Doc Type:	Working group document
Title:	Response to *China NB comments on WG2 N5155* (UTC document L2/21-050)
Status:	Liaison contribution
Source:	Unicode Consortium
Source reference:	L2/21-098
Action:	For consideration by JTC1/SC2/WG2 experts
Date:	April 9, 2021

Abstract

Document WG2/N5155 (= UTC L2/20-289) proposed various editorial changes—glyph changes and additional annotations—for certain Arabic-script characters used for Kazakh, Krygyz and Uyghur languages. Before that document was submitted to WG2, it was reviewed by an ad hoc committee of Unicode Technical Committee experts (the UTC Script Ad Hoc), which supported the proposed editorial changes. After the document was submitted to WG2, it was reviewed by China NB experts, who responded with comments (WG2/N5160 = UTC L2/21-050) expressing concerns and disagreement regarding changes for some of the characters. It appears, though, that some of the intent in N5155 may not have been clearly expressed and understood.

This document responds to the China expert comments in N5160 and aims to clarify some relevant details in relation to N5155 that may influence the position of the China experts.

Summary of China NB comments

N5155 proposed changes in code chart glyphs and annotations for certain characters, and China NB experts agreed or disagreed with the proposed changes as follows (taken from N5160):

SN	Characters	China NB feedback
1	U+0626	Disagree
2	U+0641	Disagree
3	U+06C5	Agree
4	U+0677, U+06C7	Only agree to the annotation
5	U+0674..U+0678	Disagree

Items 1, 2, 4 and 5 will each be considered in the following sections.

U+0626 YEH WITH HAMZA ABOVE

N5155 proposed the addition of an annotation for U+0626 regarding placement of the hamza mark for isolate and final forms of this character when used for the Kyrgyz language. It also proposed additional text in Chapter 9 of The Unicode Standard.

The China NB expert comments provided a few statements. The third statement is the most relevant in relation to the proposal:

> "3) In Kyrgyz, the hamza (ء) could be positioned at the top in right or middle for typography. Therefore, it is not better to add the paragraph suggested by the author to UCS and Unicode Core Spec."

While there may be some variation in positioning of hamza in fonts used for Kyrgyz, information available to Unicode experts is that the *preferred* positioning for Kyrgyz users is as described in N5155. Thus, a slightly modified annotation is recommended:

- in Kyrgyz, the preferred position for hamza is top right in isolate and final forms

If China experts can provide evidence that the *preferred* positioning for the hamza in Kyrgyz varies among users or publishers, that can be considered.

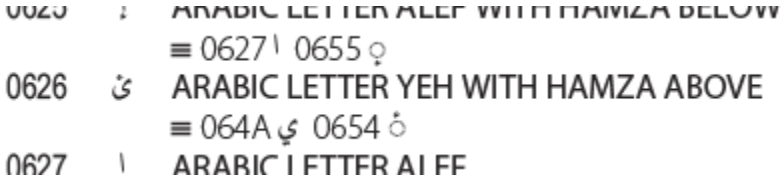
Aside: Clarification regarding normalization and U+0626

The discussion in N5155 cited a decomposition mapping for U+0626 and also included reference to related compatibility characters U+FE89 and U+FE8A. This appears to have led to one of the other statements in the China comments. This doesn't directly affect the proposed annotation, but does reflect misunderstanding on an important detail in UCS.

The first statement in relation to U+0626 begins,

> "1) The first important problem is that U+0626 (ئ) is not equal to <U+064A,U+0654> (ئ) clearly."

This is an important point requiring clarification: ISO/IEC 10646 has explicitly stated that U+0626 is considered *canonically equivalent* to the character sequence, < 064A YEH, 0654 HAMZA ABOVE >. (Note that this sequence is given in logical order, not visual, right-to-left order.) This canonical equivalence relationship has been defined in Unicode since 1999 (Unicode version 3.0), and in UCS since ISO/IEC 10646:2003. This equivalence relationship has been explicitly stated in UCS code charts since ISO/IEC 10646:2011.

