Multilingual Processing in Asia: Past, Present, and Future

Koichi Kato

Advent of Electronic "Paper"

Much has been said about the realization of a "paperless society" with the advent of computers. In reality, computerization has only produced an ironic result, an increase in paper documents. This is because CRT (cathode-ray tube) and LCD (liquid crystal displays) are sufficient for browsing a document, but are not suitable for a careful reading of large amounts of text. A new technology under development, "electronic ink," may solve this shortcoming of currently available displays. Used with its exclusive display, electronic ink uses only a small amount of energy to re-write the display, and it is retained after the power is cut off. It is easy on the eye because it does not emit light as CRT and LCDs do, but reflects light like paper. It has only a few manufacturing steps, and is hoped that the cost will decrease with mass production.

The first product to use electronic ink was introduced in 1999 by E Ink Corp. (http://www.eink.com). It can only display large characters of 2.75 inch square, yet it has a thickness of 3 mm and can be rolled up like paper. Many companies are now competing to develop electronic ink technology, and it is said that by the year 2005, there will be a product that will replace paper.

Electronic ink is a boon not only for paper-wasting industrialized nations, but also for educational scenes in developing countries. It may be difficult to deliver new textbooks to schools in remote or mountainous areas, but with electronic ink, one only needs one PC and a cellular phone to receive the latest learning materials in the classroom.

Digitizing a Document

To digitize a document, there are 2 methods: image and character. The image method treats a document as an aggregate of light or dark dots, as in a fax machine. The character method treats a document as a collection of numbers by replacing each character with a number. The number representing each character is called "coded character," and the assignment of a number to each character is called "coded character set." Operations like deletion, addition, duplication, paste, search/replacement of characters in a computer are called "editing." It is impossible to edit a document with the image method, while it is possible with the character method. With it, one can even locate a single document among a great number available on the Internet (computer search).

In most of Asian countries, text-processing technology of their mother tongues began and developed for Computer Typesetting System (CTS). We took interest in printing quality, and not in inputting and editing process, which was so complicated that remained in the typesetters' sphere. The personal computers (PCs) make text-processing work familiar to common people, who now publish a lot of essays, diaries and stories on the Internet. Text is going to fly over paper media; however, there are many languages in which text cannot be displayed properly or be easily edited. In addition, most systems can handle 2 languages, such as English and Japanese, but not multiple languages. Problems remain in international communications.

Various Kinds of Characters

The computer developed in the United States. Modern English has a remarkably simple script, with a set of capital and small letters, numbers, and punctuation marks. All English words can be written with a combination of about 70 characters. The most basic coded character set is the ASCII (American National Standard Code for Information Interchange), on which coded character sets of various countries are based.

The alphabet is a phonemic system, using a combination of consonants and vowels to express words. Alphabetic languages other than English use diacritical marks to express different sounds with the same character, e.g. 'è' or 'ç' in French. In the beginning of coded character set development, a character with a diacritic such as ' è ' was encoded as a combination of two characters, e.g. 'e' and ' ` '. At present, however, such characters are given a separate code number. 10 to 30 such code numbers in addition to ASCII are sufficient for European languages, but nearly 200 additions are necessary in Viet Nam's Quoc Ngu.

There are a great number of Chinese characters (Hanji in Chinese, Hanja in Korean, Chu Han in Vietnamese and Kanji in Japanese) used in China, Taiwan, Korea, Viet Nam and Japan. About 3000 characters are needed in daily life. A common character dictionary contains 10,000, and the world's largest character dictionary contains 80,000 characters.

Technical restrictions limit the code to a number between 0 and 127 expressed in 7-digit binary code (called 7-bit). With Chinese characters, there are so many that a code consisting of a combination of 2-code numbers, called 'multiple-byte code' is assigned.

When there is a need to express a character which a coded character set does not contain, the user may define the character with an unused code number in the coded character set. Such a character is called a "user-defined character" or an "external character". It can be printed without difficulties, but in a remote computer on a network such as Internet, it appears as a different character from the original.

One reason for the complexity and large number of Chinese characters is that in many cases, one word is expressed by one character, e.g. 山 for "mountain" or 川 for "river." Some characters have changed slightly over the course of history, developing into a different character, just as English spelling has changed. Stephenson, the inventor of the steam

engine, and Stevenson, the author, are different words, even though they have the same pronunciation. In Japanese, there are several ways of writing "Takasaki": 高崎 and 高﨑; however, this may be taken to be only a stylistic difference. The problem of where to draw the line between 2 different characters and the same character written in 2 different styles will persist for some time.

