L2/14-136

The right *heh*s for Arabic script orthographies of Sorani Kurdish and Uighur

Roozbeh Pournader, Google Inc.
May 8, 2014

Summary

The Arabic letter *heh* has some variants in the Unicode Standard, which has led to some confusion for users of the Arabic script. This documents requests clarification on their usage in Sorani Kurdish (hereafter “Sorani”) and Uighur from the Unicode Technical Committee.

Proposal

1. Clarify that consonant /h/ in Sorani should be represented as U+0647 ARABIC LETTER HEH, and the vowel /æ/ as the same character, followed by a Zero Width Non-Joiner in most cases.
2. Clarify that the consonant /h/ in Uyghur should be represented as U+06BE ARABIC LETTER HEH DOACHASHMEE, and the vowel /æ/ as U+06D5 ARABIC LETTER AE. (*This recommendation is also consistent with GB 21669.*)
3. Clarify that the U+06BE ARABIC LETTER HEH DOACHASHMEE is a *heh* letter with stems conceptually connecting to the left side, and should use the same conceptual shape for its medial and final forms and the same conceptual shape for its initial and isolated forms, but the sets of conceptual forms mentioned above may be different, based on language preference and font style.

Background

Sorani and Uighur, among other languages, have taken the Arabic abjad and are using it mostly as a phonetic alphabet, writing almost every vowel and spelling most Arabic loanwords with unified consonants (e.g. Arabic loanwords originally using *sad* may use a *seen*). There are exceptions to the general concept, for example Sorani doesn’t explicitly write the vowel /ɛ/, or sometimes keeps letters as spelled in original Arabic.

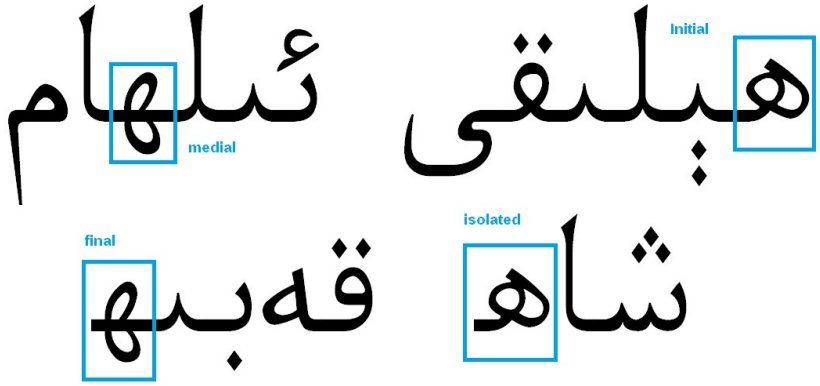
They share an interesting feature, which is the use of a final and isolated-only form of the letter heh to write a mid-word vowel, as if inserting a ZWNJ after all such mid-word usage of the letter: ئەينەكلەردە (“in the mirrors”, in Uighur). In word initial

position, a *yeh with hamza above* character (as the example above) would be used to "carry" the vowel, which may actually be pronounced as a glottal stop in Sorani, but perhaps not in Uighur.

Both languages also have a /h/ consonant, although they are written differently.

Sorani uses a normal *heh* shape for the consonant in all forms. The final and isolated forms of the consonant /h/ in Sorani are visually identical to the vowel /æ/, which may create some confusion to a novice reader. This ambiguity is of course not unique to Sorani, and also exists in the word-final /æ/ phoneme in Persian, for example.

Uighur avoids the ambiguity by reusing the initial and medial forms of heh for isolated and final forms of the consonant /h/:



The four contextual forms of Uighur /h/
(from personal communication with Waris Abdukerim Janbaz)

The words in the last line, transcribed to Sorani, would be written as شاه and قەبیە in Sorani.

The Unicode Standard has not clarified the right character to be used for each of these forms, which has resulted in disputes and confusion among the developer and user communities. The user community cannot be expected to figure out the right character to use either, since none of the Unicode *heh* characters appear to be the clear answer at the first glance. (The author has had a CLDR ticket assigned to him since July 2012 to research this.)

The image on the next page (from Core Specification, version 6.2) lists the four heh characters already encoded in Unicode. The letter AE is basically the same as the letter HEH, except that it is right-joining as opposed to dual-joining. For all practical purposes, AE is equivalent to <HEH, ZWNJ>.

The situation with HEH DOACHASHMEE is more complicated. While the Core Specification shows the character with the same conceptual shape in all forms, languages like Urdu that use the character for aspiration may actually use a shape similar to the medial HEH in some of all forms. In practice, the identity of the HEH DOACHASHMEE character appears to be a *heh*-like letter that always connects to its left side, with shapes similar to initial and/or medial forms of the default Arabic HEH.

