-utils	4.0 KB	Fo

Figure 9: Thai Linux Distribution Developed by Thai Linux Working Group (TLWG) [20]

Status of Advanced Applications

Thailand has significantly progressed in Thai language processing, with R&D at NECTEC, the statutory government organization under Ministry of Science and Technology. The Center has established R&D cell to develop the fundamental resources such as sorting, line breaking and word breaking, lexicons and font development to support use of Thai language in software and operating systems.

NECTEC has developed the following projects (also see [23]):

- Khian Thai and Khat Thai, two Thai word processors
- Thai Sorting Program
- Thai word segmentation and line breaking
- Orchid, Thai text corpus [21] and POS annotated corpus
- Thai part-of-speech tagging program based on Orchid
- Lexitron, [22], an online English-Thai-English dictionary which is based on the Royal Institute Electronic dictionary also developed by NECTEC.
- Initial work on thesaurus and grammar checker for Thai has also been done
- ARN Thai, a Thai optical character recognition application [24]
- Parsit, an English-to-Thai machine translation system [25]
- Vaja, a Thai text to speech system [26]
- Sansarn, a Thai-English search engine [27]

Significant more work continues on further maturing these applications at NECTEC, universities and organizations across Thailand.

References

[1] http://www.ethnologue.com/show_language.asp?code=tha
[2] http://www.omniglot.com/writing/thai.htm
[3] http://www.tisi.go.th/eng/tisi.html
[4] Karoonboonyanan, T. and Koanantakool, T. "Standardization Activities and Open Source Movements in Thailand." http://www.nectec.or.th/it-standards/mlit99/mlit99-country.html
[5] http://www.nectec.or.th/it-standards/std620/std620.html
[6] http://www.nectec.or.th/it-standards/mlit97/country/gii2.htm
[7] Koanantakool, T., Tanprasert, C. and Viravan, C. "Country Report – Thailand." In the Proceedings of International Symposium on Standardization of Multilingual Information Technology, Singapore. http://mozart.inet.co.th/cyberclub/trin/thairef/tis620-iso10646.html, 1997
[8] Karoonboonyanan, T. "Standardization and Implementations of Thai Language." http://www.nectec.or.th/it-standards/thaistd.pdf.
[9] http://www.into-asia.com/thai_language/thaifont/?PHPSESSID=a9df4eab71d3e863f06aa8dc17f89641
[10] http://www.travelphrases.info/gallery/fonts_Thai.html
[11] Windows XP Thai LIP, http://www.microsoft.com/downloads/details.aspx?displaylang=th&FamilyID=0db2e8f9-79c4-4625-a07a-0cc1b341be7c
[12] http://www.microsoft.com/typography/otfntdev/thaiot/default.htm
[13] http://www.nectec.or.th/users/htk/it-standard/TISKB551.gif
[14] http://www.nectec.or.th/it-standards/
[15] http://linux.thai.net/~thep/
[16] http://www.open-std.org/jtc1/sc22/wg20/docs/n668.pdf
[17] http://linux.thai.net/~thep/th-locale/
[18] http://www-306.ibm.com/software/globalization/topics/thai/locale.jsp
[19] Open TLE, http://www.opentle.org/
[20] http://linux.thai.net/

[21] Orchid Thai Corpus, http://www.links.nectec.or.th/orchid/
[22] Lexitron Thai Dictionary, http://lexitron.nectec.or.th/index.php
[23] http://www.links.nectec.or.th/web_service.php
[24] ARN Thai OCR System, http://arnthai.nectec.or.th/arnthai.htm
[25] Parsit, http://suparsit.com/index1.php
[26] Vaja, Thai Text-to-Speech, http://vaja.nectec.or.th/onlinedemo/
[27] Sansarn, Thai Search Engine , http://sansarn.com/
[28] http://software.thai.net/locale/locale/14651/n537e.pdf
[29] ISO/IEC 14652 Cultural Convention Specification,
http://software.thai.net/locale/locale/14652/14652fcd.doc
[30] ISO/IEC 15435 Internationalization APIs, http://software.thai.net/locale/locale/15435/n536.pdf
[31] http://www.microsoft.com/globaldev/reference/wincp.mspx

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.

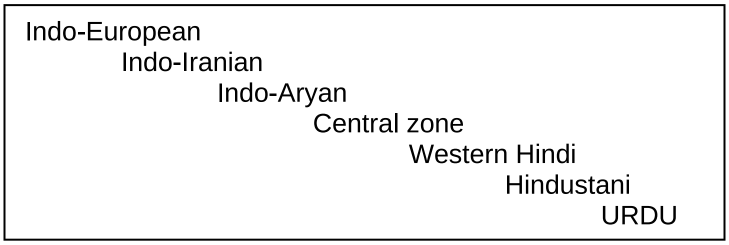


Figure 1: Language Family Tree for Urdu [1]

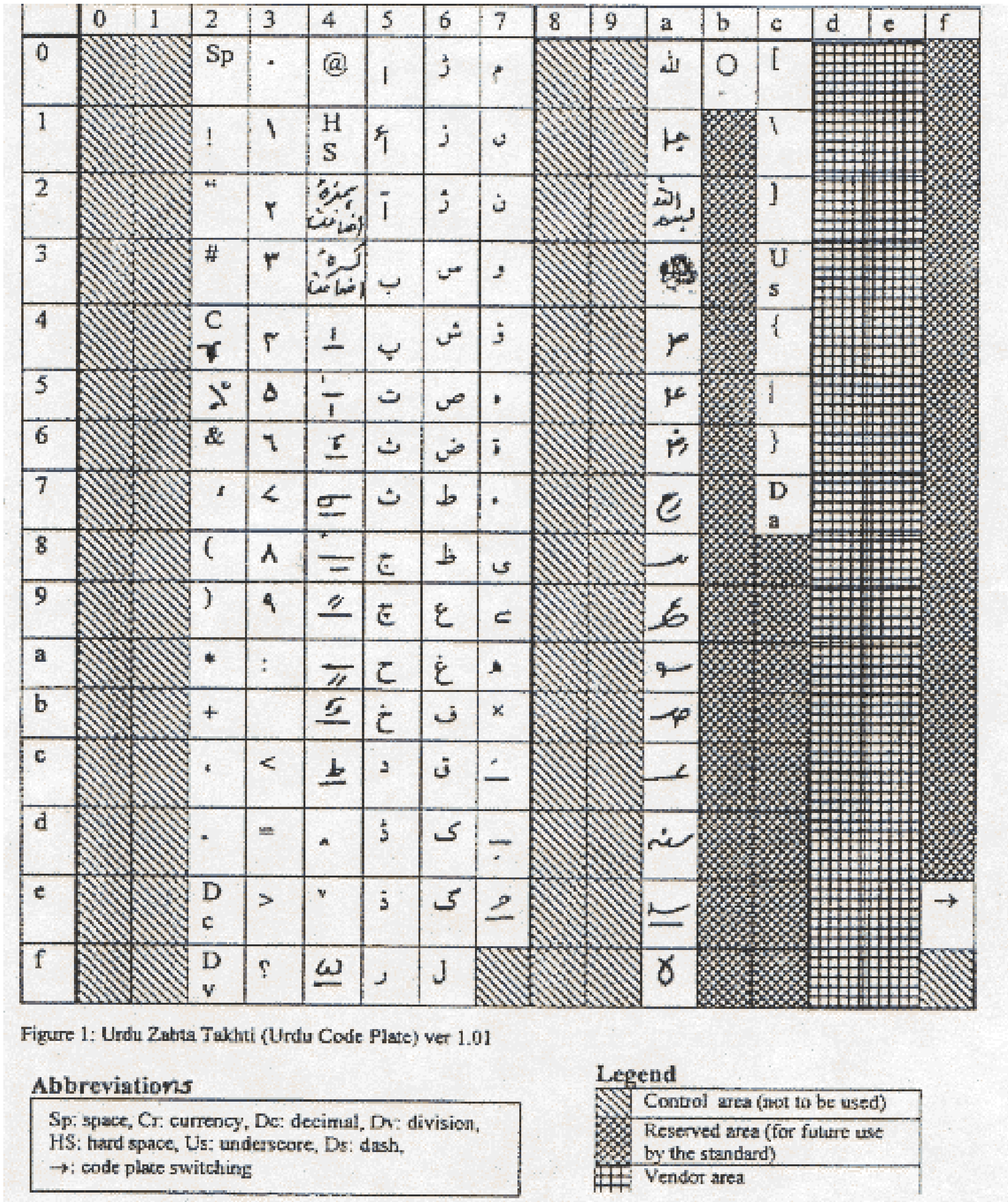
Perso-Arabic script written in Nastalique style is widely used for Urdu orthography [3].

