TO: UTC **L2/13-164**
FROM: Ralph Cleminson and David Birnbaum (via Debbie Anderson, SEI, UC Berkeley)
TITLE: Expert Feedback on Cyrillic proposals (L2/13-153, L2/13-139) and Slavonic Punctuation proposal (L2/13-140)
DATE: July 25 2013

Below are comments from Ralph Cleminson and David Birnbaum, both of whom have contributed or co-authored other proposals for Cyrillic and/or Glagolitic and have long been involved in digital projects.

1. L2/13-153 Proposal to Use Standardized Variation Sequences to Encode Church Slavonic Glyph Variants in Unicode

CLEMINSON:
We've been through this before, ad nauseam. These are glyph variants. It is perfectly proper for palaeographers and others to distinguish between them when creating electronic texts, but it is not the function of Unicode to do so, and there are other mechanisms for this. (See my article on the subject, http://manuscripts.ru/mns/docs/unicode_cleminson.pdf) Variation selectors are to be used when a particular form of a character has to be used in a certain context (and to use a different form would be wrong), but the context is not determined by position (as in Arabic, for example), but by some other factor such as grammar, so that no application could determine purely from a string of plain-text characters which variant is correct. This is the case, as I understand it, in Mongolian, where anyone who is fully literate knows which variant is required in any given instance, and therefore that information, not otherwise available in plain text, is supplied by a variation selector.

This is not the case here. The proposers admit that the distribution of the glyphs is totally haphazard. It would make no difference whatsoever if they were redistributed. What they are trying to do is encode fancy text, and they should be told in no uncertain terms not to present this proposal to Unicode again.

BIRNBAUM:
I've read the proposals and Ralph's response [above], and he and I have corresponded a bit behind the scenes, and come to somewhat different conclusions. We hope that our input will nonetheless prove useful when the UTC meets and evaluates the proposals.

Ralph disagreed with the appropriateness of the "variation sequences" proposal because it amounted to encoding glyphic variants that are neither systematic (as, e.g., Arabic positional variants) nor orthographically obligatory (it would not be incorrect, from the perspective of the orthographic system, to write the base form instead of the variant, although it might be undesirable in a transcription). Neither of us is knowledgeable about Mongolian writing, for which, if I understand correctly, the concept of variation sequences was first introduced, so I'm not sure I understand the requirements and assumptions correctly. I think the relevant documentation is at http://www.unicode.org/faq/vs.html. In particular, "[f]or historic scripts, the variation sequence provides a useful tool, because it can show mistaken or nonce glyphs and relate them to the base character" seems applicable to the current situation, and the reference to "nonce" suggests, at least to me, that the variant need not be systematic and that it need not be obligatory in the position in which it occurs. Or, more precisely, it need not be orthographically obligatory, but it might be regarded as obligatory by transcribers because they prioritize preserving this type of variant during transcription.

Ralph is correct in stating that these are glyphic variants, but he and I understand differently whether that's an appropriate use for variation sequences. One of the issues is, I think, that when we transcribe existing written sources that make distinctions, we may want to retain those distinctions because they are often part of the information included in a manuscript description. That's a different set of circumstances than creating new documents in a modern writing system. Whether the distinctions should be considered paleographic (handwriting, which may have small idiosyncratic variation) or orthographic (deliberate use of conspicuously different abstract letterforms representing the same letter) isn't always clear. However we understand the distinctions, though, they can be encoded through markup, so the issue becomes whether the UTC intends variation sequences to be used as a way of also enabling their encoding in plain text. Ralph's response asserts that such an encoding would be an abuse of the character/glyph distinction, while it seems to me as if variation sequences are intended to serve as a compromise that permits such abuse by mandating a conformant way to engage in it. Whether either of us has understood the situation properly is probably best determined by the UTC, and especially by those members who understand the Mongolian precedent.

2. L2/13-139 Proposal to Encode Combining Half Marks Used for Cyrillic Supralineation

CLEMINSON:
This proposal depends on a single sentence, "a tilde differs from a titlo both in visual appearance and in function." If you accept that, you have to accept the proposal, otherwise not. Appearance in fact may be a typographical issue (and there is indeed considerable variation in the appearance of a titlo), so it very much depends on current thinking on the functions of combining diacritics. It is nevertheless worth noting that the standard distinguishes between tilde and titlo over a single character, so it would be inconsistent not to allow us to make the same distinction over multiple characters.

BIRNBAUM:
Concerning this "half marks" proposal, I've also fretted over how to encode long titlo, that is, titlo that spans multiple characters. Their proposal is sensible, and it's consistent with what has been done elsewhere in the standard. My instinct would have been to propose three characters: beginning, middle, and end; they propose only beginning and end, specifying that those don't have to be contiguous, and that "middle" is implied over anything (zero or more characters) between the two. I have no objection to that; it's economical, although it achieves its economy

by assigning meaning to a zero sign, which means that interpretation becomes more complicated (mechanically, anyway, even if not conceptually).

Ralph has drawn my attention to an important detail: the titlo, no matter how many characters it spans, is fundamentally a single character. There is precedent, cited in the proposal, for this type of approach to long diacritics, and it's hard to come up with a graceful alternative strategy in plain text. In my own work, when I encounter a titlo that spans multiple base characters, I use the available (narrow, unitary) titlo character, put it over (= after, in the backing store) one of the base characters, and regard it fundamentally as a diacritic on the word, rather than on a particular character in the word. That's not wholly satisfactory, either, especially for rendering.

I am concerned, with respect to the first two proposals [L2/13-153 and L2/13-139], about whether there will ever be any software support. That concern is more acute in the case of the long titlo than the variation indicators because the long titlo spanning more than two characters means that there will be no overt indication of any superscript mark over the middle characters. Software support isn't the business of the UTC, to be sure, and users could address the issue by piping their documents through a routine that would map from the character stream to a glyph stream, but what we really want is a font engine (probably OpenType) that does that quietly and automatically. From an informational perspective, we want to regard middle characters (where a titlo spans three or more characters) as under the titlo, even if no titlo-associated Unicode character follows them immediately in the backing store.

When I consider the first two proposals together, I wonder whether the beginning and ending titlo characters should be variation-indicator variants of base titlo, instead of fully independent base characters. There seems to be precedent elsewhere in Unicode for encoding them completely separately from "normal" titlo, but they might alternatively be regarded, fundamentally, as different graphic representations of the meaning of the core titlo character.

3. L2/13-140 Proposal to Encode a Slavonic Punctuation Mark

CLEMINSON:
This character certainly exists, and has the function they describe. Unless there is a close analogue somewhere in the standard (and I can't find one), I am broadly in favour of adding it. You may wish to reconsider the name for it, since if something of this kind exists in other scripts (and I don't know whether it does, but it might), one would not wish to deter non-Slavs from using it.

BIRNBAUM:
The "spear" proposal seems straightforward and sensible. Like Ralph, I'm unaware of any existing character in Unicode that fills this role, although I haven't looked systematically.