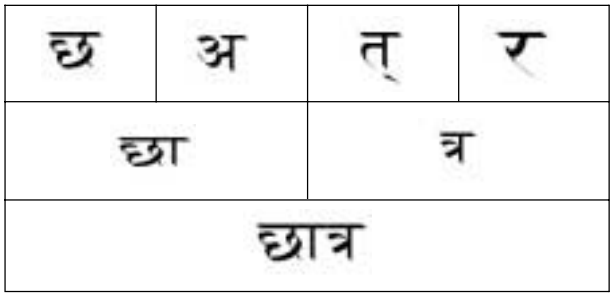
Arabic is mostly written with consonants only due to the nature of the language. In the Arabic script, the shape of a character changes with the position within a word. The character ب for the "b" sound is بـ at the beginning, ـبـ in the middle, and ـب at the end of a word. There is also a phenomenon called "ligature" — ل for "l" and ا for "a" together are written as لا; as in the case of German, "ß" is written for the combination "ss."

If these variations of presentation forms are treated as distinct characters, it will be difficult to perform editing tasks, like delete, add, and search. At present, the recommended solution is to use separate coded character sets for editing and display. For editing purposes, an independent character, or one in word-initial, word-internal, or word-final position, is represented by the same symbol. لا is treated as 2 separate characters, ل and ا for editing, but is replaced with one character for display.

Arabic script also is used in Indo-European languages such as Persian, Urdu, and Altaic languages such as Uighur. These languages have their own method of writing consonants and vowels, and are considered to have a script different from Arabic.

The world's first alphabet was developed by people speaking a Semitic language. Semitic languages can be written with consonants only; therefore, no symbols for vowels were invented. When this alphabet was transmitted westward, characters for vowels were added in Greece and became the basis of the modern alphabet. In India, marks representing vowels were added to the consonant characters to express syllables—this is called a "combinatory syllabic system." Sanskrit, Devanagari, Thai, Myanmar, Tibetan scripts, and the Hangul script of Korea are all combinatory syllabic systems.

As in the case of consonants, vowels may have a different shape according to whether they are used independently or combined with consonants to express syllables. They may also form a ligature. In Devanagari script, क for "k" and ष for "sa" turn to क्ष. Hangul, invented in the 15th century, is a relatively new combinatory syllabic system with a simple arrangement of sounds. Even so, theoretically possible combinations of all characters number over 10,000. At the request of Korea, Unicode=ISO 10646 contains all possible combinations of Hangul characters. In other combinatory syllabic scripts, the numbers of possible combinations reach astronomical proportions, and all cannot be realistically contained. These languages must be expressed by a set of consonant-vowel combinations, but strict definitions on what combinations are considered one character would be necessary for editing task.



छात्र is a word for "student" in Devanagari script.
त्र is a ligature for त् and र.

ISO 2022 and ISO 10646

The official coded character set of each country is based on the method called ISO 2022 defined by the International Standard Organization (ISO). In the Opening Ceremony of the Olympics, athletes of each country march in with a placard displaying the name of their country. Similarly in ISO 2022, a distinguishing mark is placed at the beginning of strings of characters, showing in which coded character set the text was encoded. ISO 2022 was established in 1973. At the strong request of Hiroshi Wada of Japan, the only country with non-alphabetic country present, an allowance was made for multiple-byte expansion, which made possible the systematic development of character codes for Chinese characters.

ISO 2022-based coded character sets are developed for information interchanges, so they are not suitable for internal processing of computers. Until the wide popularization of the Internet, only bilingual processing—English and one other language—was needed. In Japan, internal processing codes such as Shift JIS and Extended UNIX Code (EUC) were developed by modifying the JIS X 0208, the national coded character set of Japan based on ISO 2022. Shift JIS is common in PCs and EUC in UNIX workstations. In China and Korea, PCs use their national coded character set with EUC encoding; in Taiwan and Hong Kong, Big 5, a coded character set influenced by Shift JIS, is common.

When sending electronic mail and data to remote computers, it is necessary to convert the internal processing code into an ISO 2022-based official coded character set; however, much of the text in the internal processing code had been sent out without conversion. This began to be a problem with the wide use of the World Wide Web. Without any distinguishing marks at the beginning of text, computer systems were not able to specify which coded character set was used.

