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.

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].

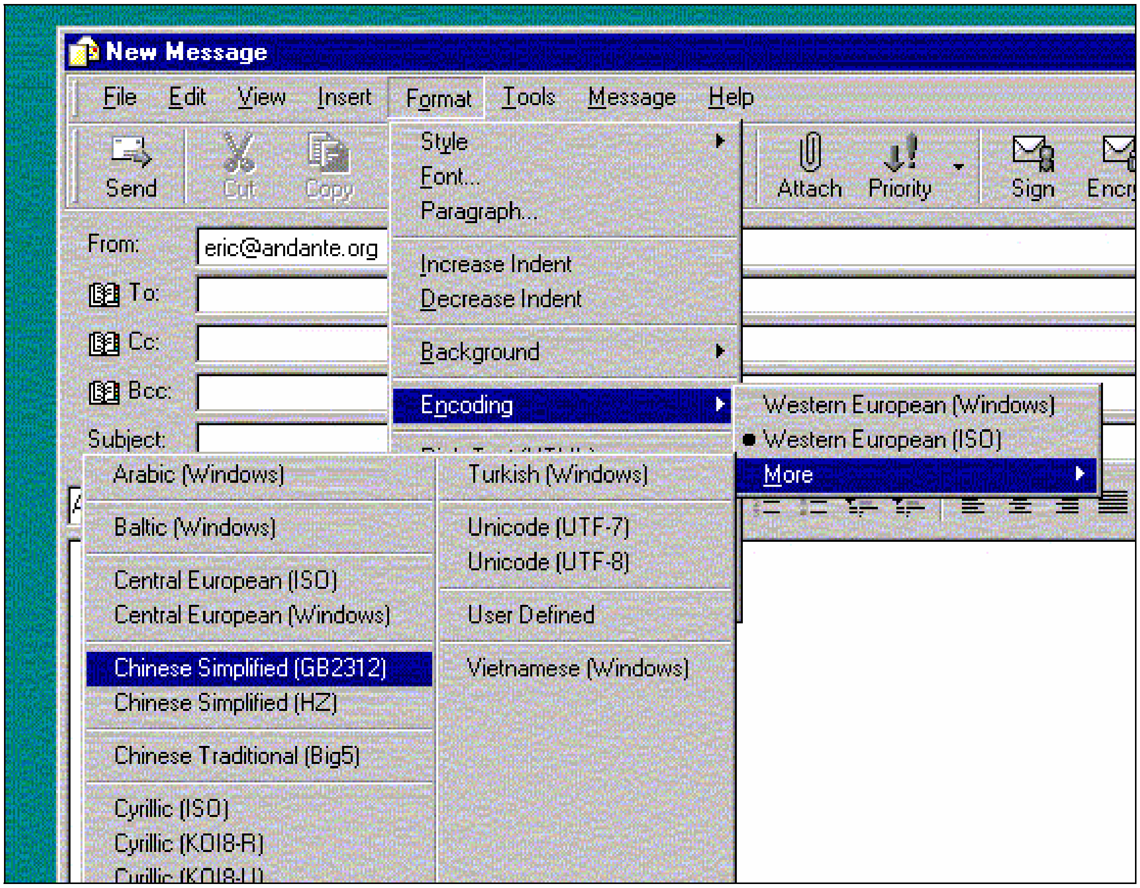
Character Set and Encoding

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.

