# A few miscellaneous comments on 14651 CTT (and UCA default table) ordering rules

**Kent Karlsson**

**2003-10-15**

## Digit cluster characters which happen to have just one digit (instead of two)

In order to get the expected order for digit cluster characters, those that have just one digit should be seen as having an implicit leading zero. It is sufficient with one leading zero, since all digit cluster characters have one or two digits, never more. Thus the following characters should be treated as if they had *collation* decompositions with an (additional) leading zero. I.e. instead of:

24EA;CIRCLED DIGIT ZERO;No;0;EN;<circle> 0030;;0;0;N;;;;;

the *collation* decomposition should include a leading 0 (U+0030):

24EA;CIRCLED DIGIT ZERO;No;0;EN;<circle> 0030 0030;;0;0;N;;;;;

Similarly, instead of:

32C6;IDEOGRAPHIC TELEGRAPH SYMBOL FOR JULY;So;0;L;<compat> 0037 6708;;;;N;;;;;

the *collation* decomposition should include a leading 0:

32C6;IDEOGRAPHIC TELEGRAPH SYMBOL FOR JULY;So;0;L;<compat> 0030 0037 6708;;;;N;;;;;

None of these characters take part in forming numerals, but stand for themselves. The set of such characters is such that they at most decompose to two digits, never more.

Here's the list of character for which this should be done:

24EA;CIRCLED DIGIT ZERO;No;0;EN;<circle> 0030;;0;0;N;;;;;
2460;CIRCLED DIGIT ONE;No;0;EN;<circle> 0031;;1;1;N;;;;;
2461;CIRCLED DIGIT TWO;No;0;EN;<circle> 0032;;2;2;N;;;;;
2462;CIRCLED DIGIT THREE;No;0;EN;<circle> 0033;;3;3;N;;;;;
2463;CIRCLED DIGIT FOUR;No;0;EN;<circle> 0034;;4;4;N;;;;;
2464;CIRCLED DIGIT FIVE;No;0;EN;<circle> 0035;;5;5;N;;;;;
2465;CIRCLED DIGIT SIX;No;0;EN;<circle> 0036;;6;6;N;;;;;
2466;CIRCLED DIGIT SEVEN;No;0;EN;<circle> 0037;;7;7;N;;;;;
2467;CIRCLED DIGIT EIGHT;No;0;EN;<circle> 0038;;8;8;N;;;;;
2468;CIRCLED DIGIT NINE;No;0;EN;<circle> 0039;;9;9;N;;;;;

2474;PARENTHESIZED DIGIT ONE;No;0;EN;<compat> 0028 0031 0029;;1;1;N;;;;;
2475;PARENTHESIZED DIGIT TWO;No;0;EN;<compat> 0028 0032 0029;;2;2;N;;;;;
2476;PARENTHESIZED DIGIT THREE;No;0;EN;<compat> 0028 0033 0029;;3;3;N;;;;;
2477;PARENTHESIZED DIGIT FOUR;No;0;EN;<compat> 0028 0034 0029;;4;4;N;;;;;
2478;PARENTHESIZED DIGIT FIVE;No;0;EN;<compat> 0028 0035 0029;;5;5;N;;;;;
2479;PARENTHESIZED DIGIT SIX;No;0;EN;<compat> 0028 0036 0029;;6;6;N;;;;;
247A;PARENTHESIZED DIGIT SEVEN;No;0;EN;<compat> 0028 0037 0029;;7;7;N;;;;;
247B;PARENTHESIZED DIGIT EIGHT;No;0;EN;<compat> 0028 0038 0029;;8;8;N;;;;;
247C;PARENTHESIZED DIGIT NINE;No;0;EN;<compat> 0028 0039 0029;;9;9;N;;;;;

