TECHNICAL UPDATE

David Herren
Middlebury College

WORLDSCRIPT, UNICODE, AND NISUS

WorldScript is a relatively new piece of the Macintosh operating system that is particularly interesting to those who word process in foreign languages, and of even greater interest to those who work in the less commonly taught languages such as Arabic, Chinese, Japanese, and Russian. Unicode is a proposed "replacement" for ASCII. Finally, I'll take a look at a word processing program for the Macintosh that takes particular advantage of WorldScript.

WorldScript: What Is It? Part I: Background

Well, I really wish I could just tell you and be done with it, but it's not "just that simple," contrary to a well-known politician from Texas. Before I describe WorldScript, I want to cover why it was needed and some computing history.

Computers and Memory

The average computer today uses memory to store and work with information. This much is common knowledge. Remember, however, that a computer is really a stupid box and the only thing that it truly understands are zeros and ones. Thus, all of the information that a computer stores is recorded in binary code, something which looks like this: 01101011. Why is it this way? Well, because computers really don't even understand zeros and ones in the same way that we do. They only know about whether a tiny voltage is off or on, and humans understand this concept better in terms of zeros and ones. Now, if zeros and ones were all that we humans needed to keep track of, that would be great, but again, it's not "just that simple." At the very least, we need to keep track of the characters we use to write our languages. How, then, are characters stored other than as zeros and ones?

To store the characters or letters that we use in our languages, the computer needs to use a series of zeros and ones to represent the letters. Most current computers use a single byte to do this. What is a byte? Easy. It consists of eight bits. OK, while true, that still doesn't tell you anything. Our number system is base 10. In other words, we count from zero to nine in a "column" before

David Herren is Special Assistant in Technology to the Vice President for Languages and the Director of the Language Schools at Middlebury College in Middlebury, Vermont.

rolling over to the next highest unit of ten. A computer is base 2 (binary), so it counts from zero to 1 before rolling over, and a byte refers to the fact that the computer uses 8 columns of zeros or ones (8 bits), and can thus "roll over" 8 times.

So, if a computer uses 8 bits to record characters, like this: 01001010, how many possible characters can be represented in this system? That too is easy (grin): it's 2 to the 8th power, or 2^8 as mathematicians might write it. That means 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2, or, 256 possible characters. Thus, in a one-byte system for representing the characters of a language, we could have 256 different characters. This is quite enough for most western languages, but clearly inadequate for the thousands of characters needed for Japanese or Chinese. Once again, however, it's not "just that simple." Computers often used to only store characters using 4 bits of information resulting in only 128 possible characters. This is due to historical reasons including the fact that computers were largely developed in the United States and not China, and that memory used to be really expensive. One hundred twenty-eight characters were still adequate for English since we only have 26 lower case letters, 26 upper case letters, 10 digits, special characters like "<>/[]{}@#," and a handful of punctuation marks. The computer itself needed a few characters for computer stuff like carriage returns, tabs, line feeds, escapes, nulls, and several other really interesting things, totaling 32 items. Thus it was possible to fit all of this stuff in 128 "slots"—a beautiful system until someone decided to work in Russian, or Polish, or Spanish, or French, etc. These languages either have more characters in the alphabets, like Russian, or include those troublesome accented vowels which the computer must consider as separate characters. Enter the 8 bit system and suddenly we're back to the 256 character system where we began our discussion.

Now along the way, some standards had to be applied: 01001101 had to mean the same thing regardless of the brand of computer, or we'd really have problems transferring information. This standard is known as the ASCII (American Standard Code for Information Exchange) table, and almost everyone agreed upon its use. (Where IBM was when this standardization occurred and why they chose a different table for their mainframe computers, EBCDIC, instead of ASCII, is a mystery to me).

What's the Problem?

Well, clearly the problem is two-fold. The "A" in ASCII stands for American, but the Russians, the French and others also wanted computers that would work with their languages. By using a full byte to represent character sets, there were finally enough slots, but the problem of the many Eastern languages remained—256 characters doesn't even come close to being sufficient for Japanese or Chinese. Apple Computer made foreign character sets a central focus of development—the original Macintosh could word process in any of the common western languages right out of the box, and could print those character sets without jumping through any special hoops. Japanese and Chinese, however, remained a problem. Enter WorldScript.