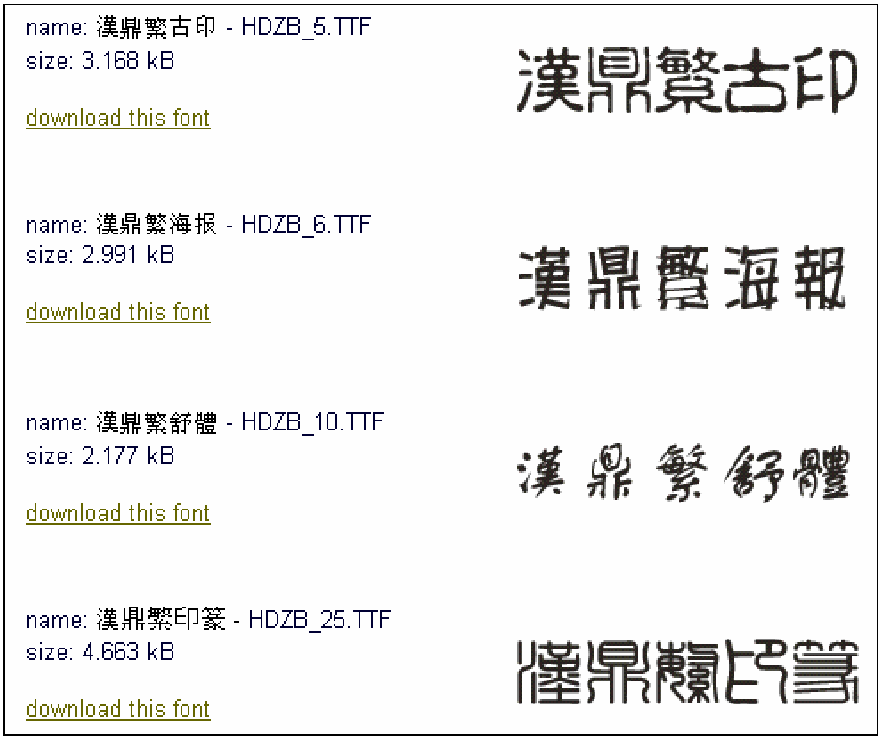


From: eric@andante.org

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.

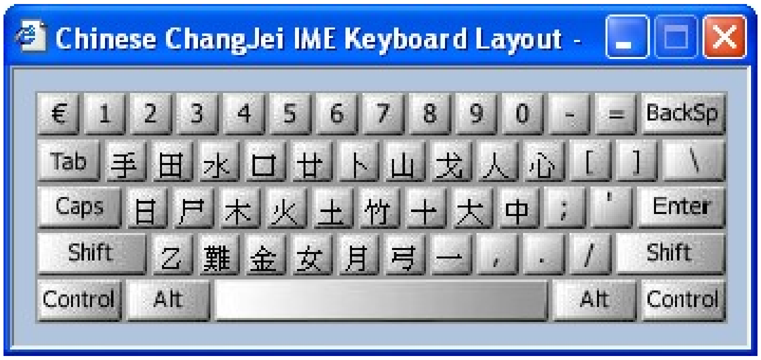

Chinese ChangJei IME Keyboard Layout -
€ 1 2 3 4 5 6 7 8 9 0 - = BackSp
Tab 手 田 水 口 廿 卜 山 戈 人 心 [ ] \
Caps 日 尸 木 火 土 竹 十 大 中 ; ' Enter
Shift 乙 難 金 女 月 弓 一 , . / Shift
Control Alt Alt Control

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].

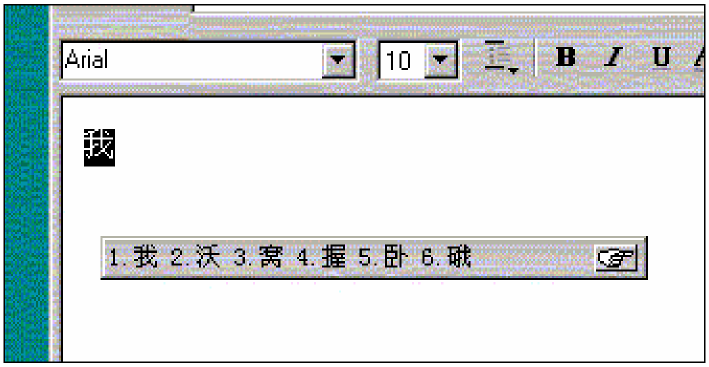
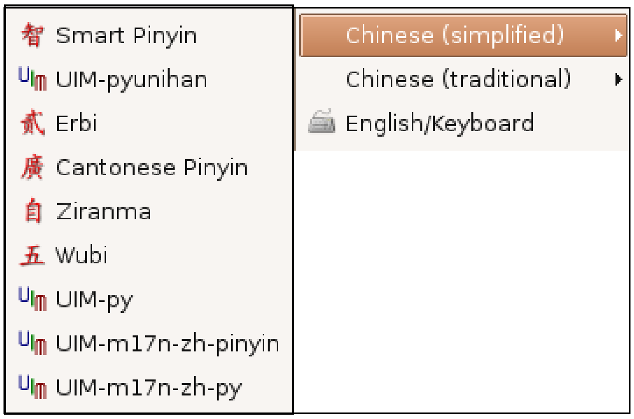


Figure 5: IME Pad for Chinese Character Input [15]

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].