2488;DIGIT ONE FULL STOP;No;0;EN;<compat> 0031 002E;;1;1;N;DIGIT ONE PERIOD;;;;
2489;DIGIT TWO FULL STOP;No;0;EN;<compat> 0032 002E;;2;2;N;DIGIT TWO PERIOD;;;;
248A;DIGIT THREE FULL STOP;No;0;EN;<compat> 0033 002E;;3;3;N;DIGIT THREE PERIOD;;;;
248B;DIGIT FOUR FULL STOP;No;0;EN;<compat> 0034 002E;;4;4;N;DIGIT FOUR PERIOD;;;;
248C;DIGIT FIVE FULL STOP;No;0;EN;<compat> 0035 002E;;5;5;N;DIGIT FIVE PERIOD;;;;
248D;DIGIT SIX FULL STOP;No;0;EN;<compat> 0036 002E;;6;6;N;DIGIT SIX PERIOD;;;;
248E;DIGIT SEVEN FULL STOP;No;0;EN;<compat> 0037 002E;;7;7;N;DIGIT SEVEN PERIOD;;;;
248F;DIGIT EIGHT FULL STOP;No;0;EN;<compat> 0038 002E;;8;8;N;DIGIT EIGHT PERIOD;;;;
2490;DIGIT NINE FULL STOP;No;0;EN;<compat> 0039 002E;;9;9;N;DIGIT NINE PERIOD;;;;

24F5;DOUBLE CIRCLED DIGIT ONE;No;0;ON;;;1;1;N;;;;;
24F6;DOUBLE CIRCLED DIGIT TWO;No;0;ON;;;2;2;N;;;;;
24F7;DOUBLE CIRCLED DIGIT THREE;No;0;ON;;;3;3;N;;;;;
24F8;DOUBLE CIRCLED DIGIT FOUR;No;0;ON;;;4;4;N;;;;;
24F9;DOUBLE CIRCLED DIGIT FIVE;No;0;ON;;;5;5;N;;;;;
24FA;DOUBLE CIRCLED DIGIT SIX;No;0;ON;;;6;6;N;;;;;

24FB;DOUBLE CIRCLED DIGIT SEVEN;No;0;ON;;;7;7;N;;;;;
24FC;DOUBLE CIRCLED DIGIT EIGHT;No;0;ON;;;8;8;N;;;;;
24FD;DOUBLE CIRCLED DIGIT NINE;No;0;ON;;;9;9;N;;;;;

24FF;NEGATIVE CIRCLED DIGIT ZERO;No;0;ON;;;0;0;N;;;;;
2776;DINGBAT NEGATIVE CIRCLED DIGIT ONE;No;0;ON;;;1;1;N;INVERSE CIRCLED DIGIT ONE;;;;
2777;DINGBAT NEGATIVE CIRCLED DIGIT TWO;No;0;ON;;;2;2;N;INVERSE CIRCLED DIGIT TWO;;;;
2778;DINGBAT NEGATIVE CIRCLED DIGIT THREE;No;0;ON;;;3;3;N;INVERSE CIRCLED DIGIT THREE;;;;
2779;DINGBAT NEGATIVE CIRCLED DIGIT FOUR;No;0;ON;;;4;4;N;INVERSE CIRCLED DIGIT FOUR;;;;
277A;DINGBAT NEGATIVE CIRCLED DIGIT FIVE;No;0;ON;;;5;5;N;INVERSE CIRCLED DIGIT FIVE;;;;
277B;DINGBAT NEGATIVE CIRCLED DIGIT SIX;No;0;ON;;;6;6;N;INVERSE CIRCLED DIGIT SIX;;;;
277C;DINGBAT NEGATIVE CIRCLED DIGIT SEVEN;No;0;ON;;;7;7;N;INVERSE CIRCLED DIGIT SEVEN;;;;
277D;DINGBAT NEGATIVE CIRCLED DIGIT EIGHT;No;0;ON;;;8;8;N;INVERSE CIRCLED DIGIT EIGHT;;;;
277E;DINGBAT NEGATIVE CIRCLED DIGIT NINE;No;0;ON;;;9;9;N;INVERSE CIRCLED DIGIT NINE;;;;

