Random Fact 4.2

International Alphabets

The English alphabet is pretty simple: upper- and lowercase a to z. Other European languages have accent marks and special characters. For example, German has three umlaut characters (ä, ö, ü) and a double-s character (ß). These are not optional frills; you couldn't write a page of German text without using these characters. German computer keyboards have keys for these characters.

A German Keyboard

Many countries don't use the Roman script at all. Russian, Greek, Hebrew, Arabic, and Thai letters, to name just a few, have completely different shapes (see The Thai Alphabet). To complicate matters, scripts like Hebrew and Arabic are written from right to left instead of from left to right, and many of these scripts have characters that stack above or below other characters, as those marked with a dotted circle do in Thai. Each of these alphabets has between 30 and 100 letters.

The situation is much more dramatic in languages that use Chinese script: the Chinese dialects, Japanese, and Korean. The Chinese script is not alphabetic but ideographic—a character represents an idea or thing rather than a single sound. Tens of thousands of ideographs are in active use.

The inconsistencies among character encodings have been a major nuisance for international electronic communication and for software manufacturers vying for a global market. Between 1988 and 1991 a consortium of hardware and software manufacturers developed a uniform encoding scheme called Unicode that is expressly designed to encode text in all

	จ	ฐ	ธ	ภ	ศ	ะ	ุ	เ	่	๐	๘		เ◌
ก	ฉ	ท	น	ม	ษ	ั	ู	แ	้	๑	๙		แ◌
ข	ช	ฒ	บ	ย	ส	า	◌	โ	๊	๒	ๆ		โ◌
ข	ช	ณ	ป	ร	ห	ำ		ใ	๋	๓	๛		ใ◌
ค	ฌ	ด	ผ	ฤ	ฬ	ิ		ไ	์	๔			ไ◌
ค	ญ	ต	ฝ	ล	อ	ี		ๅ	◌	๕			
ม	ฎ	ถ	พ	ฦ	ฮ	ึ		ๆ	๎	๖			
ง	ฏ	ท	ฟ	ว	ฯ	ื		็		๗			

The Thai Alphabet

written languages of the world (see www.unicode.org). In the first version of Unicode, about 39,000 characters were given codes, including 21,000 Chinese ideographs. A 2-byte code (which can encode over 65,000 characters) was chosen. It was thought to leave ample space for expansion for esoteric scripts, such as Egyptian hieroglyphs and the ancient script used on the island of Java.

Java was one of the first programming languages to embrace Unicode. All Unicode characters can be stored in Java strings, but which ones can actually be displayed depends on your computer system. The primitive type `char` denotes a 2-byte Unicode character.)

Unfortunately, in 2003, the inevitable happened. Another large batch of Chinese ideographs had to be added to Unicode, pushing it beyond the 16-bit limit. Now, some characters need to be encoded with a pair of `char` values (see Special Topic 4.5).

Chinese Ideographs