SC2/WG2 N4448

L2/13-053

Transition Considerations

March 29, 2013

submitted by Lisa Moore

The Latin script is used to write texts in a wide variety of languages, with much variation in typographical usage. Letters with cedillas and commas below in some cases are substituted one for the other, depending on the font technology that is available. However, there are certain languages with well-defined practices that have strong requirements for one form or the other.

The goal of the Unicode Standard is to provide an unambiguous representation for all of the forms.

In ISO legacy character sets there have been at times a single character that could take two different representations (in Turkish and Romanian, for example). This practice created great ambiguity in legacy text data, and it carried over into early versions of the Unicode Standard. The result was an incorrect form often being used in Romania. At the request of the Romanian user community, in Unicode Version 6.0, the forms to be used for each language were clearly spelled out. Today, there is a clear form for Romanian and a clear form for Turkish. Mapping between legacy character sets and Unicode for Romanian and Turkish text data requires specific implementation choices to ensure that the appropriate Unicode character is used.

Legacy ISO Latvian character sets created a different kind of ambiguity. The character name WITH CEDILLA was used, while the actual character image was that of a COMMA BELOW. In an effort to maintain compatibility with Latvian legacy practices, the Unicode Standard, Version 3.0 changed its representative glyph for d, k, l, n, and r WITH CEDILLA to show a comma below, and g WITH CEDILLA to show a turned comma above, while maintaining a decomposition to U+0327 COMBINING CEDILLA.

This creates confusion as to the identity of the characters, and makes it impossible to actually represent the true d, k, l, n, r, or g with cedilla. Such is the situation now impacting the Marshallese as they move their computing environments to be based on Unicode, only to discover that there is no way to actually represent n with cedilla.

The Unicode Technical Committee is considering returning to an unambiguous representation of characters either WITH CEDILLA or COMMA BELOW. If this action is taken, the correct way to represent Latvian letters with comma below will be <00xx, 0326 COMBINING COMMA BELOW>. Thus, the Latvian letter ņ would in the future be represented by <006E LATIN SMALL LETTER N, 0326 COMBINING COMMA BELOW>.

If this change is made, there would be period of time during which implementations would need to support both forms of representing Latvian letters (this would be similar to the transition experienced by the Romanian community). During the transition time, there could be challenges with text searching and sorting, legacy text might not display correctly, and character mapping implementations would not work correctly unless they were updated.

Over time, all users of Latin would benefit from greater clarity in the identity of the characters their language uses, and all forms of Latin letters that use cedilla or comma below could be represented. There would be a clear and unambiguous form for Latvian and Livonian, and a clear and unambiguous form for Marshallese.