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	٠	@	ا	ڑ	م	▨	▨	ﷲ	O	[	▦	▦	▩
1	▨	▨	!	١	HS	أ	ز	ں	▨	▨	ﷻ	▩	\	▦	▦	▩
2	▨	▨	"	٢	ہمزہ اضافت	آ	ژ	ن	▨	▨	﷽	▩	]	▦	▦	▩
3	▨	▨	#	٣	کسرہ اضافت	ب	س	و	▨	▨	[illegible]	▩	Us	▦	▦	▩
4	▨	▨	Cr	۴	◌ٰ	پ	ش	ؤ	▨	▨	[illegible]	▩	{	▦	▦	▩
5	▨	▨	٪	۵	◌ٖ	ت	ص	ہ	▨	▨	[illegible]	▩	|	▦	▦	▩
6	▨	▨	&	٦	◌ٔ	ٹ	ض	ۃ	▨	▨	[illegible]	▩	}	▦	▦	▩
7	▨	▨	'	۷	◌ً	ث	ط	ھ	▨	▨	[illegible]	▩	Da	▦	▦	▩
8	▨	▨	(	٨	◌ٍ	ج	ظ	ی	▨	▨	[illegible]	▩	▩	▦	▦	▩
9	▨	▨	)	٩	◌ٌ	چ	ع	ے	▨	▨	[illegible]	▩	▩	▦	▦	▩
a	▨	▨	*	:	◌ِ	ح	غ	ء	▨	▨	[illegible]	▩	▩	▦	▦	▩
b	▨	▨	+		◌ُ	خ	ف	×	▨	▨	[illegible]	▩	▩	▦	▦	▩
c	▨	▨	،	<	◌ط	د	ق	◌َ	▨	▨	[illegible]	▩	▩	▦	▦	▩
d	▨	▨	-	=	۔	ڈ	ک	◌ِ	▨	▨	[illegible]	▩	▩	▦	▦	▩
e	▨	▨	Dc	>	˅	ذ	گ	◌ُ	▨	▨	[illegible]	▩	▩	▦	▦	→
f	▨	▨	Dv	؟	◌ٓ	ر	ل	▨	▨	▨	[illegible]	▩	▩	▦	▦	▨

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

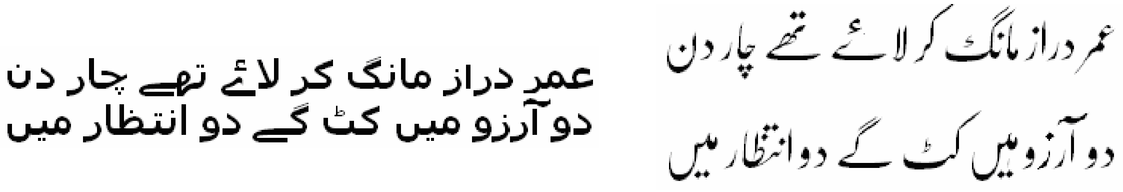
▨	Control area (not to be used)
▩	Reserved area (for future use by the standard)
▦	Vendor area

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.


عمر دراز مانگ کر لائے تھے چار دن
دو آرزو میں کٹ گے دو انتظار میں

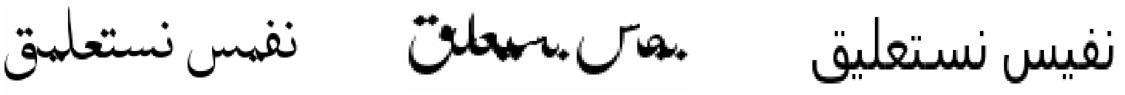
(a) (b)

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.



(a) (b) (c)

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.

□□ □□□□□□ □□
نفیس نستعلیق

(a) **(b)**

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.

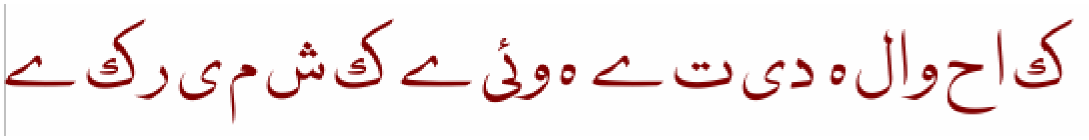


Figure 6: Urdu Text Written in Nafees Naskh in Open Office Version 1.1.0

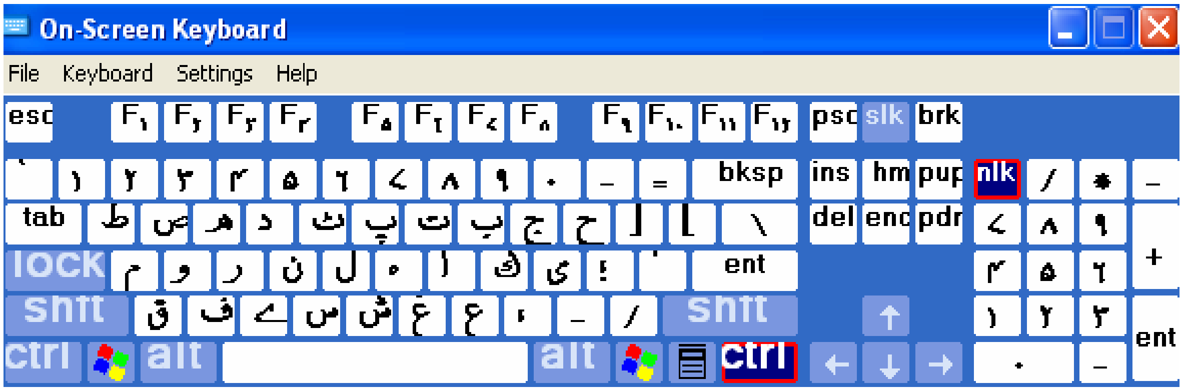
Results of rendering Urdu Open Type fonts on Mozilla are similar to those obtained through KDE's rendering engine However, Mozilla, instead of displaying boxes when rendering Open Type fonts (as in KDE), displays the same text in a simple four-shaped Naskh font.

Keyboard Layout

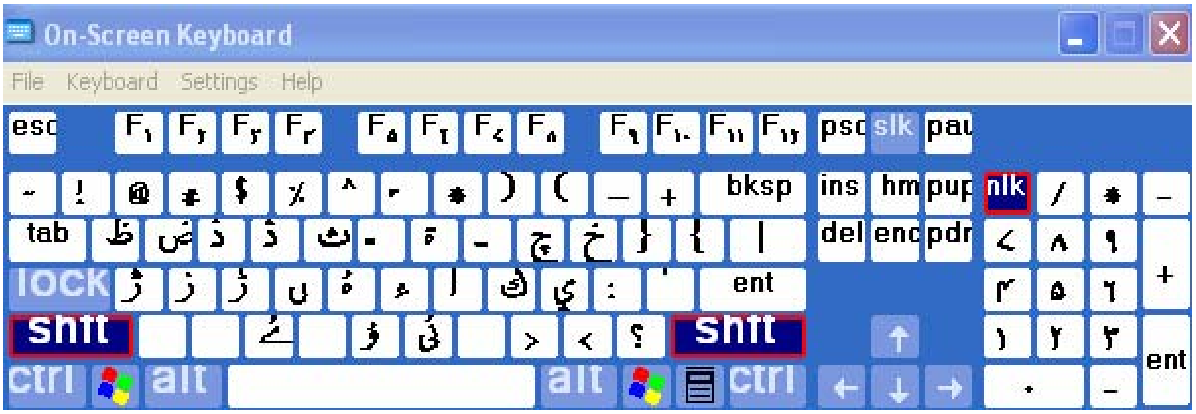
Work is underway to standardize Urdu keyboard layout in Pakistan. Similar to Urdu character set encoding formats, different software vendors have implemented different layouts for sequencing Urdu characters on the keyboard e.g. Katib, Rakim , Inpage, IBM, NLA and Microsoft keyboard layout [8, 9]. The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


On-Screen Keyboard
File Keyboard Settings Help
esc F۱ F۲ F۳ F۴ F۵ F۶ F۷ F۸ F۹ F۱۰ F۱۱ F۱۲ psc slk brk
` ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۰ – = bksp ins hm pup nlk / * –
tab ط ص ھ د ٹ پ ت ب ج ح] [\ del end pdn ۷ ۸ ۹ +
lock م و ر ن ل ہ ا ک ی ؛ ' ent ۴ ۵ ۶
shft ق ف ے س ش غ ع ، ۔ / shft ۱ ۲ ۳ ent
ctrl alt alt ctrl . –

