Input specification of Arabic inline characters in Unicode

1 Problem Statement

Unicode currently lacks a consistent and complete specification for handling Arabic inline characters. In this document we will consider two such characters:

1. *Hamzaḧ*
2. *ʾAlif k̲hanjariyyaḧ* "Dagger Alef"

1.1 *Hamzaḧ*

The case for inline *hamzaḧ* has been discussed in some detail in (Pournader 2009; Milo 2013 Section IV, 2014) but, it appears, without much official resolution. There has also been an action item 139-A60 to address the topic.

We will attempt to summarize the problem here.

Being a later addition to the Arabic script, *hamzaḧ* is written as a diacritic on the basic skeletal text. This *hamzaḧ* can be written in two different ways:

1. Above a base letter, known as the "seat of the *hamzaḧ*": In writing the Arabic language *hamzaḧ* may be seated above a vowel letter: أ ؤ ئ. It may also be sometimes written below vowel letters, depending on the writing style and vowel mark: إ ي. In Persian language writing, it may also appear above a U+0647 ARABIC LETTER HEH: ۀ.
2. After (to the left of) a base letter in the basic skeletal text. This has two sub-cases:
   a. Standalone *hamzaḧ*: After a non-joining base letter (ا د ذ ر ز و) or at the end of a word. In this case, *hamzaḧ* is written at the baseline. Examples: دعاء, سوء, عِبْء.
   b. Inline *hamzaḧ*: After a joining base letter in the middle of a word. In this case *hamzaḧ* is written between the two joining base letters, above the baseline, and without affecting the joining of the two base letters. Examples: خطيئة, شيئًا, بريئين

For a set of rules to determine the seat of the *hamzaḧ* and whether *hamzaḧ* is to be written inline, see this article: https://adamiturabi.github.io/hamza-rules/

For case (1), Unicode provides ◌ٔ U+0654 ARABIC HAMZA ABOVE and ◌ٕ U+0655 ARABIC HAMZA BELOW as combining diacritics.

For case (2a), Unicode provides ء U+0621 ARABIC LETTER HAMZA.

It is case (2b) which is problematic and has not been given sufficient attention by the Unicode standard. It would have been sufficient to use ء U+0621 ARABIC LETTER HAMZA and specify that it shall not break the joining of the letter before it with the letter after it. Unfortunately, Unicode specifies that ء U+0621 ARABIC LETTER HAMZA is a breaking character. OpenType shaping engines enforce this (as they rely on Unicode for joining behaviour). Fonts that want to change this behaviour will have to jump through many hoops to achieve it.

Futhermore, a set of "simplified" *hamzaḥ* rules have been proposed by Arabic Language Academy in Cairo where inline *hamzaḥ* is approximated with ئ U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE, which between two joining letters looks similar to an inline *hamzaḥ* ـئـ.

Nevertheless, the need still exists to input and display inline *hamzaḥ* according to the proper rules without approximating it with ئ. In order to facilitate this, some fonts propose the use of ـ U+0640 ARABIC TATWEEL with ◌ٔ U+0654 ARABIC HAMZA ABOVE. For example, خطيئة will be input with the following sequence:

Sequence Order	Arabic character	Encoding	Description
1.	خ	U+062E	ARABIC LETTER KHAH
2.	ط	U+0637	ARABIC LETTER TAH
3.	ي	U+064A	ARABIC LETTER YEH
4.	ـ	U+0640	ARABIC TATWEEL
5.	◌ٔ	U+0654	ARABIC HAMZA ABOVE
6.	ة	U+0629	ARABIC LETTER TEH MARBUTA

This solution is not ideal because it introduces a character in the basic skeletal text (U+0640 ARABIC TATWEEL) which is not part of the word's spelling. This can also adversely affect computerized search and indexing of text encoded in this manner.

Besides this being an improper hack, this solution will simply not work when inline *hamzaḥ* comes between *lām* and *ʾalif*. In this case *hamzaḥ* properly ought not to break the mandatory *lām-ʾalif* ligature لا. For example, the word الآخرة *alʾāk̲hiraḥ* has a *hamzaḥ* between the *lām* and *ʾalif*. When input with a U+0640 ARABIC TATWEEL, it will be displayed incorrectly as الأخرة.

Such words are common in some writing styles of Qurʾānic text. Even in non-Qurʾānic text there are some words which require inline *hamzaḥ* between *lām* and *ʾalif*. For example مَلَأً *malaʾan* and لَّأَل *laʾʾāl*.

1.2 Dagger Alef

Dagger Alef can also be considered an inline character. Depending on the writing style and intended pronunciation, it can appear either:

1. Above a base letter: هٰذا *hādhā*, ذٰلك *dhālika*, صلوٰة *ṣalāẗ*, علىٰ *ɛalā*
2. After a base letter. This two has two sub-cases:
    a. After a non-joining base letter: صلوٰت *ṣalawāt*, صرٰط *ṣirāṭ*, ذٰلك *dhālika*
    b. After a joining base letter: هٰذا *hādhā*, أولٰئك *ʾulāʾika*

Unicode provides ◌ٰ U+0670 ARABIC SUPERSCRIPT ALEF to encode Dagger Alef. Being a combining character, U+0670 ARABIC SUPERSCRIPT ALEF is sufficient to encode case (1) above.