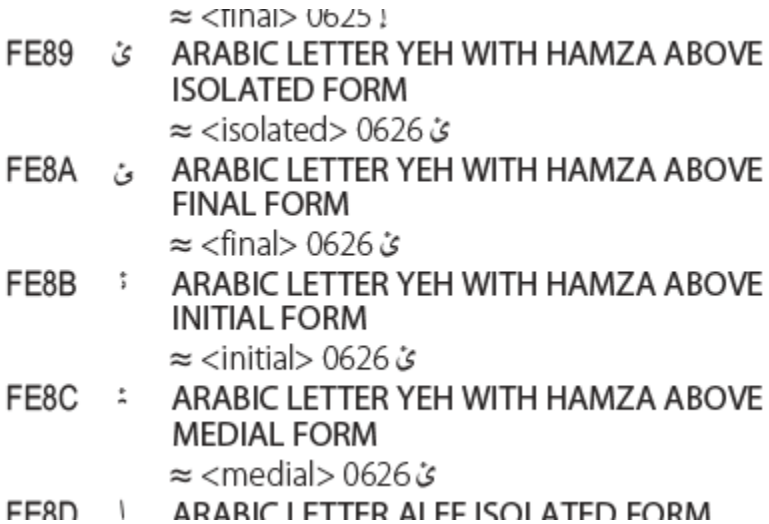


Thus, there is no question that, in UCS, *U+0626 is considered to be equivalent to < 064A, 0654 >*.

The China NB comments go on to state,

> "It is not also to treat U+FE89 (ﺉ), U+FE8A (U+FE8B (ﺋـ) and U+FE8C (as the compatibility characters of <U+064A,U+0654> (ﻲٔ)."

But again, this compatibility equivalence has existed in UCS for many years, and explicitly stated in UCS code charts since ISO/IEC 10646:2011:



Note that, because U+0626 is *canonically equivalent* to the sequence < 064A, 0654 > and U+FE89, etc. are compatibility equivalent to U+0626, the effects of Unicode Normalization require that U+FE89, etc. have a compatibility equivalence relationship to the sequence < 064A, 0654 >.

Again, these issues are important to correct understanding of UCS, but do not appear to affect the proposed changes related to U+0626.

U+0641 FEH

N5160 includes a brief section discussing U+0641. A preliminary draft of N5155 had included discussion of this character, as well as U+06A7 QAF WITH DOT ABOVE, but this was removed from the final version of N5155 submitted to WG2, to be "addressed in another document" (still forthcoming).

Although discussion of U+0641 and U+06A7 was removed from N5155, the China expert comments in N5160 appear to be responding to the preliminary draft of N5155, and a change to the annotations for U+06A7 is proposed. (The China experts submitted a separate document discussing these issues to the Unicode Technical Committee, UTC document L2/20-293.)

Since N5155 does not propose changes related to U+0641 or U+06A7, and given that there are complexities related to these character, face-to-face discussion among experts is recommended as the best way to progress through the issues.

U+0677 U WITH HAMZA ABOVE, U+06C7 LETTER U

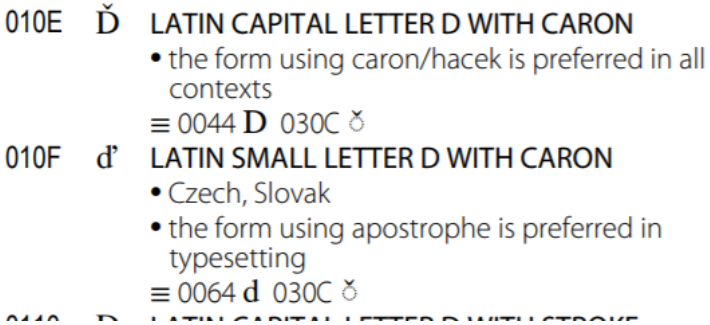
N5155 proposes glyph and annotation changes for U+0677 U WITH HAMZA ABOVE and U+06C7 U, as well as changes to the “schematic” character names for these characters given in the ArabicShaping.txt character properties data file. The proposed glyph changes are based on an understanding that the “damma” component of these characters should always have a comma-like appearance with the loop filled in.