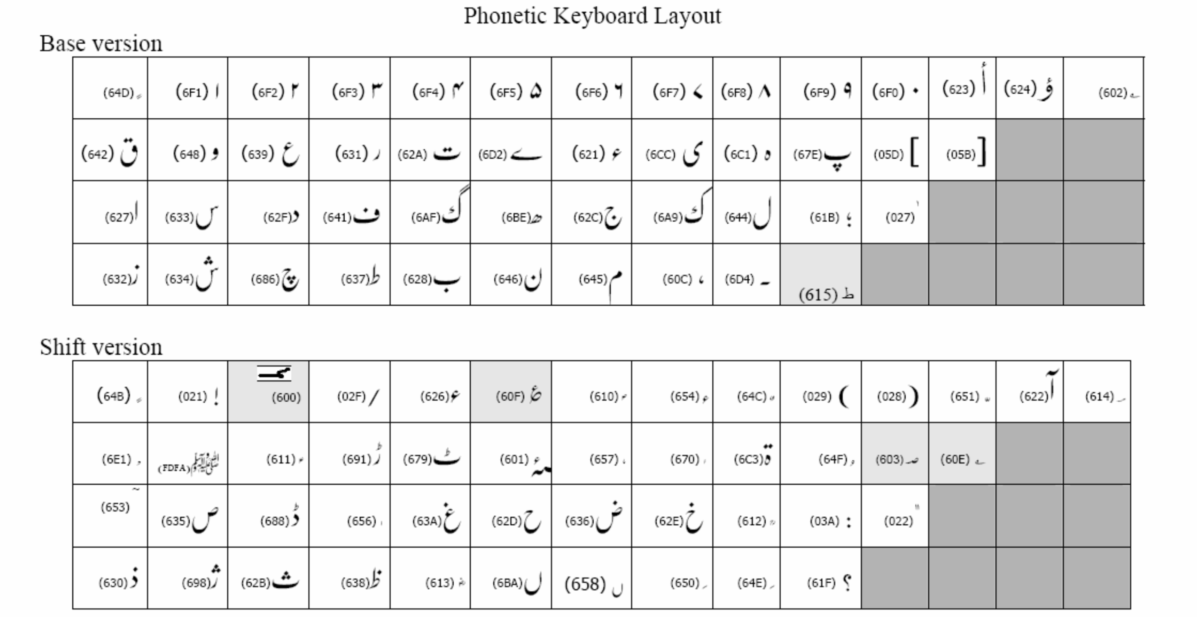
(a)


On-Screen Keyboard
File Keyboard Settings Help
esc F۱ F۲ F۳ F۴ F۵ F۶ F۷ F۸ F۹ F۱۰ F۱۱ F۱۲ psc slk pau
~ ! @ # $ % ^ & *) (_ + bksp ins hm pup nlk / * –
tab ظ ض ذ ڈ ث ۃ ـ چ خ } { | del end pdn ۷ ۸ ۹ +
lock ژ ز ڑ ں ۂ ء آ گ ي : " ent ۴ ۵ ۶
shft ۓ ؤ ئ > < ؟ shft ۱ ۲ ۳ ent
ctrl alt alt ctrl . –

(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) ٍ	(6F1) ۱	(6F2) ۲	(6F3) ۳	(6F4) ۴	(6F5) ۵	(6F6) ۶	(6F7) ۷	(6F8) ۸	(6F9) ۹	(6F0) ۰	(623) أ	(624) ؤ	(602) ؂
(642) ق	(648) و	(639) ع	(631) ر	(62A) ت	(6D2) ے	(621) ء	(6CC) ی	(6C1) ہ	(67E) پ	(05D) [	(05B)]		
(627) ا	(633) س	(62F) د	(641) ف	(6AF) گ	(6BE) ھ	(62C) ج	(6A9) ک	(644) ل	(61B) ؛	(027) '			
(632) ز	(634) ش	(686) چ	(637) ط	(628) ب	(646) ن	(645) م	(60C) ،	(6D4) ۔	(615) ؕ				

Shift version

(64B) ً	(021) !	؀ (600)	(02F) /	(626) ئ	(60F) ؏	(610) ؐ	(654) ٔ	(64C) ٌ	(029) (	(028))	(651) ّ	(622) آ	(614) ؔ
(6E1) ۡ	(FDFA) ﷺ	(611) ؑ	(691) ڑ	(679) ٹ	(601) ؁	(657) ٗ	(670) ٰ	(6C3) ۃ	(64F) ُ	(603) ؃	(60E) ؎		
(653) ٓ	(635) ص	(688) ڈ	(656) ٖ	(63A) غ	(62D) ح	(636) ض	(62E) خ	(612) ؒ	(03A) :	(022) "			
(630) ذ	(698) ژ	(62B) ث	(638) ظ	(613) ؓ	(6BA) ں	(658) ٘	(650) ِ	(64E) َ	(61F) ؟				

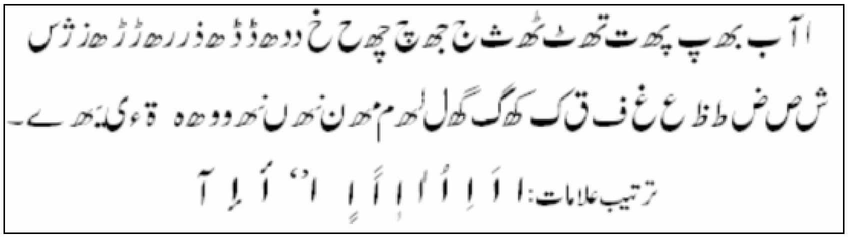
Figure 8: Urdu Phonetic Keyboard Layout by CRULP Extended From InPage Phonetic Keyboard [10]

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.



ا آ ب بھ پ پھ ت تھ ٹ ٹھ ث ج جھ چ چھ ح خ د دھ ڈ ڈھ ذ ر رھ ڑ ڑھ ز ژ س

ش ص ض ط ظ ع غ ف ق ک کھ گ گھ ل لھ م مھ ن نھ ں ںھ و وھ ہ ۃ ء ی یھ ے ـ

ترتیب علامات: ا اَ اِ اُ اْ اٖ اٰ اٍ ا“ اً اِ آ

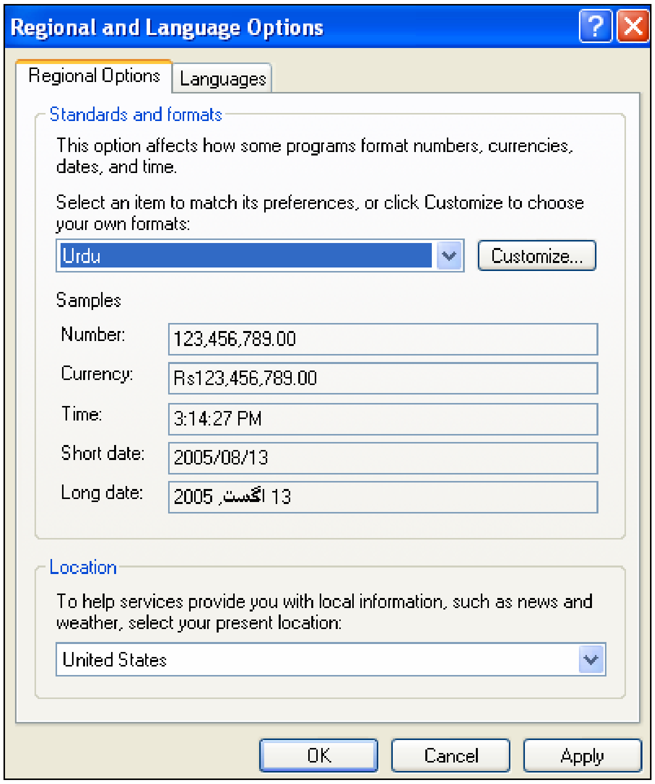
Figure 9: Standard Collation Sequence for Urdu [12]

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.



This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Urdu

Number: 123,456,789.00

Currency: Rs123,456,789.00

Time: 3:14:27 PM

Short date: 2005/08/13

Long date: 2005 ,اگست 13

To help services provide you with local information, such as news and weather, select your present location:

United States

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.