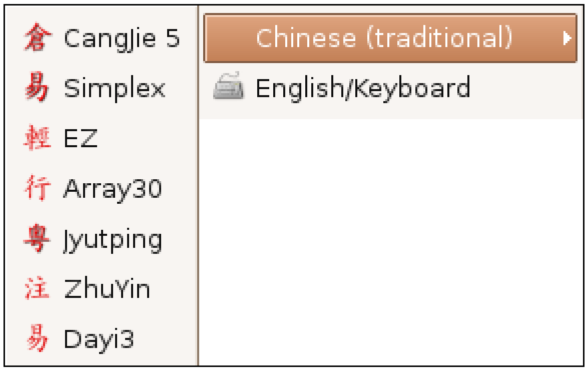


Figure 6: IME for Ubuntu CJK (Linux) for Simplified and Traditional Chinese [23]

Collation

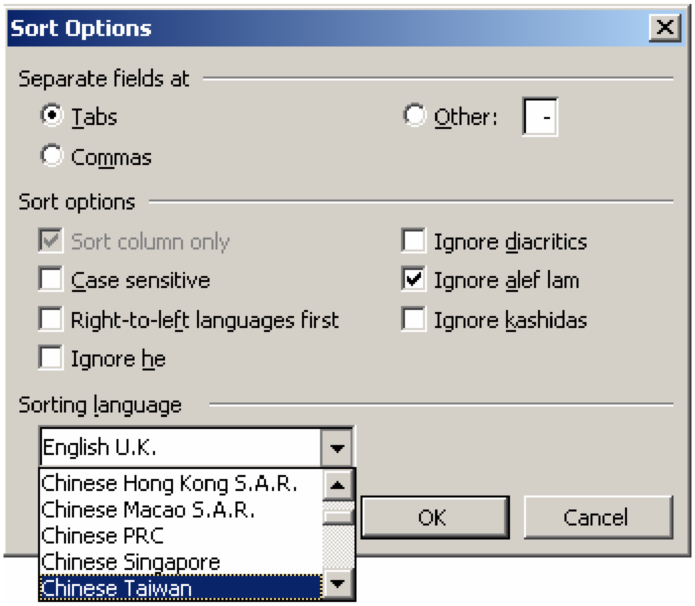
Like input methods, collation may also be done in a variety of ways, including phonetic, alphabetic (e.g. based on Latin input), stroke based or dictionary based, for example, Chinese Big5 order, PRC Chinese Phonetic order, Chinese Unicode order, PRC Chinese Stroke Count order and Traditional Chinese 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.



