[From: *Translating and the Computer 14*. Papers presented at a conference… 10-11 November 1992 (London: Aslib 1992)]

The reunion of an Eastern family - The worlds first truly integrated Chinese, Japanese and Korean DTP system

H Z Zhang, Eastword

Eastword (formerly Chinatech) has been the trade supplier of Oriental translation and typesetting for a number of years. The changeover to new technology posed serious problems for us. The only route seemed to be dedicated systems for each of the languages we specialised in. Our original intention of four years ago was to purchase a more or less off the shelf system but we were frustrated by the lack of knowledge of various consultants and suppliers who initially claimed to be able to satisfy our needs. So we set out three years ago on a long and lonely road to assemble a system that would enable us to input and output all our inhouse languages from one single integrated system.

THE THREE BROTHERS

The three main languages in question - Chinese, Japanese and Korean - could be said to be related through "marriage". Although Japanese and Korean are related linguistically to each other, they have no linguistic connection with Chinese. However, as a result of historical connections with China they all use the non alphabetic Chinese characters. These characters were adopted as written languages centuries ago and the pronunciation of many of the characters in Korean and Japanese still gives a clue to their Chinese origin. In times past it was entirely possible for Koreans to communicate with Japanese using Chinese characters even though they shared no common language. The great classic Chinese writings were an essential part of a scholar's education in Japan and Korea as well as in China. The cultures of these three countries shared many aspects but they also developed separately and this is reflected in the written languages.

Chinese characters are composed of a number of elements which are combined within a square shaped parameter to form each unique character. Most characters comprise an element which gives a clue to meaning and some have an element which gives a clue to the pronunciation, but the reading of each character must be learned individually. Each character in modern Chinese represents a syllable and two or more characters are usually linked to form words. Each homophone in the spoken language has a number of different characters with their own unique meanings that are all pronounced exactly the same. It is only in the combination of characters into word groups and the distinctiveness of each written character that the ambiguity of the spoken language is clarified. Written Chinese is therefore not an alphabetic language and various methods are used to overcome the problems of inputting these characters using the conventional qwerty keyboard.

In addition to a large base of Chinese characters, Korean and Japanese now have their own forms of alphabetic scripts which are used to transcribe the language purely as it is pronounced. Japanese has two alphabetic scripts; Hiragana for Japanese words and Katakana for borrowed Western words. Input of the various alphabetic scripts is relatively simple, this is based purely on pronunciation of the "letters" and a normal qwerty keyboard can be adapted for input. However the Chinese characters are still an integral part of both written languages and various strategies therefore need to be adopted for the input of these characters. This is particularly true for Japanese which continues to use the Chinese characters or Kanji for a high proportion of written words. A greater use of Kanji tends to indicate a higher level of education although Hiragana can be used for all Japanese words. It is the conversion of Hiragana into the appropriate Kanji which is the crucial factor in inputting Japanese.

The Korean "characters" - which were developed in the 15th century - resemble Chinese characters in the sense that each syllable occupies an equally sized square shape but the elements that compose the "characters" are in fact purely phonetic. The education policy in Korea has fluctuated in recent years with some generations of school children being taught Korean "characters" exclusively. These generations are therefore unable to read written material produced using Chinese characters. The more scholarly works are written with a higher proportion of Chinese characters as these convey a more precise meaning, thus eliminating the ambiguity of the many homophones in the language. Again it is crucial that the input system enables selection of the correct Chinese character.

These three languages with their various "alphabets" and characters sets present unique problems with regard to computerised typesetting. If the three languages had continued their joint development our problems would be greatly reduced. One problem is that the Chinese character sets used by each of these languages have continued to develop separately and many characters are now quite different from the original Chinese version. This means that a separate character set is required for each language and the space required to accommodate these is colossal. When one considers that up to three thousand Chinese characters are used in everyday Japanese it can be seen that a Japanese font, including the two Japanese alphabets, comprises considerably more characters than a typical European font.