2780;DINGBAT CIRCLED SANS-SERIF DIGIT ONE;No;0;ON;;;1;1;N;CIRCLED SANS-SERIF DIGIT ONE;;;;
2781;DINGBAT CIRCLED SANS-SERIF DIGIT TWO;No;0;ON;;;2;2;N;CIRCLED SANS-SERIF DIGIT TWO;;;;
2782;DINGBAT CIRCLED SANS-SERIF DIGIT THREE;No;0;ON;;;3;3;N;CIRCLED SANS-SERIF DIGIT THREE;;;;
2783;DINGBAT CIRCLED SANS-SERIF DIGIT FOUR;No;0;ON;;;4;4;N;CIRCLED SANS-SERIF DIGIT FOUR;;;;
2784;DINGBAT CIRCLED SANS-SERIF DIGIT FIVE;No;0;ON;;;5;5;N;CIRCLED SANS-SERIF DIGIT FIVE;;;;
2785;DINGBAT CIRCLED SANS-SERIF DIGIT SIX;No;0;ON;;;6;6;N;CIRCLED SANS-SERIF DIGIT SIX;;;;
2786;DINGBAT CIRCLED SANS-SERIF DIGIT SEVEN;No;0;ON;;;7;7;N;CIRCLED SANS-SERIF DIGIT SEVEN;;;;
2787;DINGBAT CIRCLED SANS-SERIF DIGIT EIGHT;No;0;ON;;;8;8;N;CIRCLED SANS-SERIF DIGIT EIGHT;;;;
2788;DINGBAT CIRCLED SANS-SERIF DIGIT NINE;No;0;ON;;;9;9;N;CIRCLED SANS-SERIF DIGIT NINE;;;;

278A;DINGBAT NEGATIVE CIRCLED SANS-SERIF DIGIT ONE;No;0;ON;;;1;1;N;INVERSE CIRCLED SANS-SERIF DIGIT ONE;;;;
278B;DINGBAT NEGATIVE CIRCLED SANS-SERIF DIGIT TWO;No;0;ON;;;2;2;N;INVERSE CIRCLED SANS-SERIF DIGIT TWO;;;;
278C;DINGBAT NEGATIVE CIRCLED SANS-SERIF DIGIT THREE;No;0;ON;;;3;3;N;INVERSE CIRCLED SANS-SERIF DIGIT THREE;;;;
278D;DINGBAT NEGATIVE CIRCLED SANS-SERIF DIGIT FOUR;No;0;ON;;;4;4;N;INVERSE CIRCLED SANS-SERIF DIGIT FOUR;;;;
278E;DINGBAT NEGATIVE CIRCLED SANS-SERIF DIGIT FIVE;No;0;ON;;;5;5;N;INVERSE CIRCLED SANS-SERIF DIGIT FIVE;;;;
278F;DINGBAT NEGATIVE CIRCLED SANS-SERIF DIGIT SIX;No;0;ON;;;6;6;N;INVERSE CIRCLED SANS-SERIF DIGIT SIX;;;;
2790;DINGBAT NEGATIVE CIRCLED SANS-SERIF DIGIT SEVEN;No;0;ON;;;7;7;N;INVERSE CIRCLED SANS-SERIF DIGIT SEVEN;;;;
2791;DINGBAT NEGATIVE CIRCLED SANS-SERIF DIGIT EIGHT;No;0;ON;;;8;8;N;INVERSE CIRCLED SANS-SERIF DIGIT EIGHT;;;;
2792;DINGBAT NEGATIVE CIRCLED SANS-SERIF DIGIT NINE;No;0;ON;;;9;9;N;INVERSE CIRCLED SANS-SERIF DIGIT NINE;;;;

32C0;IDEOGRAPHIC TELEGRAPH SYMBOL FOR JANUARY;So;0;L;<compat> 0031 6708;;;;N;;;;;
32C1;IDEOGRAPHIC TELEGRAPH SYMBOL FOR FEBRUARY;So;0;L;<compat> 0032 6708;;;;N;;;;;
32C2;IDEOGRAPHIC TELEGRAPH SYMBOL FOR MARCH;So;0;L;<compat> 0033 6708;;;;N;;;;;
32C3;IDEOGRAPHIC TELEGRAPH SYMBOL FOR APRIL;So;0;L;<compat> 0034 6708;;;;N;;;;;
32C4;IDEOGRAPHIC TELEGRAPH SYMBOL FOR MAY;So;0;L;<compat> 0035 6708;;;;N;;;;;
32C5;IDEOGRAPHIC TELEGRAPH SYMBOL FOR JUNE;So;0;L;<compat> 0036 6708;;;;N;;;;;
32C6;IDEOGRAPHIC TELEGRAPH SYMBOL FOR JULY;So;0;L;<compat> 0037 6708;;;;N;;;;;
32C7;IDEOGRAPHIC TELEGRAPH SYMBOL FOR AUGUST;So;0;L;<compat> 0038 6708;;;;N;;;;;
32C8;IDEOGRAPHIC TELEGRAPH SYMBOL FOR SEPTEMBER;So;0;L;<compat> 0039 6708;;;;N;;;;;