Separate fields at

● Tabs ○ Other: -

○ Commas

Sort options

☑ Sort column only □ Ignore diacritics

□ Case sensitive ☑ Ignore alef lam

□ Right-to-left languages first □ Ignore kashidas

□ Ignore he

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

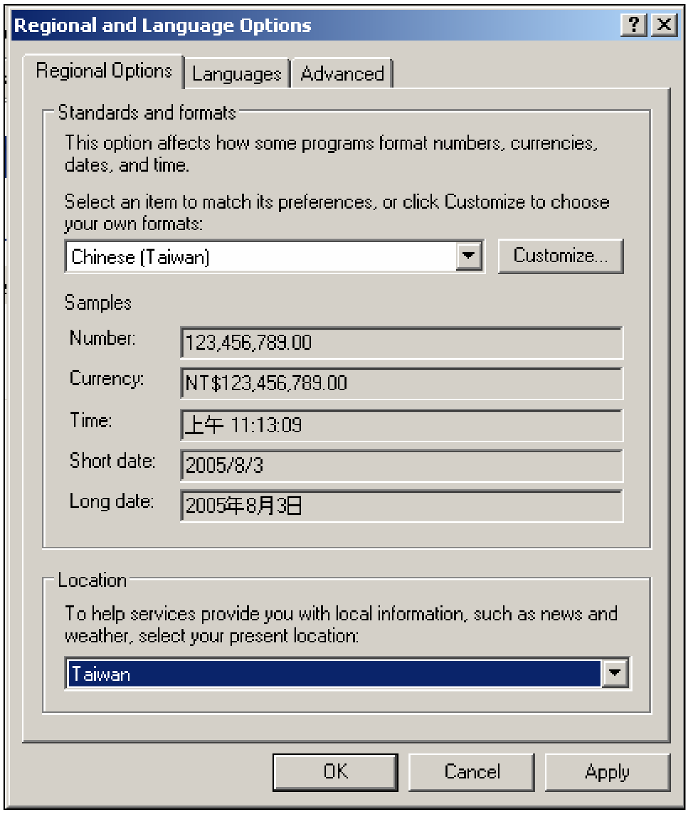
Locales for both Traditional and Simplified Chinese have been defined in IBM ICU [27]. Locales for Chinese are also available in the CLDR1.3. The locales are zh_CN (for China), zh_TW (for Taiwan), zh_HK (for Hong Kong) and zh_SP (for Singapore) (e.g. see [26, 30]).

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

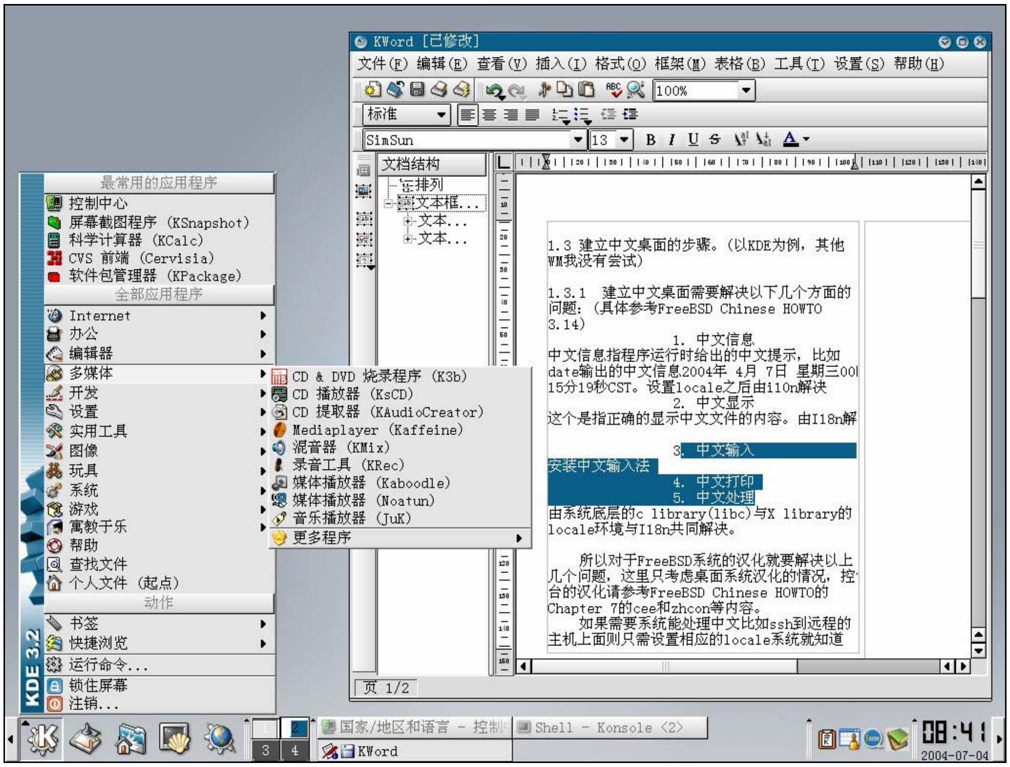
Interface terminology is available for Chinese on multiple platforms.