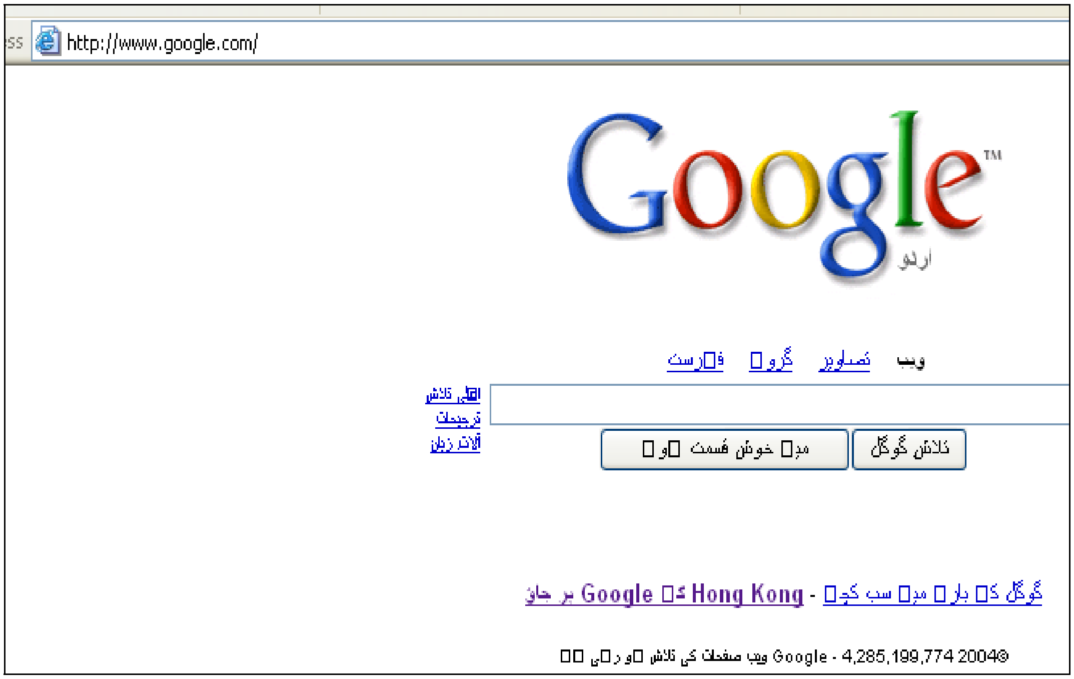
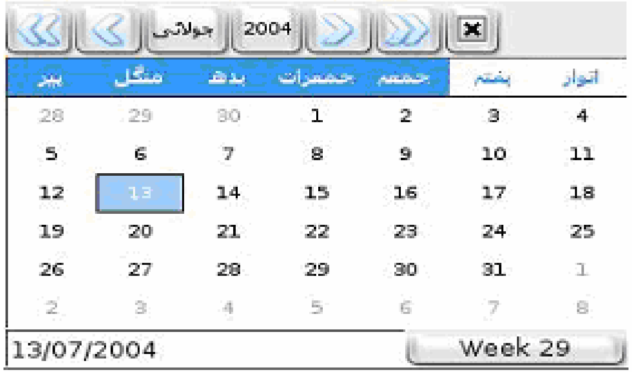

http://www.google.com/
Google™
اردو
ویب
تصاویر
گروہ□
ف□رست
اعلیٰ تلاش
ترجیحات
آلات زبان
تلاش گوگل
میں□ خوش قسمت □و□ □
گوگل ک□ بار□ می□ سب کچھ□ - Hong Kong ک□ Google پر جاؤ
Google - 4,285,199,774 2004© ویب صفحات کی تلاش □و ر□ی □□

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.



پیر	منگل	بدھ	جمعرات	جمعہ	ہفتہ	اتوار
28	29	30	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	1
2	3	4	5	6	7	8

13/07/2004

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.

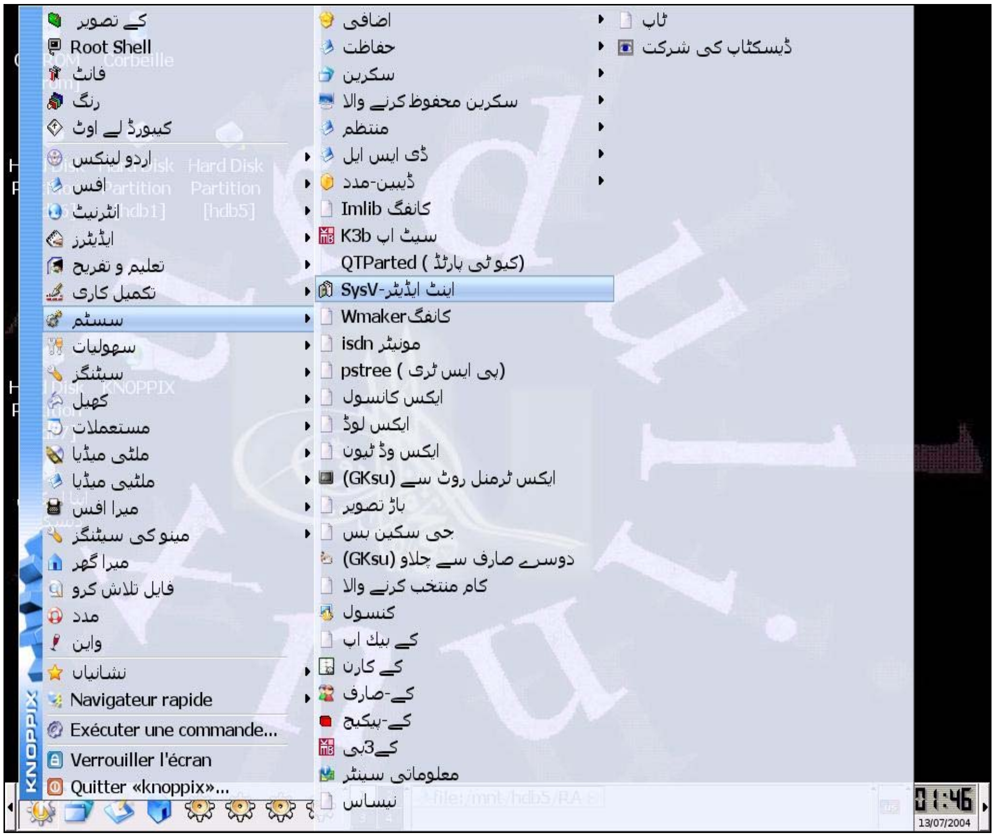


Figure 13: Localized Start-Up Menu for KDE in Urdu Distribution by CRULP

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.crulp.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.crulp.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.crulp.org
[11] http://groups.msn.com/SindhiComputing
[12] www.nla.gov.pk
[13] http://www.emille.lancs.ac.uk/home.htm

Vietnamese

Vietnamese is an Austro-Asiatic language spoken by about 68 million people of Vietnam. Vietnamese is the official language in Vietnam and is also spoken in some parts of Australia, Cambodia, Canada, China and Thailand [1, 2].

Austro-Asiatic
Mon-Khmer
Viet-Muong
VIETNAMESE

Figure 1: Language Family Tree of Vietnamese [1]

Vietnamese was initially written in classical Chinese writing Chữ-nho. This was later adapted in 10th century AD and called Chữ-nôm. In 17th centuary, western missionaries started developing a Latin based script, called Quốc Ngữ, which is now in widespread use [3]. Quốc Ngữ is the nationally adopted standard since 1945 [4]. Use of Chữ-nho and Chữ-nôm continued until about 1918. However, these are no longer in use (see [3] for more details).

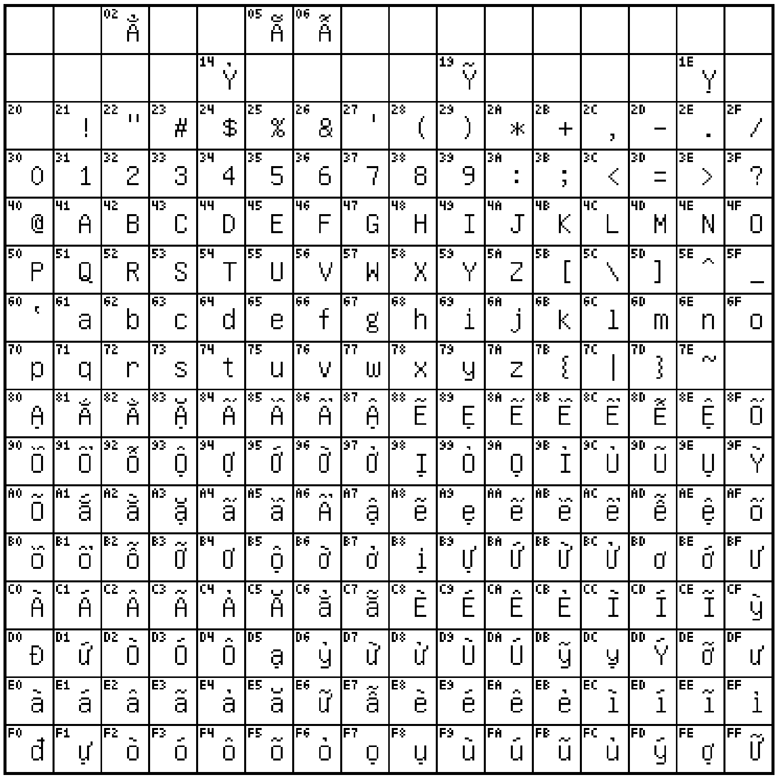
Character Set and Encoding

Initially, due to absence of any standards, more than 30 different encodings were developed by different national and international vendors. These included single byte code table, e.g., 3C30 (3C Corporation), Cyrillix, VW1, VSCII, Daisy, 2font, Vietkey, ACAD, ABC, etc., and two byte code tables, e.g., 3C25, VW2, ATM2, VNij, VNI, etc. [4].

To resolve the situation, a task force was set up in 1991 by the Ministry of Science and Technology, which produced a draft for Vietnamese Standard Code Set for Information Interchange (VSCII) or TCVN 5712:1993, which was approved in 1993 as the first national standard in IT [4, 5]. This standard has been revised to TCVN 5712:1999 in 1999 [9]. Other encodings include Microsoft's cp 1258 (developed in 1996) and IBM's cp 1129 (developed in 1997), which are slightly different from each other [4].

