Subject: Re: White paper from Dr. Raymond Doctor - Feedback

Date: Saturday, August 13, 2011 10:47:58 AM PT

From: Nicholas Ostler (sent by Nicholas Ostler <nicholas.ostler@gmail.com>)

To: Naela Sarras

CC: Steve Sheng, Francisco Arias, Dennis Jennings, Andrew Sullivan, doc, akshatj@cdac.in, nehag@cdac.in

On 10/08/2011 19:52, Naela Sarras wrote:

> Dear colleagues,
>
> The attached white paper was prepared by Dr. Raymond Doctor and shared with us during the Devanagari team meeting in Pune, India on 21&22 July 2011. Dr. Raymond updated the document based on the discussion we had during the face-to-face meeting in Pune.
>
> The variant project team members included in this message are kindly requested to review the document provide any comments, critique, or suggestions to Dr. Raymond Doctor. Specifically, I am asking Andrew Sullivan and Nicholas Ostler to review the document and give feedback . Please provide your feedback by close of business (PST time) on Monday, 15 August 2011. Once the project team has provided their feedback to this white paper, the document will be shared on thevip@icann.org mailing list.
>
> Thank you,
> Naela

I found this an excellent, and well-founded piece of work, which brings the linguistic aspects of Variant-hood into a coherent pattern, taking into account the major features of all the five scripts we are discussing. 2.3 is especially useful: it succeeds in its aim of offering a unified approach to variants, I think.

Section 2.1
There is some scope for confusion here, since the idea of "abstract character" seems first to be equated with a phoneme, such as /p/, and then equated with a glyph. But Unicode 6.0 (para 3.4 D7) explicitly denies that "abstract character" is a glyph. It seems to be what a glyph represents, but then only within the context of a given writing system, (not universally, as in phonetics or phonology). But the only property the abstract character shares with the phoneme is the fact that it is an idealized entity, used theoretically to explain a set of realizations which are closer to actual experience. It is irrelevant to our concerns that the single phoneme /p/ can be represented a variety of different scripts with different codepoints.

(Doctor, however, later (page 4) shows himself entranced by the fact that there is a parallelism between phonetics, and the coding of individual phones in the various Brahmi code pages. This is interesting, but again: this is irrelevant to the quest for variants.)

On the other hand the equivalence of "abstract character" to "glyph", a significant type of written shape, seems to be correct, or at least harmless. draft-ietf-appsawg-rfc3536bis-02 (also quoted with approval here) doesn't actually define "abstract character", but its characters are clearly the stuff of writing systems, not speech sounds. What we want to talk about is the language-entities that correlate with Unicode

code-points. These are the abstract characters, for our purposes. They each have a codepoint reference, and a name, and a physical display form. Their actual phonetic realization is irrelevant.

The discussion under a. (viz that the same phonetic vowel represent as æ "ash" in IPA correlates with two distinct glyphs, 090D in Hindi and 0972 in Marathi and Konkani) seems irrelevant. These have distinct code points, and distinct names, and distinct appearances, and even if an IPA transcription would find it hard to distinguish them, so what? They are not candidates to be variants. Perhaps Doctor's implicit point is that a user, sitting with an input device dedicated to Hindi, Marathi or Konkani, might not have the means to discriminate which she was inputting (as could well happen with the other examples Doctor gives, the different characters used for /i:/ in Arabic as against Persian/Urdu - all the characters known as yeh, or alif maksura). This may be inconvenient for such a user, but the answer seems to lie in giving her a more powerful input device, rather than pretending that two distinct characters are variants of one another (as between languages).

By contrast, the points about characters with diacritics, which can be reached either directly, or through combining the simple character with a separate diacritic, is pertinent to variants. In these cases, we have two input methods which result in what is (for all purposes) the same character. This situation he call Normalization, and it is clearly important that each of these doublets is explicitly normalized into a single (preferred) variant.