WorldScript: What Is It?
Part II: The Answer

WorldScript is a new feature of the Macintosh operating system version 7.1 which involves an extension to the system implementing two-byte character representation. In other words, instead of having only 8 columns of zeros and ones to represent the characters of a language, a Macintosh running WorldScript uses 16 columns of zeros and ones. How many characters are possible now? Well, 2 to the 16th power, or, 2^{16}, or, 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2 times 2 , or, 65,536 possible characters. That should be enough for most languages, though one report I've read said that Chinese has something on the order of 80,000 characters. I am hopeful, however, that 65,000 characters will be sufficient for most undergraduate students.

Do I Need WorldScript?

You need WorldScript if you work on a Macintosh in English and one of the following: Arabic, Chinese, Japanese, Korean, or Russian. Although some Russian colleagues may argue with me on this, if they would adopt WorldScript, their font problems would go away). You may need WorldScript for other languages as well, but I'm just not familiar with all the languages that don't use alphabets or that read right to left instead of left to right.

How Do I Use WorldScript?

Using WorldScript is very easy. Once you've decided that you need it, you just need to acquire the necessary extensions, scripts and keyboard layouts. (See the section below and the section about Nisus for the easiest way to get these.) Drop all the icons on top of your System 7.1 system folder and reboot your computer. You won't notice much that is different except for a small blue diamond icon at the right end of the menu bar. Depending upon how many scripts you've installed, you'll find several selections in the new menu, probably starting with a small blue diamond icon, followed by the icons representing the language of any installed scripts. On my own computer now, for example, I see a blue diamond, indicating that I have the U.S. system and keyboard map installed; a Russian flag; a green crescent representing Arabic; and a small rising sun and Apple logo icon representing Japanese.

To create texts using these alternate scripts, all I have to do is open my WorldScript compatible word processor (see the final section on Nisus) and begin typing. Since I generally work in English, my computer assumes that I want to begin most documents in English. To switch into Russian, for example, I can do any one of three things: (a) I can select one of the Russian fonts that came with WorldScript (older Russian fonts don't work correctly for the most part), and suddenly the icon in the menu bar changes to the Russian flag and my typing comes out in Russian; (b) I could select the Russian flag I find in the new flag menu, or; (c) I could hold down the command key (⌘) on the keyboard and press the spacebar. This last system will toggle me through all of the installed scripts. I find this to be the easiest way to change scripts since my hands are already on the keyboard. When working in just two languages like Arabic and English, for example, toggling back and forth is quite easy. In the case of Arabic, the direction of your word processing changes with each switch as well (for example, left to right switches to right to left).

WorldScript comes with an assortment of TrueType and bitmap fonts for each of the supported languages. It's intelligent enough to know that if I choose a particular Russian font earlier in the document, then I'll probably want to return to that font when I toggle back and forth from English to Russian, and it does this for me each time I press (⌘)–spacebar.

The real beauty here is that one can word process in just about any language without having to reboot the computer to run the Arabic, Chinese, or Japanese version of the Macintosh operating system. This makes it much easier to use in laboratories since anyone can read the menus—regardless of the primary language of the machine, the menu selections for switching to the language you want always appear in your language without rebooting.

How Do I Type in Japanese or Chinese?

Typing in Japanese and Chinese is unbelievably easy. In fact, one could type in simple Japanese without being able to read Japanese. All you'd need to know is the pronunciation of the words you want to type. For example, to type Fujiyama, I type: "fu" and the computer converts the letters "f" and "u" into the Hirigana character for the sound "fu." To continue, I type "ji" which is then converted to Hirigana, "ya" which is also converted, and finally "ma." What appears on screen at this point are the four Hirigana characters of Japanese which represent the syllables of the word. Of course most Japanese wouldn't use the syllabary representation; they would use the Kanji characters (the borrowed Chinese characters). To convert to Kanji, all I do is press the spacebar and those four characters are converted to the two Kanji which represent Fujiyama. In ambiguous homophonic cases, WorldScript allows the user to select the particular Kanji they intend for the context.