Now Unicode is increasingly used. As Latin script is used, Vietnamese does not have a separate code table within Unicode, but characters within the Latin tables are used. These characters are spread across multiple tables within Unicode. For one byte, VSCII provides an alternative. However, there is a national standard for two-byte standard as well, TVCN 6909:2001.

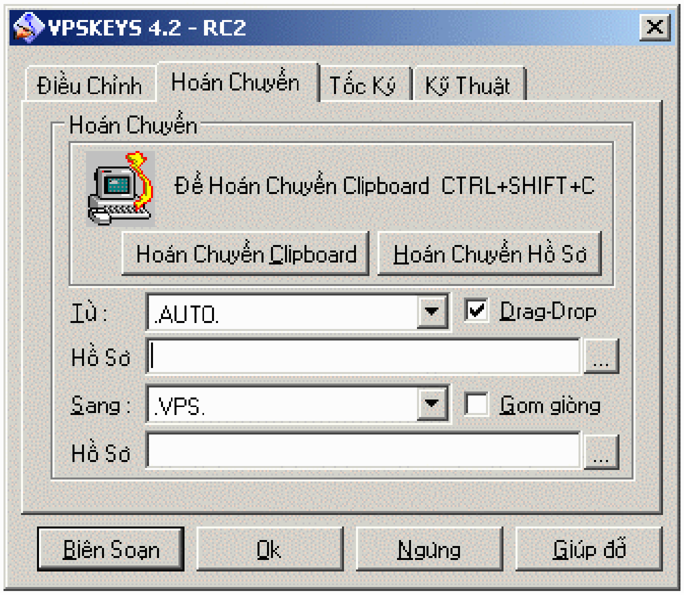
There has also been work to revive the earlier writing systems. Relevant national standards include Chu Nom 16-bit character encoding standard TCVN 5773:1993 and Chu Nom Han 16-bit character encoding standard TCVN 6056:1995 [9].



		02 Ẳ			05 Ẵ	06 Ẫ									
				14 Ỷ					19 Ỹ					1E Ỵ	
20	21 !	22 "	23 #	24 $	25 %	26 &	27 '	28 (	29)	2A *	2B +	2C ,	2D -	2E .	2F /
30 0	31 1	32 2	33 3	34 4	35 5	36 6	37 7	38 8	39 9	3A :	3B ;	3C <	3D =	3E >	3F ?
40 @	41 A	42 B	43 C	44 D	45 E	46 F	47 G	48 H	49 I	4A J	4B K	4C L	4D M	4E N	4F O
50 P	51 Q	52 R	53 S	54 T	55 U	56 V	57 W	58 X	59 Y	5A Z	5B [	5C \	5D]	5E ^	5F _
60 `	61 a	62 b	63 c	64 d	65 e	66 f	67 g	68 h	69 i	6A j	6B k	6C l	6D m	6E n	6F o
70 p	71 q	72 r	73 s	74 t	75 u	76 v	77 w	78 x	79 y	7A z	7B {	7C |	7D }	7E ~	
80 Ạ	81 Ắ	82 Ằ	83 Ặ	84 Ấ	85 Ầ	86 Ẩ	87 Ậ	88 Ẽ	89 Ẹ	8A Ế	8B Ề	8C Ể	8D Ễ	8E Ệ	8F Ố
90 Ồ	91 Ổ	92 Ỗ	93 Ộ	94 Ợ	95 Ớ	96 Ờ	97 Ở	98 Ị	99 Ỏ	9A Ọ	9B Ỉ	9C Ủ	9D Ũ	9E Ụ	9F Ỳ
A0 Õ	A1 ắ	A2 ằ	A3 ặ	A4 ấ	A5 ầ	A6 ẩ	A7 ậ	A8 ẽ	A9 ẹ	AA ế	AB ề	AC ể	AD ễ	AE ệ	AF ố
B0 ồ	B1 ổ	B2 ỗ	B3 Ỡ	B4 Ơ	B5 ộ	B6 ờ	B7 ở	B8 ị	B9 Ự	BA Ứ	BB Ừ	BC Ử	BD ơ	BE ớ	BF Ư
C0 À	C1 Á	C2 Â	C3 Ã	C4 Ả	C5 Ă	C6 ẳ	C7 ẵ	C8 È	C9 É	CA Ê	CB Ẻ	CC Ì	CD Í	CE Ĩ	CF ỳ
D0 Đ	D1 ứ	D2 Ò	D3 Ó	D4 Ô	D5 ạ	D6 ỷ	D7 ừ	D8 ử	D9 Ù	DA Ú	DB ỹ	DC ỵ	DD Ý	DE ỡ	DF ư
E0 à	E1 á	E2 â	E3 ã	E4 ả	E5 ă	E6 ữ	E7 ẫ	E8 è	E9 é	EA ê	EB ẻ	EC ì	ED í	EE ĩ	EF ỉ
F0 đ	F1 ự	F2 ò	F3 ó	F4 ô	F5 õ	F6 ỏ	F7 ọ	F8 ụ	F9 ù	FA ú	FB ũ	FC ủ	FD ý	FE ợ	FF Ữ

Figure 2: VSCII 1993 Character Set Encoding [5]

Text conversion utilities among various encodings are also available [10, 11]. Figure 3 shows one such utility.



Để Hoán Chuyển Clipboard CTRL+SHIFT+C

Từ : .AUTO. ☑ Drag-Drop

Hồ Sơ

Sang : .VPS. ☐ Gom giòng

Hồ Sơ

Figure 3: Encoding Conversion Utility for Vietnamese [10]

Fonts and Rendering

Latin script is supported by most platforms and many Vietnamese fonts are available for use. These fonts are based on a variety of encodings discussed above (e.g. see [6]).

Windows Platform

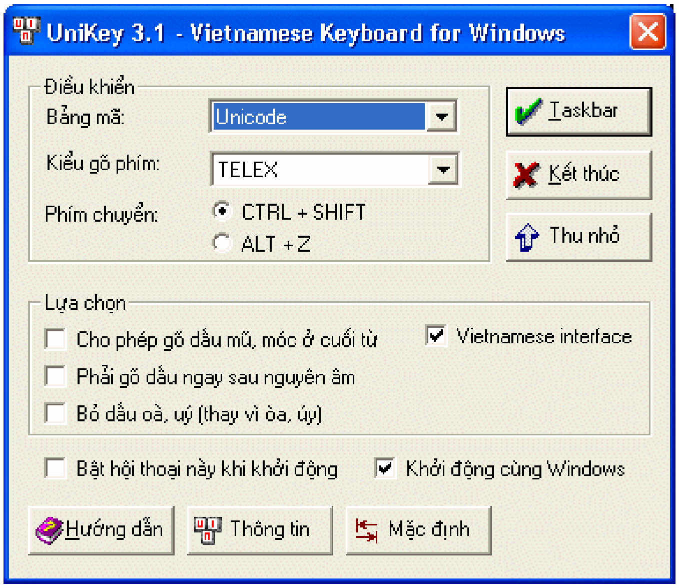
Microsoft XP does not include specific fonts for Vietnamese language but fonts for Unicode (e.g. Arial Unicode MS) cover the characters in Vietnamese as well. However many Vietnamese Unicode Open Type and True Type fonts have been developed by different research groups [4].

Linux Platform

On the Linux platform, fonts based on Unicode [6], TCVN, VNI and VPS [7] encodings can be adequately used to input Vietnamese text.

Keyboard

A national standard for Vietnamese keyboard TCVN 6064:1995 has been developed. It is based on international keyboard layout ISO 9995. TELEX and VNI are two other popularly used standards [4]. These keyboards are available, e.g. Unikey keyboard which allows user to configure any of these layouts on the keyboard for any of the encodings, as shown in Figure 4.



UniKey 3.1 - Vietnamese Keyboard for Windows

Điều khiển

Bảng mã: Unicode

Kiểu gõ phím: TELEX

Phím chuyển: ● CTRL + SHIFT ○ ALT + Z

Lựa chọn

□ Cho phép gõ dấu mũ, móc ở cuối từ

☑ Vietnamese interface

□ Phải gõ dấu ngay sau nguyên âm

□ Bỏ dấu oà, uý (thay vì òa, úy)

□ Bật hội thoại này khi khởi động

☑ Khởi động cùng Windows

Figure 4: Unikey, Vietnamese Keyboard Layout Utility [11]

Keyboards use different schemes to generate the characters. Some generate pre-composed characters while others generate base character and diacritics separately. The variation in output has to be normalized before further processing. Normalization utilities are also available.

Microsoft Platform

Microsoft provides built-in Vietnamese keyboard. The basic Vietnamese characters are in similar position as in default QWERTY keyboard layout. Figure 5 below shows the normal state Vietnamese keyboard layout on MS platform.