3358;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR ZERO;So;0;L;<compat> 0030 70B9;;;;N;;;;;
3359;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR ONE;So;0;L;<compat> 0031 70B9;;;;N;;;;;
335A;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR TWO;So;0;L;<compat> 0032 70B9;;;;N;;;;;
335B;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR THREE;So;0;L;<compat> 0033 70B9;;;;N;;;;;
335C;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR FOUR;So;0;L;<compat> 0034 70B9;;;;N;;;;;
335D;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR FIVE;So;0;L;<compat> 0035 70B9;;;;N;;;;;
335E;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR SIX;So;0;L;<compat> 0036 70B9;;;;N;;;;;
335F;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR SEVEN;So;0;L;<compat> 0037 70B9;;;;N;;;;;

3360;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR EIGHT;So;0;L;<compat> 0038 70B9;;;;N;;;;;
3361;IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOUR NINE;So;0;L;<compat> 0039 70B9;;;;N;;;;;

It is very unlikely that a tailoring will correct the weighting for these characters, and even if so, that might happen only for some tailorings, but this should apply to all. Not even a special handling for correct numeric ordering of numerals for natural numbers in base ten will address the ordering of these characters.

## Braille

Braille should not be ignored at levels 1-3. The default collation should be significant from level 1, even though the ordering is nonsensical for all applications of Braille to code letters (and punctuation, as well as "state shifts" which are employed in Braille). For an alphabetically correct ordering, a tailoring is still needed.

## Subsyllabic collation clustering (Brahmic derived scripts and Hangul)

For all the Indic scripts, incl. Khmer, Thai, Lao, etc.: The dependent vowels should have weights heavier than all scripts. (Could be made heavier than <TFFFF>, but heavier than <SFFE1> is sufficient.) The Viramas should be made heavier still. The reason for this is that the Indic scripts are syllabically ordered. E.g.,

Correct ordering (*not* produced by current tables):
<KA>
<KA, latin>
<KA, han>
<KA, II>
<KA, II, latin>
<KA, II, han>
<KA, VIRAMA, JA>
<KA, VIRAMA, JA, latin>
<KA, VIRAMA, JA, han>
<KA, VIRAMA, JA, II>
<KA, VIRAMA, JA, II, latin>
<KA, VIRAMA, JA, II, han>

Current, incorrenct, ordering (though it is correct if no other script intervenes):

<KA>
<KA, latin>
<KA, II>
<KA, II, latin>
<KA, II, han>
<KA, VIRAMA, JA>
<KA, VIRAMA, JA, latin>
<KA, VIRAMA, JA, II>
<KA, VIRAMA, JA, II, latin>
<KA, VIRAMA, JA, II, han>
**<KA, VIRAMA, JA, han>**
**<KA, han>**

Note how han (CJK) characters violates the clustering principle. The suggested change fixes this problem with the current Unicode default ordering for the Indic scripts. This applies also to the Thai, Lao and Khmer scripts (see separate proposal papers, N1076, N1077). There is a similar, though more intricate, issue that applies to Hangul. For detail see the separate proposal paper on Hangul ordering (N1051).

Note that the use of a "lightweight" cluster separator weight can for all practical cases be avoided. Only if Hangul is to be handled in full generality (due to the many unnecessary letter cluster Jamo characters that have been allocated) without the use of several thousand contractions, would such separator weights be needed.

## Indic vowel piece characters

Indic vowel piece characters *that do not form a vowel by themselves*, e.g. 0BD7, 0C55, 0C56, 0CD5, and more, should be ignored at levels 1-3 when not part of a properly constructed dependent vowel. This is because then they are meaningless symbols (and in an ideal world, they should not have been encoded).