Latvian and Livonian glyphs with commaaccent in the Unicode Standard

Roberts Rozis, Tilde
June 07, 2013

The subject of this document is about the glyphs of the Latvian and Livonian languages:

- U+0122, U+0123, currently named LATIN CAPITAL/SMALL LETTER G WITH CEDILLA
- U+0136, U+0137 currently named LATIN CAPITAL/SMALL LETTER K WITH CEDILLA
- U+013B, U+013C currently named LATIN CAPITAL/SMALL LETTER L WITH CEDILLA
- U+0145, U+0146 currently named LATIN CAPITAL/SMALL LETTER N WITH CEDILLA
- U+0156, U+0157 currently named LATIN CAPITAL/SMALL LETTER R WITH CEDILLA
- U+1E10, U+1E11 currently named LATIN CAPITAL/SMALL LETTER D WITH CEDILLA

and their definition in the Unicode Standard.

Background

In 1990 representatives of Latvia addressed the standardization organization ISO asking to provide standardized code places for the glyphs of the Latvian language. At that time Latvia was just regaining its independence and it was not a working member of ISO so we did not have influence about the decisions made by ISO. New standards ISO 8859-4 (was not commonly used) and ISO 8859-13 (standardized in Latvia) established 8 bit code places for the Latvian letters as well as linked them to Unicode code positions, most important, essential and well established.

Unicode standards 1.0 through 6.2.0 (years 1991-2012) define these code points and glyphs as relating to the Latvian and Livonian languages only.

“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.”[1] There is no discussion about the design of these accented glyphs, neither in Latvia or internationally.

It was a correct action to replace the visual representation of the glyphs from having a cedilla to that of having a commaaccent changed from Unicode 1.0 and 2.0 to Unicode 3.0. This was a correct step towards establishing these glyphs as actually used by Latvians/Livonians.

Unicode standards 1.0 through 6.2.0 (years 1991-2012) do not define code points for the Marshallese language. “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.”[2]

“As of 2013, there are no dedicated precomposed glyphs in Unicode for the letters M̧ m̧ N̄ n̄ O̧ o̧.”[3] It means that the Unicodes for the glyphs of Marshallese language may still have to be assigned and standardized. The acceptance and deployment of the new codes in Marshallese language is in development and changes must take place anyway.

General comments about the issue from the Latvian side

Had this discussion taken place in 1991, we would be quite open to various scenarios in the early stages of standardization for the Latvian language. However, the standards for the

[1] "Cedillas and commas below", Eric Muller, Adobe, January 29, 2013

[2] Ibid

[3] http://en.wikipedia.org/wiki/Marshallese_language#Display_issues

Latvian and Livonian language have a 22 year background, and they are deeply applied and integrated in the data, document turnover and in the systems developed and interconnected. Several dozens of state-level databases have required more than a decade to be established and interconnected. Content of libraries has been undergoing digitization, and vast amount of data has been produced. Many operating systems have been built with Latvian UI and Latvian language support. Many old and unsupported systems are also in use in Latvia. The changes may have legal consequences and affect businesses, people in travel and beyond. Uttermost caution and most undisputable arguments are needed to do a change, if really there is no other way, if the consequences of NOT doing the change are of much higher level.

The Latvian side admits that the Unicode Standard has done a great job by establishing and keeping proper code points for the Latvian and Livonian languages, and keeping them intact for 22 years. However, the Latvian side admits its regret that the description of the glyphs has been incorrect for 22 years.

There seems to be unanimity that the actual accented letters in the Latvian and Livonian languages bear the visual appearance of a comma below, also called commaaccent. The composition of the glyphs (we question its relevance in the real systems; how many systems do actually use it as a base of information processing vs. the systems which use Unicodes?) actually consists of the base glyph modified by a U+0312 COMBINING COMMA ABOVE / U+0326 COMBINING COMMA BELOW. These inaccuracies need to be corrected urgently as they prove to potentially create consequent problems.

The Latvian side should not be made a victim of the errors mutually made by ISO and Unicode or caused by legacy issues resulting in incorrectly named/described the glyphs of Latvian and Livonian language in 1991. If the established glyphs conflict with those of a new language, solutions must be sought for the new language instead.

Comments about some of the possible approaches enlisted by Eric Muller of Adobe, based on the viewpoint of the Latvian side:

Approaches (1-4) proposed by Eric Muller of Adobe	**Comments**
1) do nothing	Agree. This approach does create confusion and may create even bigger confusion in future.
2) declare that comma below and cedilla are essentially two different renderings of the same abstract character.	Agree. This contradicts the real world and destabilizes the composition.