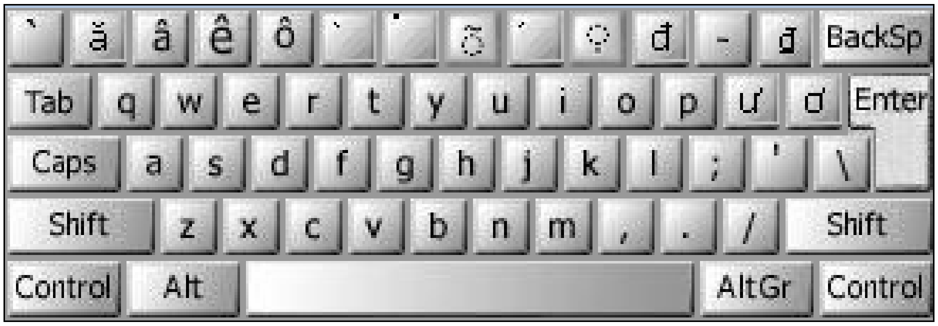


Figure 5: Keyboard Layout for Vietnamese [8]

Linux Platform

XVNKB is a Vietnamese keyboard input for X-Window. It provides a useful way for editing Vietnamese on X-Windows environment with popular input methods and character sets. This software has been released under GPL license [12].

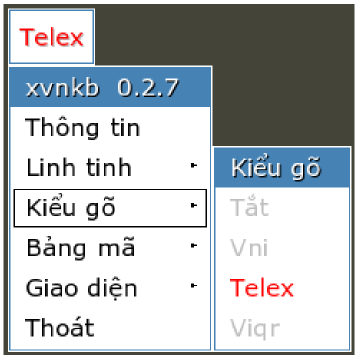


Figure 6: XVNKB Menus [12]

Collation

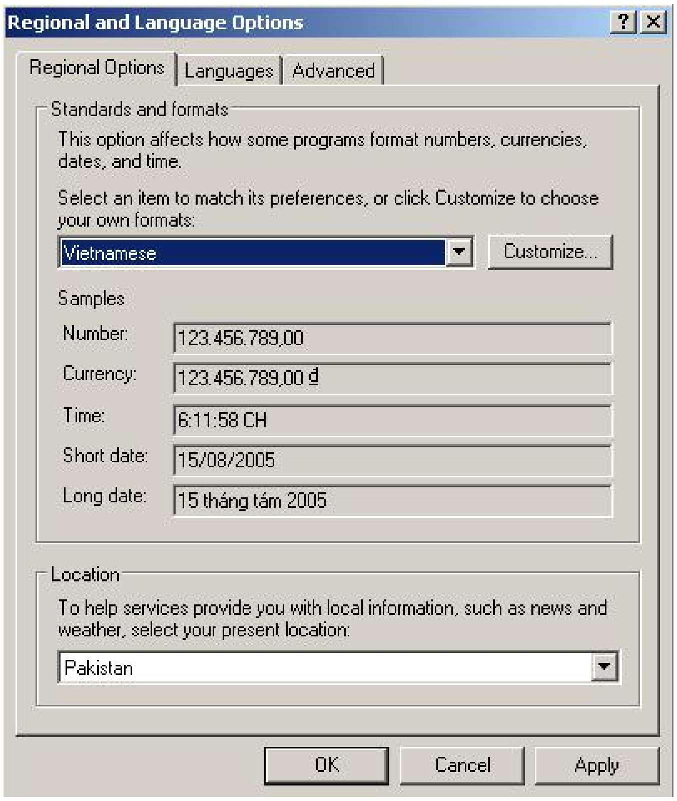
Collation order for Vietnamese written in Quoc Ngu has been developed as part of TCVN 5712:1993 standard [4] (also see [13, 14] for details of collation sequence). Sorting for Vietnamese has been enabled on Microsoft and Linux platforms, with minor problems still persisting [15].

Locale

Locale for Vietnamese (vi_VN) has been developed. TCVN/JTC1 and ITSC developed basic locale definition such as date, time, length, volume and weight measurements, and a conversion between the Gregorian (Western) and Lunar (Vietnamese) calendars. Standard locale for Vietnamese has also been included in IBM ICU and CLDR 1.3.

Microsoft Platform

Microsoft Windows XP provides support for Vietnamese locale. If system locale is switched to Vietnamese, changes may be monitored in date, time, and currency symbol within all application of Microsoft. The following figure shows Vietnamese locale on MS platform.



Standards and formats

This option affects how some programs format numbers, currencies, dates, and time.

Select an item to match its preferences, or click Customize to choose your own formats:

Vietnamese

Samples

Number:	123.456.789,00
Currency:	123.456.789,00 ₫
Time:	6:11:58 CH
Short date:	15/08/2005
Long date:	15 tháng tám 2005

Location

To help services provide you with local information, such as news and weather, select your present location:

Pakistan

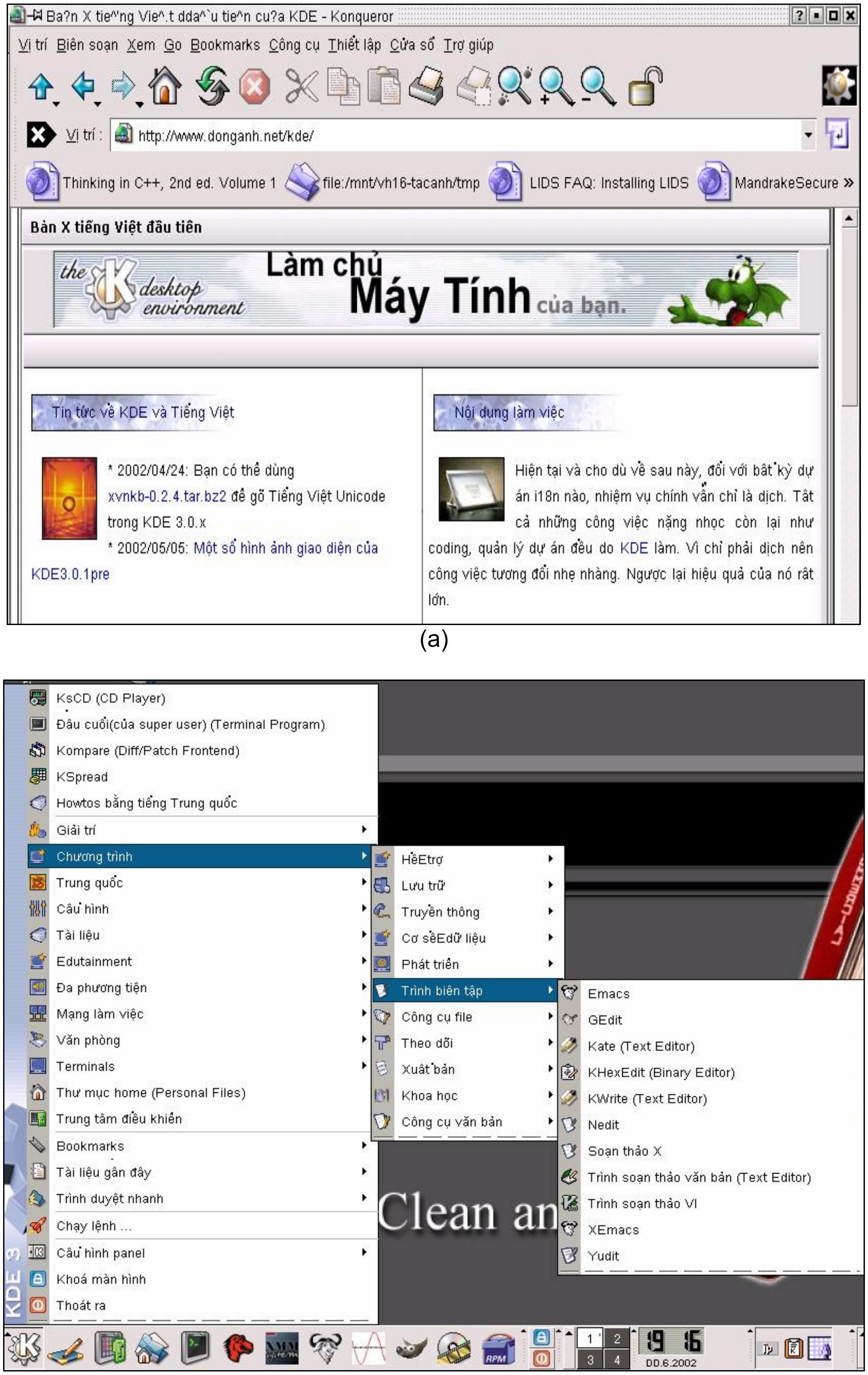
Figure 7: Vietnamese Locale on MS Platform

Interface Terminology Translation

TCVN/JTC1 also established glossary of Vietnamese and English terms for all the graphical user interface (GUI), icon names, dialogue boxes in 2000 through the standard TCVN 6695-1:2000.

Localized interface in Vietnamese is available on both Microsoft and Linux platforms. Microsoft support has been available for a long time, since Windows 95. Microsoft Language Interface Pack for Vietnamese is available [17].