The Chinese system is similar, in that one can type in Pinyin, a system for representing the sounds and tones of Chinese. Thus, to type the character which represents the act one performs when petting one's cat, I would type "fu3"—the sound "fu" and the 3rd tone. If I don't know the tone, I can just type "fu" and I'll be offered all the characters which are pronounced "fu" and I'll have to select the correct one. The Chinese system supports not only Pinyin entry, but Dayi, Parrot, Zhuyin, and Cangjie input methods for Traditional Chinese, and Pinyin, Wubi Xing, Wubi Hua, and Qüwei for Simplified Chinese.

Where Can I Get WorldScript?

There are at least two ways to get WorldScript. If you work in Japanese and Chinese, you'll have to buy the extensions—they aren't free—though this might change in the near future. I had a heck of a time finding the Chinese and Japanese extensions, and so I am going to include the official Apple part numbers here. Even if your campus computer reseller is not familiar with WorldScript, these part numbers should get the right software:

- Japanese Language Kit: M1648/LL/A
- Chinese Language Kit: M2368LL/A

Pricing will vary of course, but these kits should be available for under $150 each. Be prepared when they arrive—the Japanese Kit includes something on the order of 12 megabytes of fonts and the whole installation is approximately 17 megabytes. The Chinese Kit is even larger depending upon whether or not one installs both simplified and/or traditional systems, and both come with several optional, and very large, TrueType fonts. The Chinese Kit was only officially released in October of 1993 and first appeared on the Apple Higher Education price sheet on October 21, 1993.

If you work in Russian, Arabic, or any of the western languages, it's even easier to get WorldScript, and it's free—sort of. All you have to do is buy an academic copy of Nisus, the greatest word processor ever written for the Macintosh (in my opinion, of course). See the section on Nisus for more information.

UNICODE: A STANDARD

Unicode is actually a company, Unicode, Inc., formed by a group of computer companies whose members include Microsoft, IBM, Aldus, NeXT, Apple, GO, Sun, Metaphor, Lotus, Novell, and the Research Libraries Group. This group has been working on a two-byte standard different from

WorldScript. WorldScript grew out of earlier solutions for the Macintosh and Japanese or Chinese, but is internally "unicode-aware" so that migration to Unicode won't be all that difficult once the standard is fully agreed upon. Reportedly, however, Unicode will support only 27,000 possible characters, rather than the 65,000 theoretically possible. Nonetheless, it is relatively complete and includes many mathematical symbols, shapes (like fancy "bullets", etc.) and support for many languages including Gurmukhi, Oriya, and Bopomofo. (If you know what these languages are, write to me and tell me something about them.) For more information about Unicode and the proposed standard, contact the Unicode Consortium at asmusf@microsoft.uucp.

As an interesting aside, Apple's new handheld computer line, the Newton, is Unicode-based and not ASCII- based. Thus, though the current 1.04 release of the Newton operating system doesn't implement alternate script recognition, the core technology is there and it shouldn't be too difficult to add support for languages other than English. When you also take into account the fact that the Newton's handwriting recognition system is at least partly supported by carefully watching the stroke order as you write upon its screen, you can begin to imagine what a wonderful tool it is for Chinese and Japanese.

Well, all of this is wonderful and you're probably ready to run right out and buy WorldScript (or a Newton) to begin word processing in multiple script systems. That would be great except for the fact that to the best of my knowledge, there is only one word processing program that fully supports WorldScript—Nisus from Nisus Software (formerly Paragon Concepts).

Nisus: The Incredible Word Processor

One Macintosh word processing program stands far above the others: Nisus from Nisus Software. Nisus is simply incredible. At the time of this writing it is the only Macintosh word processing program capable of producing Arabic, Chinese, Japanese, Hebrew, Korean, and Russian all in the same document using the English version of the Macintosh operating system and WorldScript. (Microsoft Word, unfortunately, simply doesn't work properly with WorldScript. For example, if you pull down the font menu, the WorldScript fonts appear as garbage characters. When you attempt to type in one of the installed scripts, they work in the entry window, but when placed in the document, the script returns to garbage. You can type correct WorldScript in Nisus, then paste it into Word and it appears correctly. However, if you attempt to edit it, it reverts to garbage characters again.)

