How to Solve the the Problems Addressed by Lithuania in WG2 N4191

Karl Pentzlin – 2012-01-26

> » *Without the work of the Unicode Consortium, our [the Cherokee] language would not be poised in the great position it is technologically. Our unique syllabic writing system is being used on smartphones, computers, and social media all over the world now. It … strongly continues in all the modern media communications. For that, we thank you all greatly.* «
> — ᏫᏂᏕ ᎪᎳᏍᏉ (Roy Boney), 2011-12-02 on the Unicode mailing list

Obviously, the Lithuanian community is not likewise happy regarding the support of their language by Unicode.

In WG2 N4191, they request single code points for 35 accented letters as precomposed characters, presenting evidence that otherwise it is not possible to use these characters within applications using Unicode (especially including such widespread applications as Microsoft Internet Explorer or Firefox).

The documents WG2 N4187, N4188, N4189, N4190, and N4192 all express support by several parties, including the Prime Minister of Lithuania as well as persons who are known to be aware of Unicode details.

The 35 characters requested in WG2 N4191 are:

Ą́ ą́ Ą̃ ą̃ Ę́ ę́ Ę̃ ę̃ Ė́ ė́ Ė̃ ė̃ i̇̀ i̇́ i̇̃ Į́ į̇́ Į̃ į̇̃ J̃ j̇̃ L̃ l̃ M̃ m̃ R̃ r̃ Ų́ ų́ Ų̃ ų̃ Ū́ ū́ Ū̃ ū̃

Here, they are written as they in fact are encoded in Unicode, using fully decomposed character sequences, i.e. compliant with Unicode's Normalization Form D (NFD). The font used here is Doulos SIL (obtainable free of charge at http://scripts.sil.org/cms/scripts/page.php?item_id=DoulosSILfont). This text was edited using Microsoft Word 2010 (the character list was edited in Emurasoft's EmEditor, using the keyboard driver from http://www.europatastatur.de , and copied/pasted into this document), and this document was finally generated using Adobe Acrobat CS5.

Thus, it is proven that it is possible to represent the requested Lithuanian characters, at least if you use the adequate font and the adequate software. (And you need an adequate input method. In fact, the new German keyboard standard DIN 2137:2012 to be published in April allows the input of all official E.U. languages using the Latin script, *including the Lithuanian letters requested in WG2 N4191*. However, as a side remark, this standard cannot be implemented using the Microsoft Keyboard Layout Creator only.)

While there may be room for subtle improvements (like the too high position of the tilde over the dotted E, or the side bearings of narrow letters with tilde as in the sequence Į̃į̇̃J̃j̇̃l̃), the characters are clearly readable. In fact, Doulos SIL yielded one of the best results within the arbitrary selection of fonts I had tested.

A similar good result was obtained with Microsoft's Times New Roman (on Microsoft Windows Vista):

Ą́ ą́ Ą̃ ą̃ Ę́ ę́ Ę̃ ę̃ Ė́ ė́ Ė̃ ė̃ i̇̀ i̇́ i̇̃ Į́ į̇́ Į̃ į̇̃ J̃ j̇̃ L̃ l̃ M̃ m̃ R̃ r̃ Ų́ ų́ Ų̃ ų̃ Ū́ ū́ Ū̃ ū̃ – Į̃į̇̃J̃j̇̃l̃

Compared with these, the result using Adobe's Minion Pro (as delivered with Adobe CS5) is catastrophic:

Ą́ ą́ Ą̃ ą̃ Ę́ ę́ Ę̃ ę̃ Ė́ ė́ Ė̃ ė̃ i̇̀ i̇́ i̇̃ Į́ į̇́ Į̃ į̇̃ J̃ j̇̃ L̃ l̃ M̃ m̃ R̃ r̃ Ų́ ų́ Ų̃ ų̃ Ū́ ū́ Ū̃ ū̃ – Į̃į̇̃J̃j̇̃l̃

The eventual goal of the Lithuanian request is to get the possibility to handle their characters with the same ease and reliability as this is possible for accented Croatian (referring to U+0200…U+0217) or for Polytonic Greek. **This request is fully legitimate and has in fact to be supported by all means.**

In fact, to get standardized code points for precomposed characters is a way to accomplish this. With the font technology available in the 1990s on common office equipment, that was the way to go, thus at that time it was the right way to go to encode precomposed characters. This changed with the necessity to provide more advanced technology required for Indic and other recognizedly complex scripts, thus in 1996 it was decided that the single point encoding of precomposed Latin letters (for Yorùbá, in that case) was no longer necessary.

However, it is a shame for the whole software industry that, 16 years after that decision, the treatment of Latin letters still is as catastrophic as WG2 N4191 shows.

This is not the problem of Unicode or ISO/IEC 10646. In fact, it is possible to get the desired results with selected software and selected fonts, as the presentation of the requested letters above shows.
The sequence U+0041 U+0301 U+0328 e.g. is a perfectly working encoding of the letter Ą́, and the placements of the diacritical marks are sufficiently described by their Canonical Combining Class values.

However, it cannot be the task of font designers to ensure the correct placement of all diacritical marks. They have to provide glyphs, nothing else. They may provide special glyph selection and placement rules for special cases where they decide that this is advantageous for their special typographical issues, but not for any possible combination which may occur with any language they do not specially care of.
This has to be handled *reliably* by the basic software, which simply has to be compliant with the sufficient specifications given in ISO/IEC 10646 and Unicode.

The Lithuanian request has to be addressed by a serious commitment by each of the main players in the software industry that they will complete the necessary work on the Latin script within adequate time.

The UTC and SC2/WG2 have done their work by having provided an encoding which works, thus there is no need to fall back behind the Yorùbá decision of 1996 and to encode precomposed characters only to support software technology which corresponds to the state of the art of that time but is still delivered today.

Latin IS a complex script, not simpler than Devanagari, and the software industry was able to provide solutions for Devanagari to a degree that the UTC and SC2/WG2 are not bothered regularly with requests to assign single code points e.g. for the numerous consonant/vowel combinations.

Now, it is up to the software industry to prove that they are able to do the same for the Latin script.

(By the way, what is said about the Latin script here applies to the Cyrillic script also.
See e.g.: http://www.pentzlin.com/Orok.html).

Appendix

For the case that the PDF file does not display the same as on the computer where it was generated in all cases, screenshots were taken from a display of the PDF file for the character sequences and added here.

Doulos SIL:

Á ą̃ Ą ą̃ É ę̃ Ę ę̃ É ė́ Ė ė̃ ì í ĩ Į į́ Į̃ į̃ J̃ j̃ L̃ l̃ M̃ m̃ R̃ r̃ Ų́ ų́ Ų̃ ų̃ Ú ú Ū ũ

Į̃ į̃ J̃ j̃ l̃

Times New Roman:

Á ą̃ Ą ą̃ É ę̃ Ę ę̃ É ė́ Ė ė̃ ì í ĩ Į į́ Į̃ į̃ J̃ j̃ L̃ l̃ M̃ m̃ R̃ r̃ Ų́ ų́ Ų̃ ų̃ Ú ú Ū ũ – Į̃ į̃ J̃ j̃ l̃

Minion Pro:

Á ą̃ Ą ą̃ É ę̃ Ę ę̃ É ė́ Ė ė̃ ì í ĩ Į į́ Į̃ į̃ J̃ j̃ L̃ l̃ M̃ m̃ R̃ r̃ Ų́ ų́ Ų̃ ų̃ Ú ú Ū ũ – Į̃ į̃ J̃ j̃ l̃