read multibyte characters, convert them to wide characters for manipulation within the program, and then convert the wide characters back to multibyte form for output.

Q: Unicode and UCS seem to be pretty much the same. What's the difference between the two? [p. 650]

A: Both contain the same characters, and characters are represented by the same code points in both. Unicode is more than just a character set, though. For example, Unicode supports "bidirectional display order." Some languages, including Arabic and Hebrew, allow text to be written from right to left instead of left to right. Unicode is capable of specifying the display order of characters, allowing text to contain some characters that are to be displayed from left to right along with others that go from right to left.

Exercises

Section 25.1

1. Determine which locales are supported by your compiler.

Section 25.2

2. The Shift-JIS encoding for *kanji* requires either one or two bytes per character. If the first byte of a character is between `0x81` and `0x9f` or between `0xe0` and `0xef`, a second byte is required. (Any other byte is treated as a whole character.) The second byte must be between `0x40` and `0x7e` or between `0x80` and `0xfc`. (All ranges are inclusive.) For each of the following strings, give the value that the `mbcheck` function of Section 25.2 will return when passed that string as its argument, assuming that multibyte characters are encoded using Shift-JIS in the current locale.

   (a) `"\x05\x87\x80\x36\xed\xaa"`
   (b) `"\x20\xe4\x50\x88\x3f"`
   (c) `"\xde\xad\xbe\xef"`
   (d) `"\x8a\x60\x92\x74\x41"`

3. One of the useful properties of UTF-8 is that no sequence of bytes within a multibyte character can possibly represent another valid multibyte character. Does the Shift-JIS encoding for *kanji* (discussed in Exercise 2) have this property?

4. Give a C string literal that represents each of the following phrases. Assume that the characters à, è, é, ĉ, î, ô, û, and ü are represented by single-byte Latin-1 characters. (You'll need to look up the Latin-1 code points for these characters.) For example, the phrase *déjà vu* could be represented by the string `"d\xe9j\xe0 vu"`.

   (a) *Côte d'Azur*
   (b) *crème brûlée*
   (c) *crème fraîche*
   (d) *Fahrvergnügen*
   (e) *tête-à-tête*

5. Repeat Exercise 4, this time using the UTF-8 multibyte encoding. For example, the phrase *déjà vu* could be represented by the string `"d\xc3\xa9j\xc3\xa0 vu"`.