From: Andrew Glass (WINDOWS)
Sent: 04 February 2014 18:14
To: 'Kiyonori NAGASAKI'; Ken Lunde
Cc: suzuki toshiya; Taichi KAWABATA; Deborah W. Anderson; Anshuman Pandey; Michel Suignard; Roozbeh Pournader
Subject: Concerns about encoding variants of matras in Siddham

+Roozbeh

Dear Nagasaki-san and Suzuki-san,

Thank you for the new document that has the examples of the different vowels with base characters other than h. Thank you also for the excellent summary document from the Tokyo ad-hoc. To re-cap, the core of the agreement at the ad-hoc meeting was the following. In order to be considered a variant the following three conditions should be met:

- Both forms are semantically distinct in a logographic context
- Both forms cannot be algorithmically derived by context
- Both forms co-occur in a single source

With respect to the combining vowel signs I would like to make the following points.

1. I agree the syllables HU, HUU, HUM, HUUM, can be semantically distinct when used with the two variants of U and UU. Also, these syllables cannot be derived from context and do occur in a single source. This much is demonstrated in your document (L2/14-055).
2. What is not clear from L2/14-055 or from your additional document (pu_yu) is that ALL other combinations of base consonant plus the vowels U and UU in their cloud and warbler forms are also logographically distinct, cannot be algorithmically derived from context, and co occur in a single source.
3. What you are requesting are variant forms of the combining vowels U and UU in order to use them productively with H to produce the syllables HU, HUU, HUM, HUUM. Therefore, the request is much broader than the use case, because while the alternate forms of U and UU could be used with H to form the distinct syllables having semantic distinction, there is nothing in the Indic shaping mechanism that would prevent them from also being used productively with other bases. There are consequences to that approach, as follows:
    - FONTS – A Siddham font should have the ability to combine any consonant with any vowel including the vowels U, UU in both cloud and warbler forms. As you point out in your doc, L2/14-055 (page 6), the cloud form does not occur with some base consonants, e.g., DHU. A font developer is not in control of which sequences are rendered using the font and so a font developer would most likely have some default display form for the cloud form of DHU even though such a thing is said not to exit
    - SHAPING – Indic shaping is done based on properties. A user can supply any combination of code points to the shaping engine. While a shaping engine could prohibit a combination of a particular consonant (e.g., DH) with a particular vowel (e.g. cloud form of U/UU), I do not believe this prohibition is the responsibility of the shaping engine. Such orthographic prohibitions are the responsibility of an edit control using tools such as spell checking. The reason being that if, at some future time, a document is discovered that does use the cloud form for DHU it would require a bug fix in every rendering system to undo the block that was put in place to prevent this sequence. That would be a difficult thing for a loan researcher how may need to render such a form to accomplish. If it is difficult to accomplish, they

may opt for a different solution, such as using an image, or a different font, or some other hack. Any such hack would undermine the value of having an encoding which is intended to make documents interchangeable.

- SORTING – When used with H, and having semantic distinction, I would assume that the preference would be to sort all occurrences of HU with the cloud form together, and all occurrences of HU with the warbler form together, either before or after the cloud form occurrences on the basis that there is a primary distinction between them. However, does the same apply to the two forms with other bases, for example YU. Is *nayuta* (from your first example in your doc "pu_yu") in with warbler different from *nayuta* with cloud or are they the same? To my knowledge, there is one word "Nayuta" attested in Buddhist Hybrid Sanskrit meaning some very large number, perhaps 100 billion or so. If the variants of U are distinguished with HU they would also be distinguished with Y and all other bases --- unless special measures were taken in order to special case the behaviour, which would require additional work, and additional testing across implementations
- SEARCHING – *From an implementation point of view this may be combined with SORTING in some implementations, but it is useful to think of it separately from a user scenario point of view* - Should a search for HU with warbler find HU with cloud? If yes, then a global find and replace for HU with warbler, would also change HU with cloud to the target replacement. That is to say, from the point of view of SEARCHING, there would be no difference between them. That may not suit HU where a semantic distinction is intended, but might work better for cases where it is not. If one takes the opposite approach and says they are different then when I look up *nayuta* with warbler in an dictionary or index, or using a search engine, should I find *nayuta* with cloud?
- INPUT – Should a keyboard for Siddham include dedicated keys for both the cloud and warbler forms of U/UU? If so, then the user would be at liberty to type either in combination with any base character. If cloud U does not occur with DH, then how should that combination be prevented? If the two forms are not unified for SORTING AND SEARCHING purposes, then we would want users to be careful about which forms they used.

4. My general concern is that it seems like there is a valid case of a semantic distinction for particular syllables which is expressed in the shape of the U and UU vowel matras. This observation has led to a generalized solution due the combinatory nature of the Indic scripts with the effect that we get two matras for U and UU – which must therefore be conceived as semantic differentiators in any combination in which they can occur. This seems to me to be too broad a claim and not in alignment with the three principles that were agreed to at the adhoc meeting. That is to say, just because something is the case for a particular combination with U, it does not follow that it is also the case for all combinations with U. Therefore to my mind the solution does not fit the problem.

I would like to see a solution to this problem, but I think that the proper solution should be thought through with more time, so that it addresses the need to render distinct variants where they meet the agreed criteria, but does not spillover and have untoward effects on other aspects of the encoding.

Cheers,

Andrew