On word breaking in Chinese and Japanese

devblogs.microsoft.com/oldnewthing/20160307-00

March 7, 2016

Raymond Chen

In Western languages, you can generally break a line at whitespace. (You can also break a line within a word, subject to language-specific hyphenation rules, but let's not get into that.) People unfamiliar with other language families sometimes wonder what's up with line breaking in other languages. In particular, line breaking in Chinese and Japanese tend to elicit confused responses.

> When I put text in a static control and it does not fit, the behavior is different depending on whether I'm using Chinese characters or Latin characters. Why does the Chinese string wrap to the second line, but the Latin string does not?
>
> ㄅㄆㄇㄈㄉㄊㄋㄌㄍㄎㄏ
>
> ABCDEFGHIJKLMNOPQRSTUVWXYZ.

In Chinese and Japanese, there are no spaces between words, so if you're going to wait for a space before inserting a line break, you're going to be waiting a long time. Instead, to a first approximation, line breaks are permitted after almost any character. (You can learn the finer points of line breaking from Wikipedia.)

The static control uses Uniscribe to decide where to insert line breaks, and Uniscribe understands that in Chinese and Japanese text, you can break after almost any character. That's why you're seeing a line break in the static control with Chinese text. On the other hand, the static control cannot find a valid word break in the Latin string, so it all gets jammed onto one line (and the excess gets clipped).

The `DrawText` function also has rudimentary understanding of line breaks in Chinese, Japanese, and Korean text. You can override the default line breaking rule of "line breaks allowed after any full-width character" by passing the `DT_NOFULLWIDTHCHARBREAK` flag, which forces the `DrawText` function to break only at whitespace. (Basically, have it treat CJK characters as if they were Latin.)

The documentation for `DT_NOFULLWIDTHCHARBREAK` notes that it may be useful to pass this flag if you know that the text is Korean, because Korean does put spaces between words, and preferring to break Korean text at whitespace can result in more attractive results. (The `DrawText` function is not very clever and does not try to autodetect whether the string is Korean. It is legal to mix Chinese characters into Korean text, and trying to figure out whether the string is "Mostly Korean with Chinese characters mixed in" or "Mostly Chinese with Korean mixed in" would require too much fuzzy logic for the simple `DrawText` function.)

Bonus chatter: You thought Chinese, Japanese, and Korean line breaking is hard. Thai is even harder. In Thai, words are run together with no spaces, but line breaks are permitted only between words. This means that in order to break lines properly, you need a Thai dictionary.

Bonus bonus chatter: On that last page I linked to, there is a reference to the Windows Intelligent Font Emulator, which went by the acronym WIFE. Somebody probably worked really hard to retrofit that acronym.

Raymond Chen

Follow