ISO/IEC JTC1/SC2/WG2 N4536

L2/13-225

Cedillas and commas below, take 2

Eric Muller, Adobe

November 7, 2013

§1. Currently, the Unicode standard specifies that the character U+0327 ◌̧ COMBINING CEDILLA can be displayed by a cedilla (e.g. as is typically done in French) and can also be displayed by a comma (e.g. as is typically done in Latvian). In other words, the character is ambiguous.

The standard also encourages, via the code charts, to display a cedilla when used with the letters *c, e, h* and *s,* and to display a comma when used with the letters *d, g, k, l, n* and *r*. There is no indication for the other letters, but the common practice is to display a cedilla. This is particularly applicable to workhorse fonts (e.g. Times New Roman), in the absence of more specific information (such as locale). This encouragement is informal, but is quite important to the Latvian users, for example.

For users which have a strong preference for a comma on letters which are by default displayed with a cedilla, the plain text solution is to use U+0326 ◌̦ COMMA BELOW, which at this point is unambiguous. This is the solution for Romanian, since the default form from *s* displays a cedilla, and it extends to the other letters with comma used in Romanian, i.e. *t*.

There is no plain text solution for the opposite case, i.e. for users which have a strong preference for a cedilla on letters which are by default displayed with a comma. This is the case for Marshallese, which uses *l, m, n* and *o* with cedillas, get what they expect for *m* and *o,* but do not get what they expect for *l* and *n*.

§2. If it is desired to offer a plain text solution for the reliable display of a cedilla, given the current situation and the desires of stability, it seems to me that the best solution is to encode a new character, an unambiguous combining cedilla, may be named COMBINING INVARIANT CEDILLA.

I do not believe that this new character would cause particular problems. It is true that it would in principle introduce an alternate representation of the French *ç*, but it seems very unlikely that the French community would start to use this alternate representation. Also, alternate representations already exists for Latvian, since in principle U+0326 ◌̦ COMBINING COMMA BELOW could be used; it is only by a self-regulation of the Latvian community that this causes no particular problem in practice.

§3. I do believe that the characters proposed in L2/13-129 (WG2 N4466) are essentially doing the same thing (encoding of a combining invariant cedilla), but in way that attempts to minimize the alternate representation problem. However, I think the form of this attempt is very akward:

- the average user of the standard will see those characters as precomposed, but is bound to be disappointed when he discovers that there is no formal canonical decomposition
- it seems inconsistent to encode pseudo-precomposed letters given our strong stance against the encoding of precomposed letters
- it does not offer a general solution for other letters

§4. We could also improve a bit our documentation, in the discussion of commas and cedillas in section 7.1, page 213, following the approach of §1, i.e. be explicit that U+0327 ◌̧ COMBINING CEDILLA is ambiguous, and listing the expected rendering for each base letter.

At the very least, I find the sentence in the last paragraph of page 213 a bit problematic: “The Unicode Standard provides unambiguous representations for all of the forms, for example, U+0219 ș LATIN SMALL LETTER S WITH COMMA BELOW versus U+015F ş LATIN SMALL LETTER S WITH CEDILLA”. *Unambiguous* seems a bit strong, since U+0327 ◌̧ COMBINING CEDILLA is ambiguous (at least for those of us who still believe in canonical equivalence).