Title: Proposal to add a HangulSources.txt data file to the UCD
Author: Ken Lunde & Jaemin Chung
Date: 2024-04-22

Proposal

This document proposes that a new data file, *HangulSources.txt*, be added to the UCD in Unicode Version 17.0. The actual data file, with proposed header text, is attached to this document as a PDF attachment, and specifies the following six semicolon-delimited fields:

0: Unicode code point

1: Johab code point in hexadecimal

2: KS X 1001 GR code point in hexadecimal

3: KS X 1002 GL code point in hexadecimal

4: KPS 9566 GR code point in hexadecimal

5: GB/T 12052 GL code point in hexadecimal

For Fields 2 and 4, code points that are prefixed with an asterisk correspond to platform-specific extensions that support the complete set of modern hangul syllables (11,172). For Field 2, the implementation is Unified Hangul Code (UHC), also known as Microsoft Code Page 949, which supports 8,822 additional modern hangul syllables beyond the 2,350 in KS X 1001 proper. For Field 4, the implementation is Red Star OS (see document L2/18-011), which supports 8,493 additional modern hangul syllables beyond the 2,679 in KPS 9566 proper. For both fields, the GR (*Graphic Right* or eight-bit) encoding is specified for the hangul syllables that are included in the corresponding standards proper, because these legacy character sets were most commonly implemented with this encoding.

For Fields 3 and 5, the GL (*Graphic Left* or seven-bit) encoding is specified, because these legacy character sets were never implemented using the GR encoding.

The hangul syllable that corresponds to Field 0 is provided as a comment at the end of each line of the data file.

The mappings to the four regional standards in Fields 2 through 5 of the proposed UCD data file—without the platform-specific extensions of Fields 2 and 4—were once provided in document L2/17-080 (Chung).

Background

Among the scripts in the Unicode Standard whose corresponding blocks include several thousand characters, only the *Hangul* script in the Hangul Syllables block lacks source information for its characters. In addition, the *Hangul* script includes the second largest number of characters in the Unicode Standard, second only to the *Han* script. This justification in and of itself should serve as sufficient evidence to add the proposed data file to the UCD.

That is all.