Linux support has been developed later. The project on the development of VNLinux consists of two sub-modules: Development of a VNLinuxCD based on Mandriva Linux, designed for desktop use, and a VNLS server oriented distribution based on EnGarde Secure Linux [16]. A Vietnamese localization team has registered for GNOME glossary translation. As reported on the website about 96.41% of the GNOME glossary translation has been completed [18] while glossary translation of KDE has also been completed [19]. Figure 8 given below shows localized GNOME and KDE based Linux in Vietnamese.



Ba?n X tie^ng Vie^.t dda^`u tie^n cu?a KDE - Konqueror
Vị trí Biên soạn Xem Go Bookmarks Công cụ Thiết lập Cửa sổ Trợ giúp
Vị trí :
Thinking in C++, 2nd ed. Volume 1
file:/mnt/vh16-tacanh/tmp
LIDS FAQ: Installing LIDS
MandrakeSecure »
the desktop environment
Làm chủ Máy Tính của bạn.

http://www.donganh.net/kde/

Bản X tiếng Việt đầu tiên

Tin tức về KDE và Tiếng Việt

* 2002/04/24: Bạn có thể dùng xvnkb-0.2.4.tar.bz2 để gõ Tiếng Việt Unicode trong KDE 3.0.x

* 2002/05/05: Một số hình ảnh giao diện của KDE3.0.1pre

Nội dung làm việc

Hiện tại và cho dù về sau này, đối với bất kỳ dự án i18n nào, nhiệm vụ chính vẫn chỉ là dịch. Tất cả những công việc nặng nhọc còn lại như coding, quản lý dự án đều do KDE làm. Vì chỉ phải dịch nên công việc tương đối nhẹ nhàng. Ngược lại hiệu quả của nó rất lớn.

(a)

(b)

Figure 8 (a) Konqueror KDE Browser, (b) Localized KDE Start-Up Menus

Status of Advanced Applications

Much work has been done on Vietnamese language processing. Vietnamese spell checkers are available for Microsoft [21] and Linux platforms, and through other vendors (e.g. [20]). There has also been considerable work on Vietnamese and Vietnamese-English lexicons. English-Vietnamese machine translation systems were available as early as early 1990's. Much more work has been done on it and on Vietnamese-French machine translation systems [22]. Vietnamese text to speech systems and speech recognition systems are also available through private vendors. Work has also started on Vietnamese hand writing recognition systems.

References

[1] http://www.ethnologue.com/show_language.asp?code=vie
[2] http://www.omniglot.com/writing/vietnamese.htm
[3] http://www.omniglot.com/writing/chunom.htm
[4] Chuong, T., Hoang, N., Nhan, N., Phuoc, D., Viet, D. "Current Status of Vietnamese Language Processing Multilingual Processing," in Proceeding of MLIT '97, Tokyo, Japan, 1997. Also available at http://www.informatik.uni-leipzig.de/~duc/software/misc/viet.html
[5] http://czyborra.com/charsets/vietnamese.html
[6] http://www.vietgate.net/fonts/
[7] http://www.fedu.uec.ac.jp/~vuhung/linux/font-rpms/
[8] http://www.microsoft.com/globaldev/keyboards/kbdvntc.htm
[9] "Human Resource Development Policies of Information Technology in Vietnam." http://www.cicc.or.jp/japanese/kouryu/pdf_ppt/vietnam.pdf
[10] http://www.vps.org/article.php3?id_article=274
[11] http://unikey.sourceforge.net/
[12] http://xvnkb.sourceforge.net/
[13] Lương, V. "Vietnamese Sorting Rules for Dictionary Entries." Vietnam Lexicography Center. http://vietunicode.sourceforge.net/charset/quytacABC_en.html
[14] http://vietunicode.sourceforge.net/charset/vietalphabet.html
[15] http://blogs.msdn.com/michkap/archive/2005/08/27/457224.aspx
[16] http://distrowatch.com/table.php?distribution=vnlinux
[17] http://www.microsoft.com/downloads/details.aspx?FamilyID=0db2e8f9-79c4-4625-a07a-0cc1b341be7c&DisplayLang=vi
[18] http://l10n-status.GNOME.org/GNOME-2.8/vi/index.html
[19] http://i18n.KDE.org/teams/index.php?a=i&t=vi
[20] http://www.vnisoft.com/?http://www.vnisoft.com/xpfeatures.html
[21] http://www.worldlanguage.com/ProductScreenShots/104793.htm
[22] Dien, D. and Kiem, H. "State of Art of Machine Translation in Vietnam," in AAMT Journal, Special Issue, September 2005, Thailand.

Summary

The survey has looked at the extent to which twenty Asian languages and their scripts have been localized, reporting on the development of standards and basic and advanced localization applications in these languages. The study shows that these languages have a varying degree of support. Because a diverse set of development is surveyed, it is not easily possible to develop a quantitative scale to measure this development. Nevertheless, it is still possible to qualitatively gauge the level of maturity in language computing across these languages.

For many of these languages, national standards required for localization have been defined. These standards have been reviewed and revised over time and generally agree with the current international standards. However, there are also languages for which the national standards have only recently been devised or still remain undefined. The survey indicates that Chinese, Japanese and Korean have very mature standards. There has also been significant amount of work on Thai and Vietnamese standards. Arabic, Bengali, Nepali, Hindi, Mongolian (Cyrillic script) and Urdu standards are mostly defined, though some of these standards have been recently developed and therefore will mature over next few years. Some work on standards for Burmese, Dzongkha, Farsi, Indonesian, Sinhala and Telugu has been done, but more work is needed. Finally, much more work is required for Lao, Khmer, Mongolian (Mongolian script) and Pashto languages.

Localization support for Chinese, Japanese, Korean and Thai is well developed and deployed. Such applications are also available for Arabic, Bengali, Hindi and Vietnamese, though need further development. To a slightly lesser extent, these applications are also available for Burmese, Dzongkha, Nepali, Farsi, Khmer, Sinhalese, Telugu and Urdu. Most work needs to be done for Indonesian, Lao, Mongolian and Pashto.

Advanced applications require most resources, effort and expertise. Consequently, these applications are not available for most languages. Again, Chinese, Japanese and Korean have most of the advanced applications available, and Arabic is almost similarly developed. There is also reasonable amount of work in progress on Hindi, Thai, Urdu and Vietnamese. Most of the other languages, including, Bengali, Indonesian, Sinhalese and Telugu have little work done in this area. And there is almost no available work for development of advanced applications for Burmese, Dzongkha, Nepali, Farsi, Khmer, Lao, Mongolian and Pashto.

Overall, Chinese, Korean and Japanese language computing is very mature. Arabic and Thai also have reasonably mature language computing. There is also reasonable progress being made for Hindi, Urdu and Vietnamese. Other languages, including, Bengali, Sinhalese, Telugu are lagging but still there is some work being done. However the remaining languages covered in the survey have very limited computing support and significant work needs to be undertaken for the maturity of language technology in these languages.

Currently there is considerable focus on local language computing, as the technology is finally ripe to support it and ICT is increasingly getting significant in daily life of Asian populations. The implications of local language computing on development are tremendous, but so are the challenges to achieve the goals. Even for the twenty languages surveyed a lot of ground has to be covered to enable computing for most of them, though these hold national and official language status in Asian countries. For the remaining about 2180 languages of Asia, especially the non-official languages, the challenges are even more severe. Enabling local language computing for Asia, therefore, requires long term commitment and significant focus within ICT policy. Only sustained investment in language technology will enable us to completely harness the promised benefits of ICTs for the development of Asian populations.

Glossary

A

American Standard Code for Information Interchange (ASCII): A 7-bit (128 character) standard to code English text in computers. Also adopted by International Standards Organization as ISO IEC 646, and similar to 8-bit ISO 8859-1 standard. ASCII has also been adopted within Unicode standard from 0000-00FF.

Apple Advanced Typography: Font formalism by Apple which can be used to model complex scripts.

ASMO: ASMO 449 is a 7-bit character encoding standard for Arabic. ASMO 708 is another code page based on ISO 8859. These are created by Arab Organization for Standardization and Metrology.

Automatic Speech Recognition (ASR): Automatic Speech Recognition is a technology that recognizes human speech input into text in a computer.

B

Bidirectional Scripts: Some scripts (and languages) are written from right to left and with digits written from left to right. These are known as bidirectional Scripts (and languages) and include Arabic, Hebrew and Syriac.

Bidirectional Algorithm: Unicode has defined a standard algorithm to hand the bidirectional behavior of different scripts. This is called bidirectional algorithm.

C

CJK: Chinese, Japanese and Korean. All these languages share similar writing systems. Sometimes Vietnamese is also grouped, acronym CJKV, when written in traditional script.

CODAR-U: A 7-bit character encoding standard for Arabic developed by the Institut d'Etudes et de Recherches pour l'Arabisation (Rabat) in 1982.