HEH	ه	ه	ه	ه
AE	ە	ە		
HEH DOACHASHMEE	ھ	ھ	ھ	ھ
HEH GOAL	ہ	ہ	ہ	ہ

The author recommends different approaches for Sorani and Uighur representation of the vowel and the consonant:

- In Sorani, as the writing system does not distinguish the final and isolated forms of the consonant and the vowel, only one character is sufficient to represent both. U+0647 ARABIC LETTER HEH would be used for both the vowel and the consonant, to avoid the trouble of hidden code differences. The mid-word forms of the vowel would be represented as <HEH, ZWNJ>. Since Sorani already needs ZWNJ for other purposes, this is not a burden on implementations. Smart keyboards could potentially have different keys for the vowel and the consonant, with the vowel key generating the sequence <HEH, ZWNJ> in mid-word.
- In Uighur, since ZWNJ is not commonly used and the shapes are visually distinct, U+06BE ARABIC LETTER HEH DOACHASHMEE should be used for the consonant and U+06D5 ARABIC LETTER AE should be used for the vowel. This would need clarification in the Core Specification that the medial and final shapes of HEH DOACHASHMEE do not necessarily share the same conceptual form as its initial and isolated forms.

Alternatively, UTC can specify that the Uighur /h/ should be encoded as HEH, to be followed by a ZWJ at word-end positions. This would not require any change to the Unicode standard, but would create a burden for Uighur implementations, especially since proper support for ZWJ is still shaky in modern implementations of the Arabic script.

The alternative for Sorani would be specifying that its vowel /æ/ should be represented as AE. Such a solution would create a burden for applications such as matching and searching, as the letters appear identical but have no formal equivalence relation. But since ZWNJ is widely supported in implementations of the Arabic script (especially because it's very common in Persian, Urdu, and Pashto), and Sorani uses the ZWNJ for some other words too, the author's suggestion of using a ZWNJ to represent the mid-word usage of the Sorani vowel does not carry additional complexity.

Open Issues

Further research and clarification is needed for various *heh*s used in the diverse orthographies using the Arabic script. Among those with similar challenges for *heh* are Sindhi, Kazak, Kirgiz, and Behdini Kurdish. Kew 2005 addresses some of them issues, but the author of the present proposal does not agree with all of its recommendations. Future proposals will address those orthographies.

Acknowledgments

Waris Abdukerim Janbaz, through his contributions to Uighur data for CLDR, revived this issue and provided various sources and examples. Kamal Mansour kindly notified the author of Kew 2005. Saman Vaisipour kindly notified the author of Thackston 2006. Ken Lunde kindly notified the author of GB 12050 and made a copy available to him. Mohammad-Hassan Pournader kindly reconciled the author with his physical copy of Rokhzadi 2000.

Bibliography

1. GB 12050–1989. *Information processing – Uighur coded graphic character sets for information interchange.* Beijing. In Chinese.
2. GB 21669–2008. *Information technology – Uyghur, Kazak, and Kirghiz coded character set.* Beijing: Standards Press of China. In Chinese.
3. Jonathan Kew. 2005. "Notes on some Unicode Arabic characters: recommendations for usage." Draft 2. http://scripts.sil.org/cms/scripts/render_download.php?format=file&media_id=arabicletterusagenotes&filename=ArabicLetterUsageNotes.pdf
4. Roozbeh Pournader et al. 2014. "Which Heh's to use for Uighur and Sorani Kurdish". CLDR Ticket #5122. http://unicode.org/cldr/trac/ticket/5122 (opened July 27, 2012.)

5. Ali Rokhzadi. 2000. *Kurdish – Phonetics & Grammar* (آواشناسی و دستور زبان کردی). Tehran: Tarfand. ISBN 964-92684-9-9. In Persian.
6. Imad Saleh and Waris Abdukerim Janbaz. 2006. “Web development Considerations for Unicode-based text processing in Uyghur Language.” The 30th Internationalization and Unicode Conference, Washington, DC. http://www.uyghurdictionary.org/excerpts/Waris_Abdukerim_Janbaz_IUC30.pdf
7. W. M. Thackston. 2006. “Sorani Kurdish – A Reference Grammar with Selected Readings”. http://www.fas.harvard.edu/~iranian/Sorani/sorani_complete.pdf.
8. The Unicode Consortium. 2013. *The Unicode Standard Version 6.2 – Core Specification*.