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

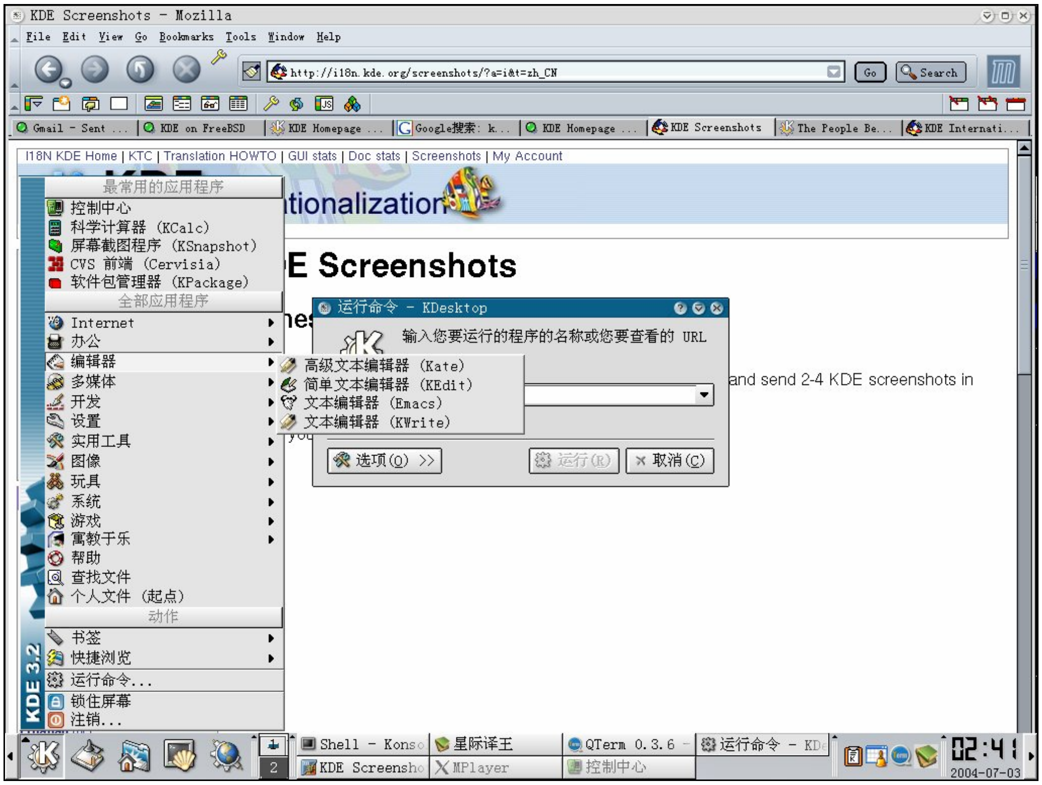
5. 中文处理

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

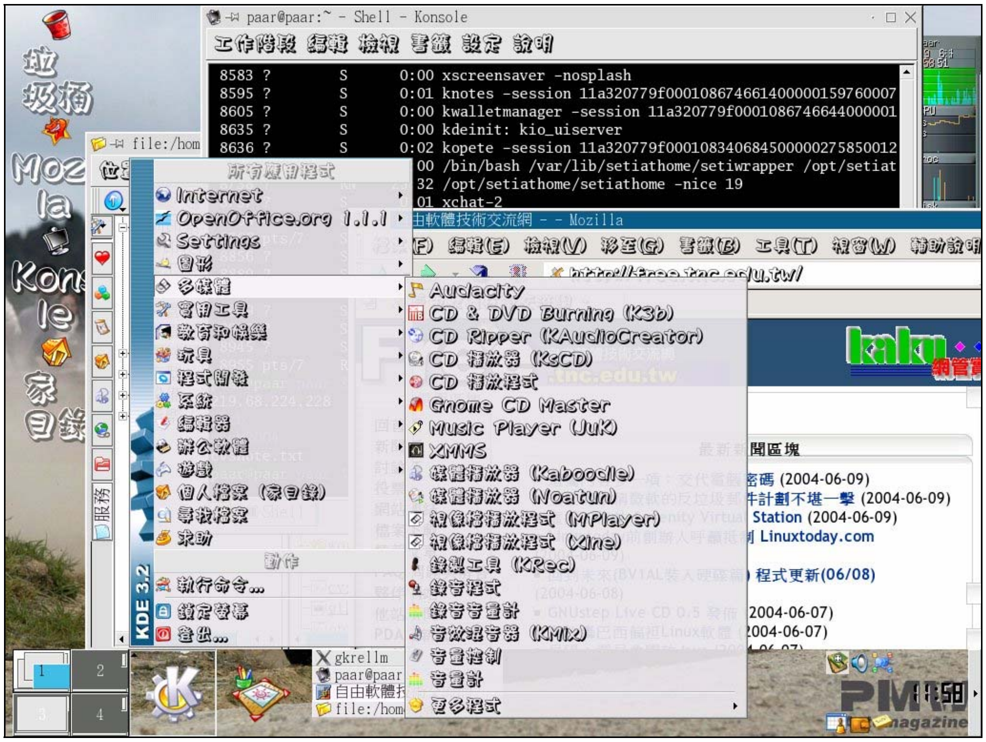
Figure 9: KDE K-Word for Simplified Chinese [36]