Why is Nisus so different? Well, for one thing it's much, no much, MUCH faster than Microsoft Word 5.1. It handles graphics much better than Word and its graphic editor is better than Word's. It will place graphics behind your text, in front of your text, within your text, or flow the text around the graphics. It includes the most sophisticated search and replace engine of any word processor I have ever encountered. For example, just the other day I instructed Nisus to search a 30-page document of mine for all text between quotation marks that appeared in the New Century Schoolbook font. I wanted Nisus to replace that text with the same text found in each instance between quotation marks (i.e., not with a set phrase), but deleting the quotation marks and changing the font to Chicago. All of this unless the text included numbers. This feat was accomplished making a few simple and intuitive menu selections, and took about 2 seconds. Nisus also implements true style sheets and not just the paragraph formats supported by Word and called style sheets.

Nisus is also different because of its very clever design. All Macintosh files have two "forks," a data fork and a resource fork. The designers at Nisus Software decided that Nisus files should contain nothing but pure ASCII text in the data fork of the file. Formatting information and graphics are stored in the resource fork of the file and the Nisus program uses the resource fork to display the file in its formatted form. Most word processing programs completely ignore the resource fork and store all formatting and data in the data fork. (That's why when you try to recover such a file when it becomes corrupted, the data that you recover is often filled with garbage—bad plan if you ask me). The real genius of Nisus' file format is that absolutely any word processing program or text editor on any kind of computer can read Nisus files directly. The files may not display all the special formatting, but they don't have to be converted before they can be read, and they won't be filled with garbage characters or little boxes. (Nisus can also open and/or save files in other common formats such as Word, MacWrite, or WordPerfect, and it supports Claris' XTND technology as well).

Another very nice feature of Nisus is its size. On my Macintosh, Word took just over 6 megabytes of disk space. For that 6 megabytes I got a spelling checker in English, a tutorial, some sample documents, and the application. My complete installation of Nisus is under 5 megabytes, but it includes the spelling and thesaurus dictionaries for English, Spanish, French, Italian and German, a tutorial, sample documents, and the application, which is only 564K. If you're tight on disk space, Nisus can run from a single 800K floppy disk.

Nisus is very popular in the Arabic-speaking world because it handles right to left text (including Hebrew) very well. In Middlebury's School of Arabic, it is the standard. It is bundled with most Macintosh computers sold in Korea and has dominated the Far Eastern market as Microsoft Word and WordPerfect have in the West.

Of course, the best thing about Nisus is the way it works with the less commonly taught foreign languages. It does. Word doesn't, and Microsoft hasn't announced support of WorldScript. The WordPerfect Corporation has announced support of WorldScript but all reports indicate that there are spacing problems with the two-byte character sets, and that it doesn't support right-to-left languages such as Hebrew or Arabic at all. In addition to Nisus, other programs which support WorldScript include: HyperCard, the FirstClass telecommunications system (we can send messages in Arabic and Japanese with ease on our FirstClass conferencing system), and a little known text editor, Style.

Getting used to Nisus if you've been using a different word processor will take a few days. Fortunately, I found it to be much more intuitive than other word processors I've used and often I could do things from a single menu that used to take me two or three dialog boxes. Any keyboard commands can be changed or added, and Nisus will even print a list of all of your keyboard shortcuts.

The only drawback I see with Nisus is that it is copy-protected for the "complete flag" edition—the version which works with all languages. In order to save or print, one has to install a hardware key on the Macintosh. That key is commonly referred to as the "dongle" and it plugs into the same port of the Macintosh as the keyboard. The keyboard cable then plugs into the dongle. Without the dongle, Nisus works as a demo only and you can't save or print. In practice the dongle hasn't been a problem for me since I plugged it in and forgot about it. If you travel with a powerbook, however, you'll have something else to carry around. Still, if you work in English and Japanese, it's the only game in town. Nisus is available from: Nisus Software, 107 South Cedros Solana Beach, CA 92075; (800) 922-2993. If you indicate an academic connection, the cost is $99 and they will include all of the necessary WorldScript extensions and foreign language dictionaries for one foreign language (other than Chinese and Japanese, to which Apple retains exclusive license).

Contributions/suggestions for the "Technical Update" column may be sent directly to David Herren. Mailing address: The Language Schools, Middlebury College, Middlebury, VT 05753; email: herren@middlebury.edu