In 1983, ISO began discussion on an international coded character set to follow ISO 2022. At the same time, big computer vendors in the United States, with a view to globalization of the economy, began to develop a 16-bit universal coded character set called Unicode to reduce the localization costs of computer. At the strong urging of the Unicode Consortium (http://unicode.org/), ISO agreed to jointly develop an international coded character set. The coded character sets completed in 1993 by the organizations were essentially the same, unifying various official character code sets of different countries at the year 1990. It is not compatible with ISO 2022-based coded character sets. This 16-bit

code can contain up to 2^{16} characters, or over 65,000 characters. It was later found that this was not enough to define all the characters in the world, and 10,000 or so code numbers were added, using the surrogate pair method. At first, Chinese characters numbered 20,000, and have been increased to 60,000 until now.

Towards the Informed Network Society

In many countries that use Arabic script, or those languages with combined syllabic system, several incompatible coded character sets are used concurrently. This situation may be due to administrative confusion or development of independent coded character sets by overseas aid agencies. It is partly caused by the nature of the scripts themselves. In many such ones, a shape of consonant/vowel combination or ligature is not a simple composite of the 2 original characters. Most designers of those coded character sets introduced easily additions for presentation forms of composite characters in order to augment the PC's poor rendering capability; these solutions are sufficient for display but make editing tasks more difficult. Even with the same basic character set of consonants and vowels, people may disagree on which presentation forms are added as separate characters or how character components are defined.

CICC (Center of International Cooperation for Computerization), a Japanese institute (http://www.cicc.or.jp/) began in 1987 a project called AFSIT (Asian Forum for Standardization of Information Technology) upon request of the Japanese government to cooperate in advancement of information technology in Asian countries. As a part of this project, it has organized a conference called Multilingual Information Technology (MLIT) since 1997 with various Asian countries. On the theme of Equal Language Opportunity and Multilingual Processing Environment, it has provided a forum of discussion on many issues with many fruitful results. Some issues discussed were the new text-processing methods for separating editing/display characters, coded character set architecture, easy-to-use input system for common people, and how to register characters and scripts with ISO. CICC opened the technical seminars called SEISA AP/IT for the countries holding serious problems in text-processing technology in 1999 and 2000. We can read all reports and thesis of symposiums in MLIT Web site (http://www.cicc.or.jp/homepage/mlit/) and find what problems Asian countries are tackling.

In countries that use Chinese characters (China, Japan, Korea, etc.), user-defined characters are the obstacles to smooth exchange of information. To combat this problem, there are now attempts to enable use of user-defined characters on networks by establishing a common character library of several tens of thousands of characters. The most widely used of these is that of *Mojikyo* Institute (http://www.mojikyo.org/). Due to international cooperation, this character library contains ancient Chinese characters, seal characters, Chu Nom characters of Viet Nam, and pseudo Chinese characters invented around China in addition to the 80,000 traditional Chinese characters. A search engine based on origin of Chinese characters has been produced by AI-Net and is distributed by Kinokuniya Company Ltd. (http://www.kinokuniya.co.jp/). *Mojikyo* Institute has the rights to distribute, free of charge, the outline fonts and the light version of the searching and inputting tools for the next 100 years from AI-Net (until Oct. 2098). The fonts and the light version of the tools can also be downloaded from *Mojikyo* Institute's website. The *Mojikyo* Institute and AI-Net also offer character images in GIF format on Internet. Inserting the URL of the target user-defined characters' image into HTML sources will enable it to be displayed on the www pages.

ISO 10646 is thought to become the leading coded character set, but at present, there is still a lack of software that uses it. Most computerized data in the world is encoded with an ISO 2022-based coded character set. With some www browsers, it has become possible to display a character in ISO 10646 within text encoded in ISO 2022 by the method of numeric reference, but editing is still a difficulty. There are still many complications to overcome before a simple and easy-to-use multilingual processing system is widely available.

(translated by Kaori Ueki)

Koichi Kato
Born in 1954. After graduating from Waseda University where he majored literature, his main interest has been two Japanese novelists, Jun Ishikawa and Kobo Abe. He has written various articles and books on literature and language. Since 1995 he has been providing a forum of discussion mainly on literature on his website 'Horagai' (http://www.horagai.com/) where he presents his research works and exchanges opinions including issues about coded character set and digitized text.
Koichi Kato
Freelance writer/literary critic, e-mail: horagai@super.win.ne.jp

The *Mojikyo* Database is constructed with the etymological system called as Takuji. 高 means "highness". In ancient characters as carved on bones, 高 is  , which represents a triumphal arch () and a sacred box ( ). 高 symbolizes a prayer for the heavenly gods in the front of a triumphal gate.