L2/12-381

Title: Initial and medial forms of Arabic Letter Noon Ghunna
Author: Roozbeh Pournader (Google)
Date: 2012-11-03
Action: For UTC and/or Editorial Committee's action

Proposed action

- Document in the text of the Unicode Standard and potentially the NamesList that U+06BA ARABIC LETTER NOON GHUNNA is dotless in all its contextual forms (form 1).

Alternative action

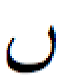
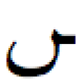
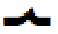
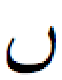
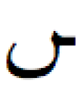
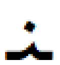
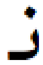
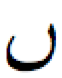
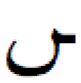
- Choose the dotted form (form 2) mentioned below as the default, create a new joining group called NOON GHUNNA, and change the joining group and schematic name of U+06BA to NOON GHUNNA, **or** make the character right-joining (form 3), create a new joining group for it, and change its joining group and schematic name;

And:

- Encode a new DOTLESS NOON for Koranic and historic usage.

Background

This character behaves differently in various fonts shipped by Unicode members:

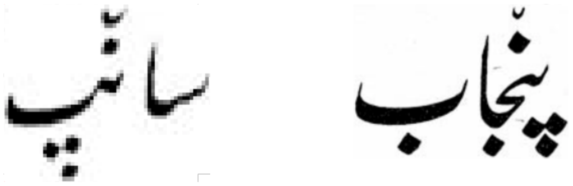
Form #	Isolated	Final	Medial	Initial	Fonts
1					Unicode-defined behavior, Android 4.2 (Droid Naskh), OS X 10.8/iOS6 (Geeza Pro), Windows 8 (Aldhabi)
2					Windows 8 (Arabic Typesetting, Arial, Courier New, Microsoft Sans Serif, Microsoft Uighur, Sakkal Majlla, Tahoma, Times New Roman, Urdu Typesetting)
3					OS X 10.8 (Al Bayan, Baghdad, DecoTypeNaskh, KufiStandardGK, Nadeem), Windows 8 (Segoe UI)

Arabic Letter Noon Ghunna (U+06BA) has been with us since Unicode 1.0. It was then called ARABIC LETTER DOTLESS NOON. Since Unicode 2.0, is has also been specified as dual-joining and has been given the joining group NOON and the schematic name DOTLESS NOON. When joining groups for NOON and YEH were split in Unicode 5.2 and dot patterns of all characters were clarified, U+06BA remained under NOON. All of these imply that the character has been specified to be dotless in all its various forms.

The character unifies two concepts: A nasalized Noon as used in Urdu, and a dotless Noon as used in earlier Arabic texts, including the Koran.

Urdu

Urdu speakers think of Noon Ghunna as they think of Yeh Barree: a subclass of another letter, whose differences are only visible in final and isolated forms. While a final or islated Noon Ghunna is most commonly written as a dotless Noon, it's also sometimes written as a dotted Noon with a ghunna mark (U+0658) above it. A initial or medial "semantic" Noon Ghunna is also commonly written with a normal Noon (U+0646), with a Ghunna mark (U+0658). The following are examples from L2/01-426, which proposed U+0658:



پن٘جاب سان٘پ

Re-typeset in Naskh style in Geeza Pro, they look like:

پن٘جاب سان٘پ

There is a disconnect from the Unicode Arabic model and the semantic mindset of Urdu speakers, but that difference exists in probably all languages written in the Arabic script (see the two-character contextual forms for Uighur, multiple forms of hamza, Yeh Barree, Pashto Yeh's, etc.) The Unicode Arabic model has been following a semi-visual model, where a combination of the written shape and semantics of a letter or word is encoded, as opposed to its pure phonemics or semantics.

Koranic and historic Arabic

To represent Koranic and historic Arabic, where all letters where dotless, U+06BA is used everywhere where a normal Noon (U+0646) would be used in Modern Arabic. Unicode completed its set of dotless Arabic forms in Unicode 3.2 with U+066E Dotless Beh and U+066F Dotless Qaf: Other historic Arabic either had a dotless version among the basic letters (Alef, Hah, Dal, Reh, Seen, Sad, Tah, Ain, Heh, Waw, Alef Maksura) or had dotless versions encoded among the "Extended Arabic Letters" (Dotless Feh and Noon Ghunna):

ا ٮ ح د ر س ص ط ع ڡ ٯ ك ل م ں ه و ى

Also, note that before Unicode 3.0.1 changed Alef Maksura to become dual-joining and Unicode 3.2 encoded the Dotless Beh, using Noon Ghunna/Dotless Noon was the only way to encode a medial or initial dotless “tooth”.

Pros and cons of each choice

None of the choices are legacy compatible. Whichever option we choose, someone’s text will break, and someone’s fonts need to be updated.

Form 1 requires no technical change to the standard. If we choose form 1, some texts created using some Windows fonts will lose dots in their renderings. But that is already happening when text created with those fonts is shown on OS X, Android, or iOS, or with some other Windows fonts. And the text was misspelled anyway.

If we choose form 2, the medial form of Noon Ghunna will not distinguishable from the medial form of Noon, both of which are available on Urdu keyboards. This creates searching and sorting problems, as users may mistakenly type a Noon Ghunna in mid-word and since there is no visual feedback, they will not know the difference. Noon and Noon Ghunna are not marked as confusable in UTR #32, and UCA assigns them different primary weights too. Also, Koranic text created following the standard will now be broken.

Option 3 is the most disruptive, as the character has been defined to be dual-joining since at least Unicode 2.0 (where the joining type was first defined). If we change the character to become right joining, all the text that used the character in medial and initial positions will be broken, as if a Zero Width Non-Joiner was inserted after each occurrence of the letter.

With both options 2 and 3, we would also need to encode a new character and migrate all the previously standard Koranic and historic usage of U+06BA to it.