Approaches (1-4) proposed by Eric Muller of Adobe	**Comments**
3) in an effort to mitigate the impact on existing data, leave the expected rendering of e.g. <U+0146 ņ LATIN SMALL LETTER N WITH CEDILLA> as a comma, and encourage <n, U+0327 ◌̧ COMBINING CEDILLA> to be rendered as a cedilla. This runs afoul of the canonical equivalence of those two sequences.	Disagree. If the Unicode standard admits that the 12 glyphs in question standardized 22 years ago and used by the Latvian language ever since, have different appearance, this should be corrected in the Unicode standard, including the correct composition sequences with U+0326 COMBINING COMMA BELOW / U+0312 COMBINING TURNED COMMA ABOVE. The canonical equivalence of both sequences will then be ensured by adjusting the glyph names and composition data of U+0122, U+0123, U+0136, U+0137, U+013B, U+013C, U+0145, U+0146, U+0156, U+0157, U+1E10, U+1E11 to the real world – see explicit definitions below.
4) declare that comma below and cedilla are two different characters, and that rendering one by the other is not correct. This approach is of course problematic for communities which have used the WITH CEDILLA characters and want to see a comma, as it changes the representation of existing text. This affects mostly Latvian and Livonian.	Caution! Impact of the changes either way must be seriously examined. In case of Latvia, it affects >2M speakers of the Latvian language, and the timespan of 22 years of standardized use of the standardized Latvian language in Unicode should be respected. Considering the integration of the Latvian language in the information systems in Latvia and in European information networks, the suggested changes would be dramatic for the Latvian side.
I personally think that approach 4), however painful it may be, is the only one that has the potential of leading to reliable ecosystem. With a bit more details:	Possibly, details to be examined
- recommend to use COMMA BELOW (combining or precomposed) when a comma is to be displayed	Right
- recommend to use CEDILLA BELOW (combining or precomposed) when a cedilla is to be displayed	Right
- change the representative glyph for the precomposed characters d/k/l/n/r WITH CEDILLA BELOW to show a cedilla	**NO!** We have to respect that these glyphs were assigned to the Latvian language 22 years ago.

Approaches (1-4) proposed by Eric Muller of Adobe	**Comments**
- replace the current annotation “Latvian” or “Livonian” on those characters by annotation similar to the one on U+015F ş LATIN SMALL LETTER S WITH CEDILLA: “the sequence <00xx, 0326> should be used instead for Latvian/Livonian”	**NO!** What systems do actually use annotations? 99% systems use code points, and the proposed workaround is of no help to them.
- document the legacy situation and in particular the implications for mappings	WILL NOT REALLY HELP.
Reverting the images of cedilla-accented glyphs to show cedillas in case of 4 or 12 glyphs.	**NO!** Unicode Consortium made a correct step towards changing the visual representation of the Latvian glyphs to display with commaaccent. Instead, additional action needs to be taken to finally and correctly establish the Latvian and Livonian glyphs, their names, composition and appearance.
Keeping L/l/N/n-accented glyphs accented with cedilla and introducing new glyphs and code places for the Latvian language	**NO!** The Unicodes for the Latvian language were established by ISO and Unicode, and are widely used. For 22 years there was no real conflict. This ecosystem should not be destroyed.

Suggestion to the Unicode Technical Committee:

1. Do not change any established code assignments. Keep all the assignments of Unicodes to the glyphs of Latvian and Livonian language intact.
2. Change the descriptions of the glyphs to reflect the actual shapes and usage, as follows:

- U+0122 (Adobe glyphname Gcommaaccent) to **LATIN CAPITAL LETTER G WITH COMMA BELOW.** Composition **0047 0326**
- U+0123 (Adobe glyphname gcommaaccent) to **LATIN SMALL LETTER G WITH TURNED COMMA ABOVE.** Composition **0067 0312**
- U+0136 (Adobe glyphname Kcommaaccent) to **LATIN CAPITAL LETTER K WITH COMMA BELOW.** Composition **004B 0326**
- U+0137 Adobe glyphname kcommaaccent) to **LATIN SMALL LETTER K WITH COMMA BELOW.** Composition **006B 0326**
- U+013B (Adobe glyphname Lcommaaccent) to **LATIN CAPITAL LETTER L WITH COMMA BELOW.** Composition **004C 0326**
- U+013C Adobe glyphname lcommaaccent) to **LATIN SMALL LETTER L WITH COMMA BELOW.** Composition **006C 0326**
- U+0145 (Adobe glyphname Ncommaaccent) to **LATIN CAPITAL LETTER N WITH COMMA BELOW.** Composition **004E 0326**
- U+0146 Adobe glyphname ncommaaccent) to **LATIN SMALL LETTER N WITH COMMA BELOW.** Composition **006E 0326**
- U+0156 (Adobe glyphname Rcommaaccent) to **LATIN CAPITAL LETTER R WITH COMMA BELOW.** Composition **0052 0326**
- U+0157 Adobe glyphname rcommaaccent) to **LATIN SMALL LETTER R WITH COMMA BELOW.** Composition **0072 0326**

- U+1E10,) to **LATIN CAPITAL LETTER R WITH COMMA BELOW.** Composition **0044 0326**
- U+1E11) to **LATIN SMALL LETTER R WITH COMMA BELOW.** Composition **0064 0326**

3. Keep in the Unicode standard the proper and actual samples of the glyphs as used by Latvians in 2013. Most Adobe/Lintoype/Tilde Pro fonts will apply. Do not revert back the graphic representations to cedillas.
4. Include new and unambiguous forms for the Marshallese language,

   LATIN CAPITAL LETTER L WITH CEDILLA. Composition 004C 0327
   LATIN SMALL LETTER L WITH CEDILLA. Composition 006C 0327
   LATIN CAPITAL LETTER N WITH CEDILLA. Composition 004E 0327
   LATIN SMALL LETTER N WITH CEDILLA. Composition 006E 0327

along with other glyphs missing.

5. As the Marshallese language is being standardized and included in the glyph definition of the Unicode Standard, introduce all the needed new glyphs and encourage the community of Marshallese people accept and apply the new standard in whole.