Title: Clarification Regarding Use of Jamo Filler Characters: US / UTC Response to N3535, N3422

Date: 2009-4-19
Source: US NB & Unicode Consortium
Status: Joint NB, Liaison contribution
Action: For review by WG2 experts
Distribution: WG2

Background

In Section 4 of N3422, the Republic of Korea national body (ROK) identifies a discrepancy between recommendations in ISO/IEC 10646:2003 and The Unicode Standard regarding the encoded representation of certain Hangul text elements in relation to the use of jamo filler characters:

> - A syllable-final letter alone is represented as "syllable-initial filler (U115F) + syllable final letter" in ISO/IEC 10646. This method is what Korea NB proposed at WG2 meeting in Seoul in 1992.
>
> - However, somehow, in Unicode, a different method is specified: "syllable-initial filler (U115F) + syllable-peak filler (U1160) + syllable final letter" (i.e., a syllable-peak filler (U1160) is added).
>
> - We need to resolve the discrepancy.

In N3535, ROK provides additional details, quoting relevant portions of The Unicode Standard 5.0 and of ISO/IEC 10646:2003 (plus subsequent amendments). The relevant portion from ISO/IEC 10646:2003, Clause 26.1, is the following:

> An incomplete syllable is a string of one or more characters which does not constitute a complete syllable (for example, a Choseong alone, a Jungseong alone, a Jongseong alone, or a Jungseong followed by a Jongseong). An incomplete syllable which starts with a Jungseong or a Jongseong shall be preceded by a CHOSEONG FILLER (0000 115F). An incomplete syllable composed of a Choseong alone shall be followed by a JUNGSEONG FILLER (0000 1160).

The concern is apparent disagreement with the following text from The Unicode Standard, Version 5.0, section 3.12 (p. 120):

> ***Transforming into Standard Korean Syllables.*** A sequence of jamos that do not all match the regular expression for a standard Korean syllable block can be transformed into a sequence of standard Korean syllable blocks by the correct insertion of choseong fillers and jungseong fillers. This transformation of a string of text into standard Korean syllables is performed by determining the syllable breaks as explained in the earlier subsection “Hangul Syllable Boundaries,” then inserting one or two fillers as necessary to transform each syllable into a standard Korean syllable…

So, for example, consider the case of an incomplete syllable with syllable-final consonant (jongseong) KIYEOK (U+11A8): Unicode 5.0 specifies a sequence using both the syllable-initial consonant (choseong) filler, U+115F, and the syllable-peak (jungseong) filler U+1160:

< U+115F, U+1160, U+11A8 >

In contrast, the text of ISO/IEC 10646:2003 is interpreted as specifying a sequence that includes only the syllable-initial consonant filler:

< U+115F, U+11A8 >

It is this discrepancy that ROK is concerned to see resolved.

Comments and recommendation

The US national body and the Unicode Consortium agree that this is a discrepancy that should be resolved.

In the discussion that follows, the following notation will be used:

L —a syllable-initial consonant (choseong) jamo

Lf —CHOSEONG FILLER (U+115F)

V — a syllable-peak (jungseong) jamo

Vf — JUNGSEONG FILLER (U+1160)

T — a syllable-final consonant (jungseong jamo)

Arguably, the text of Clause 26.1 in ISO/IEC 10646 might be considered ambiguous: It clearly requires that Lf be used if the first jamo letter is V or T (which Unicode's specification also requires), but is unclear as to the need for Vf in the case that the first jamo letter is T. On one reading, it is simply an incomplete specification that fails to mention a required Vf in this situation. In a different reading, the specification for this situation is complete, and no Vf is used. Clearly, ROK is assuming the latter interpretation. That interpretation is probably how most readers are likely to understand the text, and may well be what was originally intended, based on ROK's comment in N3422, "This method is what Korea NB proposed at WG2 meeting in Seoul in 1992."

Whatever may have been proposed in 1992, it is clear that the initial encoding for Hangul text left outstanding problems that are only now getting resolved, and that scenarios and requirements for Hangul text are understood better now than they were then. The addition of 121 new conjoining jamo characters in Amendment 5 fills in gaps that had caused issues in representing some Old Hangul syllables, and now allows a simple and consistent schema for representing all known-attested Hangul syllables: L + V (+ T).

The specification for incomplete syllables given in The Unicode Standard is consistent with that simple schema for representation of Hangul syllables: requiring both Lf and Vf fillers results in a sequence that includes one each of L, V and T elements. In contrast, the specification in ISO/IEC 10646, Clause 26.1 results in a more complex set of schemas: L/Lf + V/Vf (+T), or Lf + T. This necessitates greater complexity for implementations of various kinds of text processes that may be applied to Hangul text.

It is also noted that Section 5.1 (p. 4) of KS X 1026-1:2007 recommends the use of both Lf and Vf fillers in the case of an incomplete syllable with only a T jamo:

> If L, V, T stand for a syllable-initial letter, a syllable-peak letter, a syllable-final letter, respectively; and if LF, VF stand for a syllable-intial filler, a syllable-peak filler, respectively, these rules can be expressed as regular expressions as shown below:
>
> …
>
> 3) A representation of a syllable-final letter along: L_F V_F T

Thus, based on KS X 1026-1:2007, it appears that ROK has opted in favour of the specification in The Unicode Standard over that in Clause 26.1 of ISO/IEC 10646. (This would be understandable given that it is consistent with recommending the simpler schema, L + V + (Lf).)

Accordingly, the US national body and the Unicode Consortium recommend that the text of ISO/IEC 10646 be revised to give a clear specification of which fillers are required in each of the possible cases for an incomplete syllable. In particular, it is recommended that the specification require the use of a choseong filler whenever a choseong jamo is not present, and a jungseong filler whenever a jungseong jamo is not present, so that syllable representations always include choseong and jungseong elements. Thus, in five possible patterns for incomplete syllables, the recommended sequences would be as follows:

- An incomplete syllable with only a syllable-initial consonant (choseong): *choseong jamo + JUNGSEONG FILLER (U+1160)*
- An incomplete syllable with only a syllable-peak character (jungseong): *CHOSEONG FILLER (U+115F) + jungseong jamo*
- An incomplete syllable with only a syllable-final consonant (jongseong): *CHOSEONG FILLER (U+115F) + JUNGSEONG FILLER (U+1160) + jongseong jamo*
- An incomplete syllable with a syllable-initial consonant (choseong) and a syllable-final consonant (jongseong): *choseong jamo + JUNGSEONG FILLER (U+1160) + jongseong jamo*
- An incomplete syllable with a syllable-peak character (jungseong) and a syllable-final consonant (jongseong): *choseong filler (U+115F) + jungseong jamo + jongseong jamo*

The following is a suggested revision to the text in Clause 26.1 to achieve this result: Remove the second paragraph and substitute in its place the following text:

> An incomplete syllable is a string of one or more characters which does not constitute a complete syllable: a Choseong alone, a Jungseong alone, a Jongseong alone, a Choseong

followed by a Jongseong, or a Jungseong followed by a Jongseong. In encoded representation, any incomplete syllable will include either CHOSEONG FILLER (0000 115F), JUNGSEONG FILLER (0000 1160), or both as needed to ensure that the syllable has exactly one Choseong element (a syllable-initial character or a filler) and one Jungseong element (a syllable-peak character or a filler). For example, a incomplete syllable consisting of a Choseong along would be represented as the Choseong character followed by JUNGSEONG FILLER; an incomplete syllable consisting of a Jongseong alone would be represented as CHOSEONG FILLER followed by JUNGSEONG FILLER followed by the Jongseong character.