Survey of Language Computing in Asia 2005

Sarmad Hussain
Nadir Durrani
Sana Gul

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

www.nu.edu.pk

www.idrc.ca

Published by

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

Printed by Walayatsons, Pakistan

ISBN: 969-8961-00-3

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

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].

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.

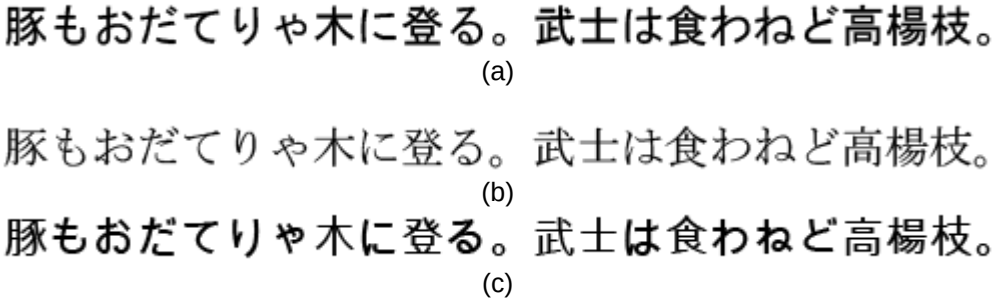

Figure 2: (a) MS Gothic, (b) MS Mincho and (c) Arial Unicode [5]

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.

Figure 3: Hiragana/Katakana Keyboard Layout [9]

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.

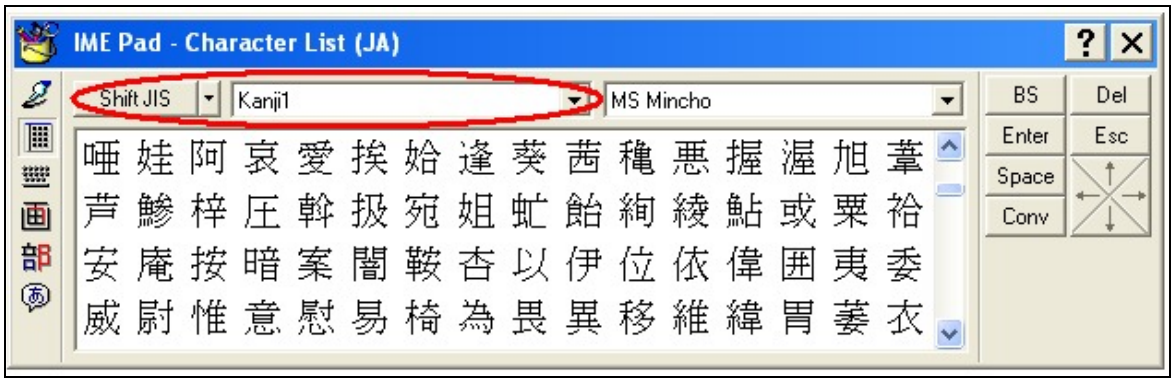

Figure 6: JIS Based Character List [9]

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.

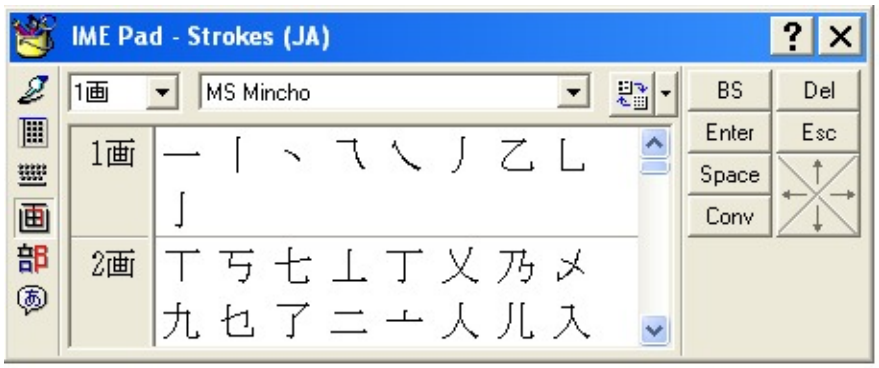

Figure 8: Strokes Utility in IME Pad [9]

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.

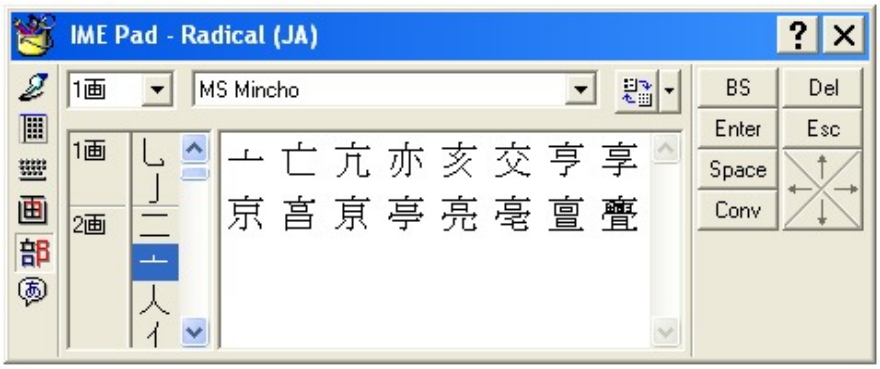

Figure 9: Radical Utility in IME Pad [9]

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.