[The last line contains a mistake as it stands. The cited codepoints refer to the Devanagari equiv. of Q, written using KA with a nuqta dot under it. But the first character displayed is 095C ड़ DEVANAGARI LETTER DDDHA, while the second is 0921 ड DEVANAGARI LETTER DDA with the nuqta dot 093C. This is subsequently corrected in the discussion on p. 22.]

On page 4, there is an omission in the last sentence in the main text on the page.
"This will become more acute when South Africa which recognizes"
It seems to suggest that South Africa's 11 official languages have a special need to mix items from various scripts' codepages, but no evidence is cited. I should be surprised if so, since all the languages use Latin script. The only complications are;
I. Sepedi and Setswana have an s with a caron 0300 over it, to palatalize it.
II. Venda adds diacritics (a subscript circumflex 032D to t, d, l and n, making them dental not alveolar) as well as a superposed dot 0323 to n, representing velar nasal.

Page 5:
1."coeval" does not mean "equivalent", which is (I think) what the author intended.
5. 02BC is the apostrophe (or one code for it). The author believes this is necessary for Boro, Dogri and Assamese. It must be presumed that the arguments for this are comparable to those for the same apostrophe (02BC) in Ukrainian and Belarusian - copiously discussed on the Cyrillic

VIP list. I cannot comment on the need for this glyph in the Inidan languages, since I know little of them, but I have done so already "On U+02BC" as for Cyrillic.
4. The request that language tags be implemented is simply an appeal, not really argued for. I find it unconvincing. The full list of identities of languages is not complete (and ultimately will be prey to becoming a political football). Security considerations make it inadvisable to distinguish realizations of a single codepoint, realizations which differ only because someone somewhere has assigned the string in which they occur a different language-tag. There is no reason to assume that Hindi-using organizations will snap up all the good names before Marathi, Konkani or Dogri get a chance.

2.3
This is where it gets interesting. Doctor seems to have provided a good characterization of the key issues that arise for all the major scripts we are considering. As such, this document seems like required reading for all us variant-walas.

2.3.1
This proposal of "archi-variant" and "variant-eme" as new technical terms (possibly synonymous?) seems gratuitous, and probably of interest only structural linguists. "Types of variant" would seem a reasonable substitute. Doctor does not actually use either of these terms much, nor does he need them.

Page 11. Doctor has a problem with his three variants names for a single kind of script (Abugida, Alphasyllabary, Akshara) since he is not subsequently consistent when he refers to it. (Devanagari is the only instance of this script-type which we are currently considering.) Strangely, I think "Abugida" (derived from the Ethiopic name for alphabet) is the best established in the linguistic literature.

2.3.3 The Problem of the Preferred Variant
Doctor maintains that this only arises as a result of particular word-context. (This is comparable with arguments that have raged about whether ë and e should distinguished in Cyrillic.) Since this is comparable with "color/colour case", (i.e. string-level variants, usually ruled out of court as a potential variant in our sense), we might decide that this is irrelevant to our current concerns.

However, he does talk of "spell-variants at character level" in this table of page 19, and claims that they exist in all types of script except the Abugida/Alphasyllabary/Akshara. Presumably he is referring to cases where different languages discriminate in favour of one variant rather than another (as yeh in Arabic/Persian, æ "ash" in various Indian languages), so it is language-context rather than word-context which establishes a preference. Arguably, in these cases, registries in all language areas would accept the same set of Unicode variants (called for some reason "alternants" on p. 22) ; but each registry (with its own preferred language) might discriminate in favour of a different one when a gTLD is to be registered in their own domain. So Arabic regustries would block or disprefer in some way gTLD with Persian yeh, and vice versa.

That is all my comments.
In sum, it appears to me that the tabular approach adopted here (in 2.3) might be adopted with advantage by all the groups. This would enable them to compare more directly the kinds of distinction and decisions that each group is proposing.

--
Nicholas Ostler

nicholas@ostler.net
+44 (0)1225-852865, (0)7720-889319

Chairman: Foundation for Endangered Languages
www.ogmios.org

Author: Empires of the Word (2005),
Ad Infinitum (2007), The Last Lingua Franca (2010)
www.nicholasostler.com