China expert comments in N5160 concur with the proposed annotation changes, but provide additional information related to the proposed glyph and character-property changes:

- That damma is written throughout China using a comma-like form.
- However, U+06C7 is also used for Azerbaijani, and that display requirements for this character in Azerbaijan should also be taken into consideration.

While U+06C7 is used for Azerbaijani, Arabic script has not been commonly used within the country of Azerbaijan since the early 20th century. However, Arabic script is still used for Azerbaijani language in the country of Iran. In that country, U+06C7 is consistently written using a comma-like form for the damma.

N5160 also suggests that a glyph change to U+0677 and U+06C7 would be a form of disunification, but that is not really true: these characters do not have decomposition mappings and so there is no formal connection between the “damma” / “comma” component of these characters and U+064F DAMMA or any other characters involving a damma-like component. Thus, potential concerns about disunification are not a significant factor in evaluating technical merits of a proposed glyph change. A similar situation with existing precedent exists in the case of U+010F LATIN SMALL LETTER D WITH CARON: the corresponding uppercase letter is always presented with a caron / hacek component, but this lowercase letter is normally presented with a comma / apostrophe component.



In that case, there is a canonical decomposition to a sequence with U+030C COMBINING CARON, yet the representative glyph for U+010F in the UCS charts has a comma / apostrophe form. In this case, the glyph used is the form preferred for the languages in question without any concern about appearance of a disunification.

China experts also comment on the proposed change to ArabicShaping.txt that there is no existing precedent for use of *comma* in the schematic names within that data file. Lack of existing precedent does not preclude introducing a term in the schematic names when appropriate. Given that the

schematic names (unlike the character names) are intended to reflect appearance, and given the evidence to have glyphs for these two characters use a comma-like form, Unicode experts consider the use of *comma* in the schematic names to be appropriate.

Thus, it remains the position of Unicode experts that the glyph changes and the changes to ArabicShaping.txt as proposed in N5155 would be appropriate and an improvement for users.

There is consensus between China and Unicode Script Ad Hoc experts regarding the annotation changes proposed in N5155, so that aspect of the proposal could be adopted directly.

U+0674 HIGH HAMZA, U+0675 HIGH HAMZA ALEF – U+0678 HIGH HAMZA YEH

N5155 proposes various changes related to the characters U+0674..U+0678:

- Glyph changes in relation to the positioning of high hamza
- Discourage use of U+0675..U+0678—to be reflected in code charts as well as in the core text of the Unicode Standard
- Remove annotations for U+0675..U+0678 mentioning use for Kazakh
- Adjust DUCET collation weights for U+0675..U+0678 (would be reflected in a new edition or amendment to ISO/IEC 14651)

China experts provide several comments in N5160.

The issues regarding these characters fall into two distinct categories:

i) Recommendation versus discouragement of use of U+0675..U+0678
ii) Compatibility equivalence relationships for U+0675..U+0678 and the general category property of U+0674—letter versus combining mark

Issue (ii) has a bearing on consideration of issue (i), so (ii) will be considered first.

