WorldScript: Background on Different Language Structures

Revised: 1/11/93
Security: Everyone

WorldScript: Background on Different Language Structures

==========

Article Created: 4 January 1993

TOPIC ----------

Many languages have structures, writing directions, or alphabetical sorting orders that differ from those of English. Also, depending upon the geographic location, things such as the calendar, date and time display, and currency conventions can also vary widely. WorldScript deals with these different language and geographic software needs.

DISCUSSION ----------

From a software standpoint, languages around the world are divided into two categories: one-byte and two-byte languages.

One-byte Languages

One-byte languages are languages that require only one byte of memory to fully represent their character set. English and Spanish are examples of one-byte Roman languages, in which A to Z and 0 to 9 are the standard characters. Arabic and Hebrew are examples of one-byte non-Roman languages. (The term non-Roman refers to their different character set.)

The characteristic shared by one-byte languages is that their character sets contain fewer than 256 characters.

Two-byte Languages

Two-byte languages are languages that require two bytes of memory to fully represent their character set. The character sets of these non-Roman languages can consist of as many as 40,000 characters, which means that they require more memory to store.

Two-byte languages are also characterized by multiple character sets. Japanese, for example, has both Kana, a phonetic language of 300 characters,

and Kanji, an ideographic language with up to 40,000 characters. A phonetic language describes pronunciation. In an ideographic language, each character has a specific meaning -— for example, a single Chinese character represents the word sun.

Specialized Text Entry and Display (Input Methods)

The Asian two-byte languages present a text-entry challenge, because they contain far more symbols than can be easily typed directly from a keyboard. The solution is to use an input method to enter text into their documents.

The input technology enables users to type the word phonetically in Kana (a relatively limited character set) and then map that to the Kanji character that they ultimately want to enter. The input method typically provides a small window at the bottom of the screen that displays the typing. An analogy might be if you were to type English by entering the pronunciation marks used in dictionaries. After typing the phonetic description of the word, this technology would return the proper English word or a range of possible words.

Multidirectional Text

Roman languages are written left to right. However, some non-Roman languages, such as Arabic, run right to left, and Asian languages, such as Japanese, run top to bottom. Languages can also require multiple text directions, for example, Japanese is written from top to bottom and right to left.

Contextual Forms

In English, certain character combinations can form what are called ligatures, or physical joins between adjacent letters. For example, in some old English text you see "ae" shown as "æ." Similarly, in Arabic and Hebrew, the appearance of a character is determined by the adjacent characters. Thai is another language that requires this type of contextual formatting.

Language Attributes

Certain languages require different formats for attributes such as their date and time or currency symbols. These needs vary from something as simple as the changing of a 12-hour clock to a 24-hour clock for French, to something as complex as changing from a Gregorian calendar to a Lunar calendar for Arabic.

The International Software Problem

System software can easily support Roman languages such as English and French, but support for non-Roman languages such as Japanese and Chinese can often require the reengineering of the base system. Previous approaches to this problem have tended not to be very user friendly. For example, some systems force users to type in English because the software won't support their language. In other cases, the software is localized to the language, but certain features simply aren't implemented. In the case of the Macintosh before WorldScript, support for non-Roman languages such as Arabic and Japanese was added through a reengineering of the Roman base.

..TIL11252-WorldScript-Background_on_Different_Language_Structures.pdf

WorldScript is a set of integrated software technologies that removes many barriers from global software development, providing a simplified platform for developers to create applications in multiple languages.

Tech Info Library Article Number:11252