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.

Many thanks for your kind remarks as well as the detailed reading of the white paper, which must have taken quite some time. I have tried to answer to some of your comments not as a justification but more as a eans of explaining the underlying motivation.

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 code point reference, and a name, and a physical display form. Their actual phonetic realization is irrelevant.

Both you and Andrew have pointed this anomaly out. I have based myself on *draft-ietf-appsawg-rfc3536bis-02,* which states as under

```
glyph
A glyph is an abstract form that represents one or more glyph
images. The term "glyph" is often a synonym for glyph image,
which is the actual, concrete image of a glyph representation
having been rasterized or otherwise imaged onto some display
surface. In displaying character data, one or more glyphs may
be
selected to depict a particular character. These glyphs are
selected by a rendering engine during composition and layout
processing. <UNICODE>
```

The way I read (past tense) this text has led me to make this statement. However if I have misunderstood the text, I stand corrected.

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).

I would like to point out that this was written as an evidence of the fact that the basic principle of Uniqueness which Unicode advocates is respected very often more in the breach. As per Unicode tow characters which are “allographic representations” of an abstract character shall not be admitted. As a matter of fact there are three ways of writing the Devanagari /la/ depending on where the stem is placed (mid, half-mid, extreme right) Unicode has accepted the extreme right case. The same occurs in the case of 090D in Hindi and 0972 in Marathi and Konkani. These are allographs, but the Marathi and Konkani speakers have demanded allogrph2 and not accepted allograph1. This is the point I was trying to make

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 code points refer to the Devanagari equiv. of Q, written using KA with a nuqta dot under it. But the first character displayed is 095C □ DEVANAGARI LETTER DDDHA, while the second is 0921 □ DEVANAGARI LETTER DDA with the nuqta dot 093C. This is subsequently corrected in the discussion on p. 22.]

Many thanks for the sharp reading and the correction

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.

Agreed.The foot-note was just to point out that as more and more languages fall into the scope of IDN’s, complexities will arise and there is an urgent need to separate LANGUAGE and SCRIPT

Page 5:
1."coeval" does not mean "equivalent", which is (I think) what the author intended.

Many thanks once again: Coeval implies coetaneous , contemporaneous

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 Indian languages, since I know little of them, but I have done so already "On U+02BC" as for Cyrillic.

I agree. I think somewhere ICANN should distinguish between a “diacritic marker” and a character that looks like a diacritic marker but has linguistic pertinence. This is why I have made on case for Dogri where the apostrophe marker (0027) functions as an abbreviation marker (like in French le+eau or popular German Elena’s)

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 code point, 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.

The white paper was never meant to be an “appeal” for Language Tags. However the issue needs to be discussed and sorted out once and for all. It is a social as well as cultural issue to degrade language and place it on a rank lower to script, with script gaining primacy. While Unicode’s preoccupation is with Script (and rightly so), somewhere language has to be given priority or else languages sharing a script. There are today 71936894 speakers of Marathi and over 366 million first-language speakers; an additional 121 million second-language speakers for Hindi, not to mention the

7 other official languages sharing Devanagari[1]. While I agree that it could become a « political football » I also feel that mixing all languages sharing a common script into one melting pot is just as bad a policy.

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.
As a linguist, the temptation of establishing parallels in structuralism was too tempting. Hope you'll understand the underlying motivation
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.
Agreed but as I have pointed out Devanagari and all Brahmi related scripts are simply NOT abugidas, since the notion of the basic building block :"akshar" is much more than that. I still feel that abugidas and alpha-syllabaries need to be separated out, and that is why in my analysis, I have kept Abugidas distinct from Alpha-syllabaries.

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 registries would block or disprefer in some way gTLD with Persian yeh, and vice versa.
Many thanks for this observation. The problem is more acute however:

Constraining rules can be deployed for Farsi where the plural marker "he" needs to be separated from the root. These conditions are specific and can be handled by a constraining rule which states that introducing ZWJ only when the root word ends in x and is followed by "he" and "he" alone would be treated as valid.

Off the record the same situation occurs in Urdu cf. our policy for Urdu where writing for example aam aachaar (mango pickle) as one word disrupts totally the reading of the word since the final meem of aam is conjoined to the initial alef madd of aachaar. We have proposed a "hyphen" in this case.

However in the case of Noun paradigms of languages such as Telugu, Nepali, Malayalam and to a lesser extent Oriya, ZWNJ placed after a Halanta (Virama) allows for disambiguation by retaining the identity of the root word and its suffixal inflection

Nepali: mananku needs necessarily a HALANTA+ZWNJ after manan to ensure that the final consonant of manan does not join to the suffx "ku". The same is the case in Oriya: tebalku" Since in theory all and every noun admits the suffixal form and the number of suffixal paradigms I very high, to

[1] http://www.oclc.org/languagesets/educational/languages/india.htm

admit or not to admit the ZWNJ within an IDN is a major issue, since admitting it can create a large number of spoofing issues. The call to be taken is similar to the apostrophe in English and French. Our Nepali colleague has been requested to propose some solution for this problem which seems to me to be extremely complex, since constraining rules are not applicable..

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