Compatibility equivalence relationships for U+0675..U+0678

In N5160, the section discussing these characters begins with the statement,

> "The author [of N5155] suggests treating U+0674 (ٔ) as a combining mark rather than a spacing character…"

This is a misunderstanding of the intended meaning of N5155. In comments 3 and 4, the China experts note an anomaly in the compatibility decomposition mappings for U+0675..U+0678:

> "3) We also noticed the orders of the equivalent sequences of U+0675 (ٵ), U+0676 (ٶ), U+0677 (ٷ) and U+0678 (ٸ) are questionable. The orders of two elements of the equivalent sequences should be interchanged…

> "4) Similar to U+0626 (ئ), the equivalent sequence of U+0678 (ٸ) should be changed to <U+0674,U+0649> (ىٔ) from <U+064A,U+0674>…"

This is, in fact, exactly the same observation that is being made in N5155, though it is expressed in a different way. The author of N5155 was not proposing that U+0674 be changed to a combining mark, but was noting that the ordering of elements in the decomposition mapping of U+0675..U+0678 would only make sense *if U+0674 were a combining mark*. There is an implied meaning intended: that the ordering of the elements is incorrect since U+0674 is categorized as a letter, not as a combining mark.

In regard to the general category of U+0674 as a letter, then, the authors of N5155 and N5160 are in complete agreement: the decomposition mappings of U+0675..U+0678 are not as expected! In relation to the following statement in N5160,

> "1) We object to change U+0674 (ٔ) to a combining mark."

this is moot (a non-issue) since N5155 is not proposing such a change.

The anomalous decomposition mappings for U+0675..U+0678 has implications: *What should be done?* This is where the authors of N5155 and of N5160 differ.

The authors of N5160 suggest changes to the decomposition mappings:

> "3) … The orders of two elements of the equivalent sequences should be interchanged, that means decomposition property values in UCD should be modified."

> "4) … the equivalent sequence of U+0678 (ٸ) should be changed to <U+0674,U+0649> (ىٔ) from <U+064A,U+0674> (ئ)..."

Unfortunately, while these changes would seem to make sense, these changes are not possible due to stability requirements in UCS and Unicode. This is stated in clause 22 of ISO/IEC 10646:2020:

> "NOTE 1 – The result of applying any of these normalization forms onto a code unit sequence is intended to stay stable over time. It means that the normalized representation of a code unit sequence consisting of characters assigned in this version of the standard remains normalized even when the standard is amended."

For the Unicode Standard, this stability constraint is stated as part of Unicode Character Encoding Stability Policies:

> "Once a character is assigned, its decomposition mapping will not change."

Stability of normalization is critical in relation to other industry specifications that depend on Unicode and UCS normalization forms, such as IETF specifications for international domain names.

Thus, for this issue affecting U+0675..U+0678, the remedy proposed in N5160 is not a viable option. This is important background to understanding the changes proposed in N5155.

N5155 proposes a change to DUCET collation weights for U+0675..U+0678. The intent here (not well explained) is to have these character collate the same as correctly ordered sequences using U+0674 (e.g., for U+0675 to collate the same as <0674, 0627> rather than <0627, 0674> as currently found in DUCET). This change is not constrained by any Unicode or UCS stability requirements and *would be possible* in a future edition or amendment of ISO/IEC 14651.

Recommendation versus discouragement of use of U+0675..U+0678

The anomalous decomposition mappings for U+0675..U+0678 are a significant defect in the UCS / Unicode encoding for these text elements. As noted, fixing the decomposition mappings is not possible. This leaves two options for vendors and end users:

i) Continue support and use of U+0675..U+0678 in spite of the anomalous decomposition mappings and problems that might arise for implementations or for end users as a result.
ii) Discourage use of U+0675..U+678 and recommend, instead, use of corresponding character sequences (with an appropriate ordering of elements) involving U+0674.

The latter is what is proposed in N5155. Given the problems associated with U+0675..U+0678, Kazakh users will encounter fewer problems by avoiding use of these characters, and using sequences with U+0674 (or U+0621 HAMZA) instead.

But evidence indicates that use of sequences instead of U+0675..U+0678 is already common practice for Kazakh content, as noted in N5155:

> "In practice, people usually use U+0674 followed by the base character."