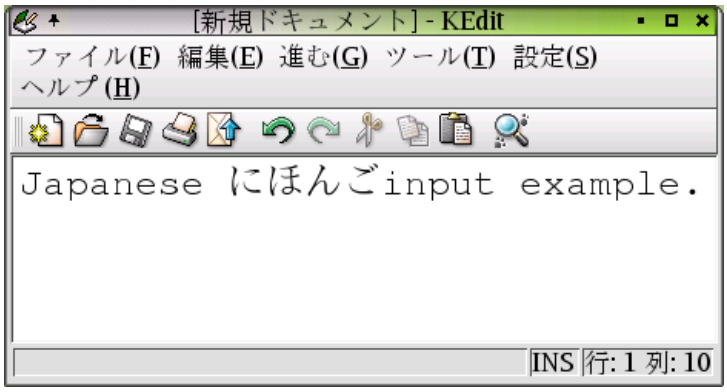

Figure 10: X-Input Method in K-Edit [10]

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]).

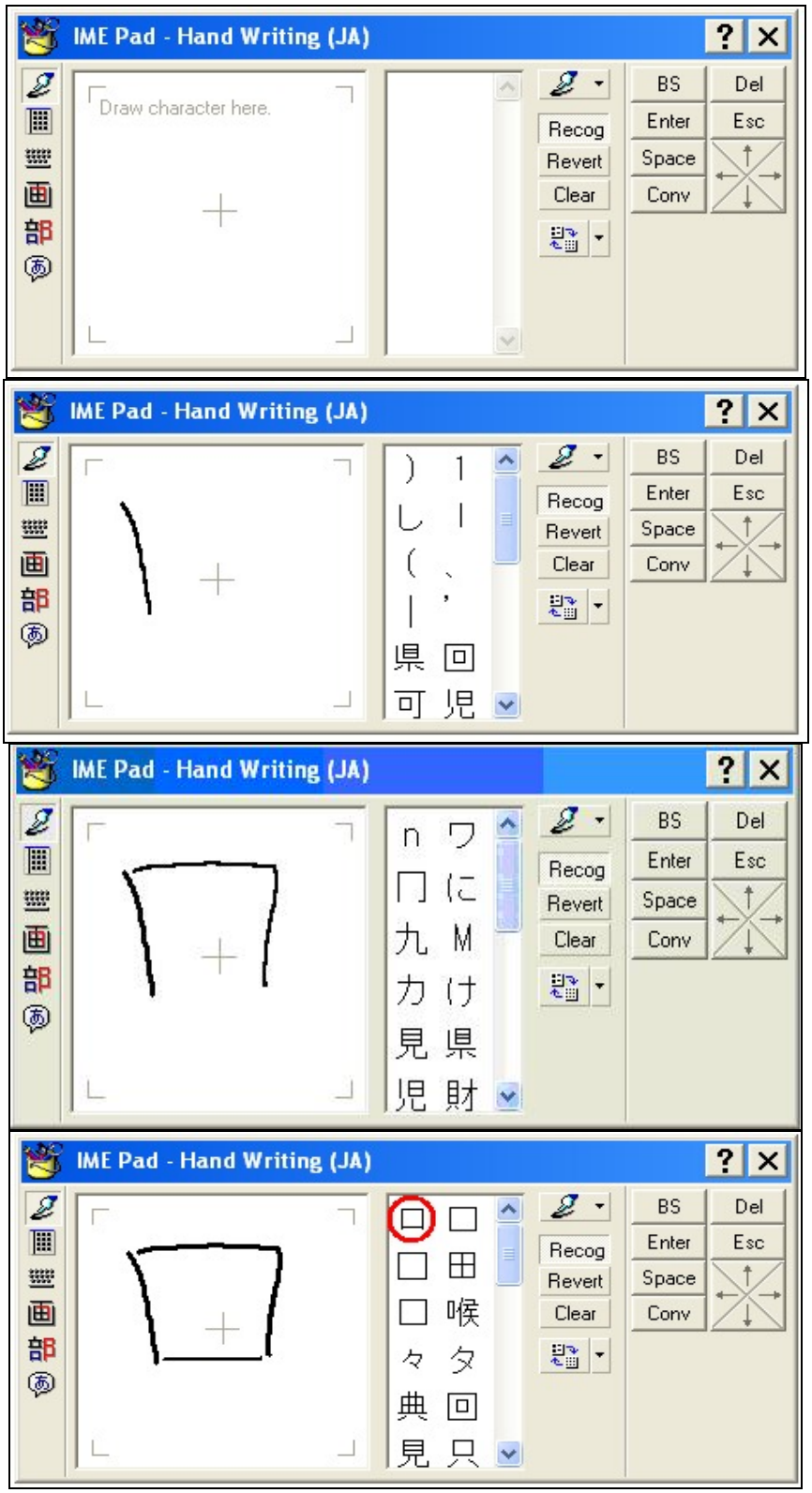

Figure 11: Handwriting IME for Kanji Characters [9]

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