L2/05-251

Hamza Issue

In this document I would like to discuss that the Hamza in Urdu and Sindhi is not the same as Arabic Hamzas 0x0626 and 0x0621. (Since my mother tongue is Sindhi therefore I am presenting some examples in it to clarify my point of view.)

Idea behind Unicode is to have a unified code chart for all scripts that should take care of all languages. Because of this reason we can easily find similar characters with some variations for different languages in the Arabic script. Some examples are as under:

ہۂ ۂ	(0x06C0)	ۓ ۓ	(0x06C2)
ۃ ۃ	(0x0629)	ہ ہ	(0x06C3)
ک ککک	(0x06A9)	ك ككك	(0x0643)
ڪ ڪڪڪ	(0x06AD)	ڳ ڳڳڳ	(0x0763)

None of these variants are used in the same language, but are still included in Unicode so that a single text file can have data of any language. These variants can very easily be handled by the Language tag in OpenType font, but then every language will need a separate font for displaying its text. Therefore if a character is used in different language and has even same context as in any other character in the same script, but since it is used in different language, a separate code point is included. For example there are three kaafs, two gaafs, two rnoons and two duls in the Unicode. All these characters have same sound like the remaining in the group but are used in different languages with different shapes:

ك	(0x0643) – Kaaf for Arabic	ک	(0x06A9) – Kaaf for Farsi, Urdu, Pashto and many other languages of the script	ک	(0x06AA) – Kaaf for Sindhi
گ	(0x06AF) – Gaaf for Farsi, Urdu and Sindhi	ڳ	(0x06AB) – Gaaf for Pashto		
ڻ	(0x06BB) – Rnoon for Sindhi	ڼ	(0x0768) – Rnoon for Siraiki and Pothohari		
ڏ	(0x068F) – Dul for Sindhi	ڍ	(0x0759) – Dul for Siraiki		

So now it is clear that even if a character with glyph variation is required for two separate languages, a separate code point is utilized, and even if there is a need of same character with different shapes in different languages, again a separate code point is introduced in Unicode. Although these features can be very easily handled by OpenType fonts using Language Tag, but we don't see any such practice to be encouraged by Unicode.

How Hamza is Different

Have a look at the examples shown blow:

In Sindhi Grammar the basic forms of all of the verbs end with 'Rnoon' which is same as we put 'to' in front of any verb in English:

Column 1	Column 2	Column 3	Column 4
Eat	کاءُ	To eat	کائڻ
Drink	پيءُ	To drink	پيئڻ
Talk	ڳالهاءِ	To talk	ڳالهائڻ

The Sindhi verbs that I have chosen are the one that end with a Hamza. If we use 0x0621 Hamza, as I have done in column 2, then the shape of Hamza will not change, as it is required after adding rnoon, so right now the only solution is to change the spelling of the verb, and use 0x0626 hamza if we need to add rnoon after it. According to dictionary rules, only Rnoon is the added alphabet in column 4, and the remaining spelling in column 2 and 4 is the same. But since right now there is no such Hamza that has isolated shape as required in Column 2 and initial and middle shapes as required in Column 4, we have to train the user to change the spelling of the verb whenever required which does not happen in English and should not happen in Sindhi as well!

As per the discussion above, variations in different languages of same characters, with slight or major differences, are dealt by adding a different code point in the Unicode Arabic Range, instead of asking the font vendors to add Language tags for their languages or even providing all the variants to the end user and asking him/her to learn which variation to be used in which form of a word. I strongly recommend that a separate Hamza should be added so that different forms of same words can use one single Hamza, instead of two. The reuired Hamza for Sindhi Language should have the following forms:

Isolated	Initial	Medial	Terminal (Non joining)
ء	ؤ	ـؤ	ء

Lateef Sagar Shaikh
Lateef_sagar@yahoo.com
www.paktype.org

September 9, 2005