For example, consider the Kazakh-language Tianshannet News site, kazakh.ts.cn: this site consistently uses sequences with U+0621 HAMZA rather than U+0675..U+0678. This can be seen in the navigation menus at the top of each page, as well as in articles:

Figure 1. Sequence with *hamza + alef maksura* (< 0621, 0649 >) (from http://kazakh.ts.cn/system/2021/04/10/036610927.shtml)

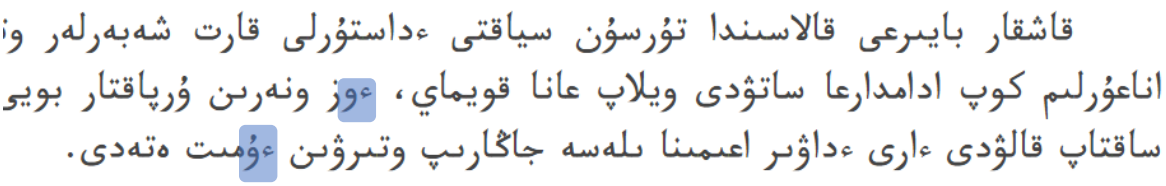


Figure 2. Sequences with *hamza + waw* (< 0621, 0648 >) and *hamza + letter u* (< 0621, 06C7 >) (from http://kazakh.ts.cn/system/2021/04/10/036610927.shtml)

Thus, while N5160 recommends maintaining current practice,

> "5) There are no problems to use the current method for so many years in many companies and industries, so that it is best to keep it stable."

the main proposal of N5155 is not, in fact, inconsistent with that recommendation: it doesn't prohibit use of U+0675..U+0678, but rather encourages use of sequences that are already in use.

Even if use of U+0675..U+0678 is discouraged, there may still be Kazakh document created using these characters. It is for that reason that N5155 also proposes changes that will improve user experience if these characters are used:

- Revise glyphs for U+0674..U+0678 so that the high hamza has the same vertical position across all of these characters.
- Revise the DUCET collation weights for U+0675..U+0678 so that they collate like the sequences that are being used in practice, and not like the problematic compatibility decomposition sequences for these characters.

N5160 does not raise objections to these changes, and so these are assumed not to be controversial.

Conclusions

Authors of N5155 and N5160 are in agreement on some points but not others. The following is a summary of findings, as discussed above:

Closed issues for which there is consensus among China and Unicode experts:

a) U+06C5 KIRGHIZ OE: add annotation proposed in N5155.

b) U+06C7 LETTER U: add annotation proposed in N5155.

c) U+0674 HIGH HAMZA: Do not change the general category from letter to combining mark.

d) U+0675 HIGH HAMZA ALEF – U+0678 HIGH HAMZA YEH: Decomposition mappings for these characters are anomalous and problematic.

Closed issues on which China experts did not comment (hence consensus is assumed):

e) U+0675 HIGH HAMZA ALEF – U+0678 HIGH HAMZA YEH:
   • Revise glyphs (vertical position of high hamza) as proposed in N5155.
   • Revise DUCET collation weights as proposed in N5155.

Open issues for which Unicode experts have arrived at a recommendation, after considering the input from China experts in N5160, and request further review by China experts. *Note that these are all editorial in nature*:

f) U+0626 YEH WITH HAMZA ABOVE: add an annotation:
   • in Kyrgyz, the preferred position for hamza is top right in isolate and final forms

g) U+0677 U WITH HAMZA ABOVE, U+06C7 LETTER U: make glyph changes and changes to ArabicShaping.txt as proposed in N5155

h) U+0675 HIGH HAMZA ALEF – U+0678 HIGH HAMZA YEH
   • Discourage use of these characters
   • Remove the annotation, “Kazakh” (to avoid suggestion that use of these characters for Kazakh is recommended)

Other open issues requiring further discussion among experts:

i) Possible changes in relation to U+0641 FEH

Finally, as suggested in N5160, face-to-face discussion among experts on the open issues is likely to be useful, and an in-person or virtual meeting between China and Unicode experts, or other WG2 experts, would be welcomed.