L2/12-227

UCA constructs ill-formed 4th level weights

Author: Markus Scherer
Date: 2012-jul-13

Reference: UCA (UTS #10) draft for UCA 6.2, section 3.7 Well-Formed Collation Element Tables

The UCA defines conditions for well-formed collation element tables. The second condition is:

All Level N weights in Level N-2 ignorables must be strictly less than all weights in Level N-1 ignorables.

- For example, secondaries in primary collation elements must be strictly less than those in secondary collation elements: given collation elements [C, D, E] and [0, A, B], where C ≠ 0 and A ≠ 0, D *must be* less than A.
- For a detailed example of what happens if the condition is not met, see *Section 4.5 Rationale for Well-Formed Collation Element Tables*.

Proposal

Change the condition to only apply to levels 2 & 3, and clarify it. Suggested text:

Secondary weights of secondary collation elements must be strictly greater than secondary weights of all primary collation elements. Tertiary weights of tertiary collation elements must be strictly greater than tertiary weights of all primary and secondary collation elements.

- Given collation elements [A, B, C], [0, D, E], [0, 0, F], where the letters are non-zero weights, the following must be true:
  - D > B
  - F > C
  - F > E
- For a detailed example of what happens if the condition is not met, see *Section 4.5 Rationale for Well-Formed Collation Element Tables*.

This retains the current, desired behavior of “Shifted” variable ordering, avoids a conflict between the L4 weight computation and the well-formedness condition, and clarifies that the condition also applies to tertiary weights of primary collation elements, and is easier to make out.

Rationale

For N=4, the condition says “All Level 4 weights in Level 2 ignorables must be strictly less than all weights in Level 3 ignorables.” In other words, the q in collation elements like p.s.t.q (where at least t is not zero) must be strictly less than the x in collation elements like 0.0.0.x.

There are two problems with this.

Problem 1

When variable collation elements are Shifted, then table 12 (L4 Weights for Shifted Variables) specifies that level 4 weights are constructed with the opposite order of what the well-formedness condition requires. Shifted-variable collation elements become 0.0.0.p but other non-ignorable collation elements get a maximum L4 weight of FFFF>p.

The Shifted column of table 13 (Comparison of Variable Ordering) shows this order:

death
de luge
de-luge
de-luge
deluge
de Luge
de-Luge
de-Luge
deLuge
demark

This seems to be the desired “ignore punctuation” behavior.

If the well-formedness condition were met for level 4 by using an L4 weight for non-ignorables below shifted primaries, e.g. 01FF, the Shifted order would be:

death
deluge
de luge
de-luge
de-luge
deLuge
de Luge
de-Luge
de-Luge
demark

This would be consistent with the behavior of secondary and tertiary collation elements: The

insertion of a character with a minor difference makes a string sort after the version without that additional character.

If the currently specified behavior is desired, then the second well-formedness condition needs to make an exception for N=4.

If no such exception should be made, then the Shifted algorithm needs to be modified and the different sorting result documented.

Problem 2

This well-formedness condition is hard to read. More importantly, for N>2, it is incomplete.

For example, with N=3 (tertiary weights), the condition is silent about tertiary weights of primary CEs (“level 0 ignorables”).

To complete the condition and keep it parameterized by N, it could be “All Level N weights in collation elements that are not N-1-ignorable must be strictly less than all weights in Level N-1 ignorables.”

If the condition is modified to exclude N=4, then it would be simpler to spell it out for N=2 and N=3:

- Secondary weights of secondary collation elements must be strictly greater than secondary weights of primary collation elements.
- Tertiary weights of tertiary collation elements must be strictly greater than tertiary weights of primary and secondary collation elements.