

Cedillas and commas below

Eric Muller, Adobe Systems

January 29, 2013

The Marshallese orthography uses the Latin script, including the letters l, m, n, o with cedilla (with the cedilla attached to the rightmost leg of the m and n). It is considered unacceptable to display the cedillas as commas.

The Latvian orthography uses the Latin script, including the letters g, k, l, n, r with comma below. It is considered unacceptable to display the commas as cedillas.

Other orthographies use letters with commas below or cedilla (e.g. French, Romanian) and also require a specific form.

In principle, Unicode makes the distinction between cedillas and commas below, by providing U+0327 ◌̣, COMBINING CEDILLA and U+0326 ◌̤, COMBINING COMMA BELOW.

However, legacy character sets have not always made that distinction (at least in practice). For example, Latvian has historically used legacy ISO character sets that named the characters WITH CEDILLA, with the understanding that the context would imply the rendering appropriate for Latvian, i.e. with a comma below.

This ambiguity in legacy character sets has spilled over to Unicode, via the mapping from those legacy sets to Unicode. The mapping of the legacy WITH CEDILLA characters to the corresponding Unicode WITH CEDILLA characters was quite natural, but it carried with it the

ambiguity of the legacy characters. This can be seen in the code charts, which mentions that for example U+0146 ņ LATIN SMALL LETTER N WITH CEDILLA is to be used for Latvian and shows a representative glyph with a comma below.

Here are the code chart entries for the precomposed characters with cedilla:

00E7 ċ LATIN SMALL LETTER C WITH CEDILLA
≡ 0063 c 0327 ◌̣

1E11 đ LATIN SMALL LETTER D WITH CEDILLA
• Livonian
≡ 0064 d 0327 ◌̣

0229 ě LATIN SMALL LETTER E WITH CEDILLA
≡ 0065 e 0327 ◌̣

0123 ģ LATIN SMALL LETTER G WITH CEDILLA
• Latvian
• there are three major glyph variants
≡ 0067 g 0327 ◌̣

1E29 ħ LATIN SMALL LETTER H WITH CEDILLA
≡ 0068 h 0327 ◌̣

0137 ķ LATIN SMALL LETTER K WITH CEDILLA
• Latvian
≡ 006B k 0327 ◌̣

013C ĺ LATIN SMALL LETTER L WITH CEDILLA

- Latvian

≡ 006C l 0327 ¸

0146 ñ LATIN SMALL LETTER N WITH CEDILLA

- Latvian

≡ 006E n 0327 ¸

0157 ŀ LATIN SMALL LETTER R WITH CEDILLA

- Livonian

≡ 0072 r 0327 ¸

015F ș LATIN SMALL LETTER S WITH CEDILLA

- Turkish, Azerbaijani, ...

- the character 0219 ş should be used instead for Romanian

→ 0219 ş latin small letter s with comma below

≡ 0073 s 0327 ¸

Note that the representative glyphs for c, e, h and s show a cedilla, the representative glyphs for d, k, l, n, r show a comma below, and the representative glyph for g show a turned comma above.

In Unicode 1.0 and 2.0, the representative glyph for N WITH CEDILLA showed a cedilla; this was changed in Unicode 3.0 to a comma below, to accommodate the use of the character in Latvian.

This situation creates a conflict: U+0146 ñ LATIN SMALL LETTER N WITH CEDILLA needs to be displayed with a comma for Latvian and with a cedilla for Marshallese. (Note that combining sequences vs. precomposed characters has no real importance, the same confusion happens with combining sequences.)

For completeness, here are the entries for the characters with comma below:

0219 ș LATIN SMALL LETTER S WITH COMMA BELOW

- Romanian

→ 015F ș latin small letter s with cedilla

≡ 0073 s 0326 ̣

021B ț LATIN SMALL LETTER T WITH COMMA BELOW

- Romanian

→ 0163 ț latin small letter t with cedilla

≡ 0074 t 0326 ̣

The same confusion was problematic for Romanian: the original recommendation for the representation of ș and ț was to use the LETTER WITH CEDILLA characters, and because those characters were often displayed with cedillas, the recommendation was changed to use the WITH COMMA BELOW characters.

The proposal is to solve the current problem in the same way:

- recommend to use COMMA BELOW (combining or precomposed) when a comma is to be displayed
- recommend to use CEDILLA BELOW (combining or precomposed) when a cedilla is to be displayed
- change the representative glyph for the precomposed characters d/k/l/n/r WITH CEDILLA BELOW to show a cedilla
- replace the current annotation “Latvian” or “Livonian” on those characters by annotation similar to the one on U+015F ș LATIN SMALL LETTER S WITH CEDILLA: “the sequence

<00xx, 0326> should be used instead for Latvian/Livonian”

- document the legacy situation and in particular the implications for mappings