Input of the various different scripts is achieved in a number of ways according to the requirements of each language and to achieve acceptable input speeds sophisticated artificial intelligence must be built in to the systems. This means that each language must run through or on top of its own system software, i.e. Chinesetalk, Kanjitalk (for Japanese) and Hangultalk (for Korean). A system switcher is used to alternate between language systems. The "talks" have been developed separately to address the respective characteristics of each language and so they each have their own distinctive methods. In other words they are not interchangeable and training in each system is required. There are now systems under development that will eventually mean it is unnecessary to switch to a new system for a new language - only the keyboard will need to be changed - but until that day we must operate several different systems on one machine.

ACHIEVEMENTS TO DATE

We now have a system that can produce postscript typesetting in Chinese (5 fonts), Japanese (5 fonts), Korean (11 fonts), as well as Thai, Cambodian and Vietnamese. Each time we have added the fonts for a new language to our repertoire there has been much celebration. We started with Japanese, next came Korean and finally Chinese.

1) Hardware

The system we finally arrived at after much research is built from a series of Apple Macintosh machines networked to a Hyphen software RIP which is installed on a llfx. This powerful platform has all the Oriental fonts downloaded and permanently available to the input devices. The RIP is in turn connected to a 600 dpi laser printer for high quality proofing and a 2400 dpi image setter. The film or bromide is then developed in the daylight processor.

2) Software

This is more problematic; after three years we are still not able to do all that the systems for phonetic languages are capable of. We are 60% of the way there, but we are restricted by the lack of interest of software developers and the disadvantage of having no precedents to learn from. The major DTP softwares have been developed for Oriental languages but a different version of the software is needed for each language. The expense and memory needed to run these different versions is prohibitive. We have instead opted for a multilingual software that can handle all the different input methods for the languages we use.

3) Early problems

In the first six months we encountered problems of all kinds virtually every day. Since all the software we are using is relatively untested it has been a difficult process to pin down the root of any of the

problems. By a slow process of elimination we have isolated most of the problems.
Some of the more interesting ones are as follows:

Overlapping of characters, Latin alphabet & numerals as a result of the clash between proportional and proportional and non-proportional spacing: Oriental characters are not proportionally spaced like the Latin alphabet, each character occupying exactly the same space. Consequently when the two types are mixed this can lead to problems. Although this has largely been ironed out by the font developers, some fine tuning is still required.

Characters disappearing when at the edge of the text box: Certain characters are very prone to disappear when they happen to be at the point where the text wraps onto the next line. When working in the Japanese system this is a persistent problem that requires careful proofreading whenever text is rearranged.

Wrong characters: The fonts we are using are very new and it will be a while before they are free of bugs. Since there are about 30,000 thousand Chinese characters it will take some time before all the anomalies are discovered. Several Chinese characters do not print out as they appear on the screen. In some cases the screen font is the wrong character, but more seriously sometimes, the printer font is incorrect. Potentially a very dangerous situation for any typesetter because What You Get Is Not What You See - our pet term for quite a few months - WYGINWYS.

Line spacing: the line spacing for Oriental characters in a given point size was different from that of the Latin characters in the same point size. This caused irregular line spacing when Latin text was input in the middle of a block of Oriental text. This was solved by an update of the software.

Inter-character spacing for some fonts was not as shown on the screen - naturally we blamed the font manufacturers, the page makeup software producers and anyone and anything that was connected to or with our system. We were given highly technical explanations for this but it turned out those fonts were "2 byte" fonts and it was merely a question of changing the printer setup to " byte Postscript printer".

Recent Problems:

The time taken to access the fonts was gradually increasing as we added more and more new fonts. The solution: a pre-loader was written into the initiation file so that the relevant fonts for each language could be accessed. This now involves preparing the RIP each time a different language system is used, which takes up to thirty minutes for each language but subsequently speeds up printing.

Our latest font set could not produce certain characters. The font manufacturer is going to update the tables as per our suggestions.

WHAT NEXT ?

The quest for a page makeup software capable not only of handling virtually any language, but also able to import and export to the market leaders for phonetic languages such as Quark Express, Pagemaker & Ventura with formatting commands intact.

To convert between "simplified" and "traditional" Chinese: the characters used in the People's Republic of China are a simplified form of the traditional full form characters still used in Taiwan and Hong Kong. At present we have to use two separate systems for these two types of written Chinese. There is as yet no conversion application available and a file input in one system becomes unintelligible when read in the other system.