Collation: Sorting of words according to lexicographic (dictionary) order of a language.

Common Locale Data Repository (CLDR): Repository that maintains locale information of all languages. Data is stored in XML format which would make the exchange of locale data convenient between different applications and keep it consistent across platforms.

Corpus: A collection of tagged text documents and sources for a language used for statistical modeling in language processing.

D

Distribution: A distribution comprises of an operating system on top of which different desktops, application suites and development tools etc. are included. A Linux distribution is GNU/Based.

Popular distributions for Linux include Red Hat (now released as Fedora Core), Debian, Caldera, SUSE and Mandrake (or Mandrivia) etc.

Default Unicode Collation Element Table (DUCET): A table which provides the default collation elements or weights for each character for each language. These are used by Unicode Collation Algorithm for sorting Unicode strings if language specific weights are not otherwise standardized.

E

Extended Binary Coded Decimal Interchange Code (EBCDIC): An 8-bit character encoding used on IBM mainframe operating systems. The first four (zone) bits represent the category of a character and the last four (digit) represents the character itself.

Extended Unix Coding (EUC): An 8-bit character encoding standard, used to represent character sets which can not be encoded within 7 bits. The structure of EUC is based on ISO 2022 and it is mainly used for CJK. EUC-JP stands for EUC-Japanese.

F

Firefox: Firefox, also known as FireBird or Phoenix, is a Mozilla based web browser. Firefox can be used on multiple platforms including Linux and Microsoft.

G

G-Edit: The formal text-editor for GNOME desktop.

GLIBC: GNU C library, also known as GLIBC, deals with system calls and is used with other important components in GNU system. Locale is also defined in GLIBC as well.

Glyph: A physical or visual representation of a character(s).

GNOME: A Unix/Linux based multilingual desktop environment shipped with browsers, editors and development tools etc.

GNU: GNU is a UNIX like operating system. The term stands for hybrid approach having Linux kernel running on GNU based systems.

GPL: The GNU General Public License (GNU GPL or simply GPL) is a license for free and open software.

H

Han Characters: These are traditionally used characters (ideograms and pictograms) for writing Chinese. They are also known as Hanzi.

Hangul: Hangul is indigenous character set used now days for Korean. They initially used the Hanja system borrowed from China. Hangul is phonemic based.

Hanja: Koreans initially borrowed some of the Han characters and incorporated them into their language. This set of characters is known as Hanja.

Hiragana: One of the three Japanese writing systems, Hiragna (ordinary syllabic script) is commonly used by children and for the words for which Kanji variant is not available.

Hinting: A process used to enhance the quality of glyph/image at smaller font sizes.

I

IBM International Components for Unicode (ICU): A software library providing support for internationalizing software using Unicode.

Indian Standard Code for Information Interchange (ISCII): Eight bit character encoding standard for Indic languages developed by Government of India.

Intelligent Character Recognition (ICR): See OCR.

Input Method Engine (IME): Software used along with keyboard to generate the required sequence of codes based on the keys typed by the user. Input method engine is widely for CJK.

Inscript: A keyboard standard for Indic languages developed by.

ISO 639: A standard that defines 2 or 3 character language codes for all languages.

ISO 9995: Keyboards are defined by the ISO 9995 standard.

ISO IEC 2022: A standard mechanism devised to include multiple encodings of variable size in a single encoding scheme

ISO IEC 8859: A series of 8-bit character encoding standards for different languages, e.g. 8859-1 (Latin), 8859-6 (Arabic).

ISO IEC 10646: See Unicode standard.

ISO IEC 14651: A standard that provides a method for sorting multilingual data satisfying their lexicographic needs. It provides a general purpose ordering template to cater to different cultural requirements.

ISO IEC 646: See ASCII

J

Jamo: The 51 letters that compose Korean language are known as Jamo or Natsori. 24 of these are atomic characters (14 consonants + 10 vowels) while other 27 are formations combining 2 or sometimes 3 characters from the first 24 characters.

K

Kana: A term used to refer to the syllable based Japanese writing systems Hiragana and Katakana.

Kanji: Japanese initially borrowed some of the Han characters and incorporated into their language. These characters are known as Kanji in Japanese language.

Katakana: Katakana is a syllable based Japanese writing system used mainly to write foreign words/names or for putting prominence as is done with italics in Latin.

KDE: KDE is a Linux based multilingual desktop suite that provides word processing facilities, office suites, browsing and e-mail clients and development tools.

K-Edit: A KDE based text editor.

Knoppix: A Debian based live CD having KDE as standard desktop. It boots and runs from CD, can be customized to add/remove packages or to localize it through a procedure known as re-mastering.

Konqueror: A KDE based web browser and file manager.

L

LaoNux: A KDE based live distribution localized in Lao.

LC_COLLATE: A part of locale in GLIBC that defines collation of strings.

LC_MONETARY: A part of locale in GLIBC that defines the monetary conventions.

Line Breaking: An algorithmic solution to determine end of line place for writing systems which do not have space between words, e.g. Thai, Lao, Khmer, Chinese. This is required for word processing.

Language Interface Pack (LIP): Language Interface Pack provides localized environment such as menus, dialogue boxes and pop up messages for Microsoft platform.

Live CD: A CD that boots operating system without needing installation on the hard disk.

Locale: A set of formats and word translations for cultural information (date, time, days of week/months, monetary values and collation etc.) in local language which define basic interface requirements for that language as spoken in a particular country.

M

Machine Translation (MT): Process of using a computer program to translate text from one language to another.

Microsoft Keyboard Layout Creator (MSKLC): It is a multilingual KB layout creator for Microsoft platform.

Morphix: A Debian based live CD. It can be customized by adding/removing packages and localized through a procedure called as morphing.

Morphological Analyzer: A computer application which can analyze a word into its morphological structure or vice versa.

Mozilla: An open source multilingual browser that can be used on multiple platforms. Mozilla suite comes up with Mozilla browser and e-mail client.

Multiple User Interface (MUI): A software package that enables installation and use of multiple language interfaces on Microsoft platform, as compared to LIP that adds one language at a time.

N

Nautilus: A GNOME based file manager.

National Language Support (NLS) File: It is an alternative name used for locale information.

O

Optical Character Recognition (OCR): software that recognizes and extracts type-written text from a scanned image and creates a corresponding text file.

Open Office: An open source multilingual office suite that can be used on Microsoft and Linux platforms.

Open Source Software (OSS): Software distributed with the source code.

Open Type Fonts (OTF): An open standard for developing intelligent fonts. The standard is developed by Microsoft and Adobe, and contains glyph outlines in TTF format and additional rules to select and join these glyphs.

P

Pango: A library that is concerned with rendering of text in GNOME.

PO Files: These files provide translations of standard strings for Linux platform. These are compiled to become MO (Machine Objects) files which are read by Operating System to display the translated strings for a particular locale instead of English strings.

Part of Speech (POS): Syntactic category of a word, e.g. noun, verb, etc.

Part of Speech (POS) Tagger: A software which guesses and tags the POS for words in input text.

Positioning: Process in OTF which use rules to determine glyph position in context.

Python: An object-oriented programming language used mostly on Unix/Linux based platform.

Q

QT: Basic engine behind KDE environment on Linux.

R

Rendering: A process which involves reading font file to generate the display on the screen against an input text.

Reordering: Process of swapping characters for proper display, required for scripts in which typing order differs from visual placement order of some characters, e.g. in Devanagari and Thai.

S

Simplified Chinese: A new approach of writing Chinese introduced in 1949. It simplified the traditional Chinese and reduced the character set into fewer and simpler glyphs to improve the literacy.

Substitution: Process in OTF which use rules to choose one of many glyph shapes depending on the context.

T

TCVN: First national character encoding standard for Vietnamese.

TLE: Thai Language Environment.

Traditional Chinese: Chinese written with Han characters.

Thunderbird: An open source Mozilla based e-mail client that can be used on multiple platforms.

True Type Font (TTF): A font format which stores glyph shapes in forms of outlines (or splines) along with basic character position.

Text-to-Speech (TTS) System: A computer software application which generates speech against an input text for a language.

U

Unicode: An international standard used to encode all scripts of the world. Also known as ISO IEC 10646. Most languages are encoded in a space of 16 bits.

Uniscribe: A rendering engine for OT fonts developed by Microsoft.

Urdu Zabta Takhti (UZT): A national 8-bitstandard for Urdu character encoding in Pakistan.

UTF-8: 8-bit format to encode Unicode text.

UTF-16: 16-bit format to encode Unicode text.

V

Vietnamese Standard Code Set for Information Interchange (VSCII): An 8–bit encoding for Vietnamese.

W

WordNet: An extensive network of words and their senses, with information on how words are related to each other in various relationships, e.g. synonymy.

X

X Input Method (XIM): An input method for Xwindows for complex language systems e.g. CJK.

XVNKB: A keyboard for Vietnamese on Xwindows.

Xwindows: A window based wrapper on Unix and Linux operating systems providing a graphical user interface.