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 11a320779f00010867466440000001
 8635 ?        S      0:00 kdeinit: kio_uiserver
 8636 ?        S      0:02 kopete -session 11a320779f000108340684500000275850012
                        00 /bin/bash /var/lib/setiathome/setiwrapper /opt/setiat
                        32 /opt/setiathome/setiathome -nice 19
                        01 xchat-2
```

Figure 11: KDE Traditional Chinese [36]

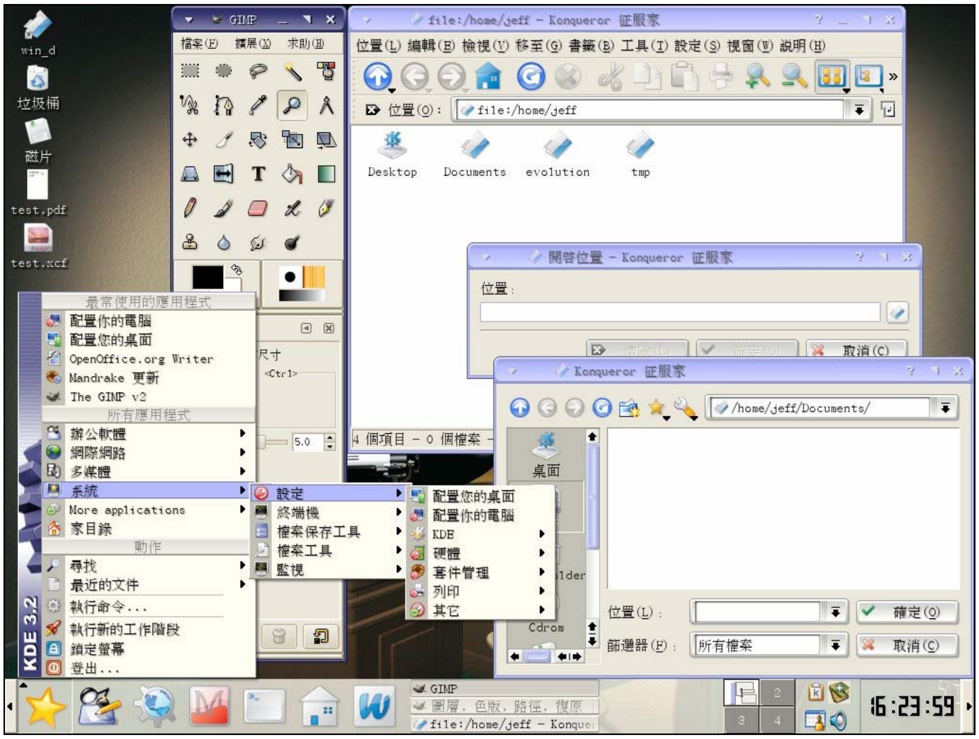


Figure 12: Simplified Chinese Mandrake [36]

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/