Cases (2a) and (2b) do not have a standard specification for being input. Many fonts propose using U+0640 ARABIC TATWEEL combined with U+0670 ARABIC SUPERSCRIPT ALEF for case (2a), and U+00A0 NO-BREAK SPACE (NBSP) or U+202F NARROW NO-BREAK SPACE (NNBSP) combined with U+0670 ARABIC SUPERSCRIPT ALEF for case (2b).

However, this solution is problematic for some of the same reasons as for *hamzaẗ*:

1. It introduces characters in the text which are not part of the word's spelling: TATWEEL and NBSP/NNBSP.
2. The user has to choose between between TATWEEL and NBSP/NNBSP simply on the basis of whether the preceding character is joining or non-joining. The computer's shaping engine already handles joining of base letters. Therefore, this choice left to the user is non-semantic and unnecessary.
3. It may adversely affect computerized search and indexing.

Other fonts will render the Dagger Alef above the base letter by default but automatically offset the Dagger Alef to the left if a ◌َ U+064E ARABIC FATHA is encoded before it. Thus these fonts can display هٰذا and هَـٰذا but (problematically) not هـٰذا.

1.3 Summary

We may condense the problem statement thus:

Arabic has the need to display inline characters as diacritics that:

1. May generally be either above or after (to the left of) a base character. The two (above and after) are to be differentiated as they may affect pronunciation.
2. Must not affect the joining of base letters.

2 Proposed solution

We propose specifying the use of U+034F COMBINING GRAPHEME JOINER (CGJ) to determine whether a diacritic appears above or after a base letter.

Also that ◌ٔ U+0654 ARABIC HAMZA ABOVE be used for the diacritic inline *hamzaẗ* and ◌ٰ U+0670 ARABIC SUPERSCRIPT ALEF continue to be used for the diacritic Dagger Alef.

If the diacritic is to appear above the base letter then it shall be input after the letter without an intermediate CGJ.

If the diacritic is to appear after the base letter then it shall be input after the letter with an intermediate CGJ before the diacritic.

CGJ and/or the diacritic shall not affect the joining of base letters.

CGJ shall prevent any other diacritics before it from normalizing/reordering with the diacritics after it.

We recommend to font developers and typographers that the combination of CGJ followed by ◌ٔ U+0654 ARABIC HAMZA ABOVE after a non-joining letter be rendered identical to ء U+0621 ARABIC LETTER HAMZA.

Here are some examples of the proposed input sequences and the displayed rendering:

Arabic text	Transcription	Character input sequence
خطيئة	*k̲haṭīʾaẗ*	KHAH, TAH, YEH, **CGJ**, HAMZA ABOVE, TEH MARBUTA
خاطئة	*k̲hāṭiʾaẗ*	KHAH, ALEF, TAH, YEH, HAMZA ABOVE, TEH MARBUTA
لَّآل	*laʾʾāl*	LAM, FATHA, **CGJ**, HAMZA ABOVE, SHADDAH, ALEF, LAM
هٰذا	*hāḏ̲hā*	HEH, SUPERSCRIPT ALEF, THAL, ALEF
هـذا	*hāḏ̲hā*	HEH, **CGJ**, SUPERSCRIPT ALEF, THAL, ALEF
هَـٰذا	*hāḏ̲hā*	HEH, FATHA, **CGJ**, SUPERSCRIPT ALEF, THAL, ALEF
ذٰلك	*ḏ̲hālika*	THAL, SUPERSCRIPT ALEF, LAM, KAF
ذ ٰلك	*ḏ̲hālika*	THAL, **CGJ**, SUPERSCRIPT ALEF, LAM, KAF
ذَ ٰلك	*ḏ̲hālika*	THAL, FATHA, **CGJ**, SUPERSCRIPT ALEF, LAM, KAF
صلوٰة	*ṣalāẗ*	SAD, LAM, WAW, SUPERSCRIPT ALEF, TEH MARBUTA
صلو ٰت	*ṣalawāt*	SAD, LAM, WAW, **CGJ**, SUPERSCRIPT ALEF, TEH
سمـو ٰت	*samāwāt*	SEEN, MEEM, **CGJ**, SUPERSCRIPT ALEF, WAW, **CGJ**, SUPERSCRIPT ALEF, TEH

3 Other inline characters

There exist other inline diacritical characters in Arabic used for displaying Qurʾānic text like:

1. ◌ُ U+08F3 ARABIC SMALL HIGH WAW. Example: لِيَسُۥٓـُٔوا *liyasūʾū*.
2. ◌ۧ U+06E7 ARABIC SMALL HIGH YEH. Example: وَلِيِّيَ *waliyyiya*.

These too are currently input using a U+0640 ARABIC TATWEEL. However, the proposal to use CGJ should work here as well. Note that, Unicode has separate non-joining characters corresponding to the above which, strictly speaking, will no longer be necessary:

1. ۥ U+06E5 ARABIC SMALL WAW.
2. ۦ U+06E6 ARABIC SMALL YEH.

But similar to the case of non-joining *hamzaḥ̈*, we recommend, for example, that the sequence CGJ, ◌ۥ U+08F3 ARABIC SMALL HIGH WAW after a non-joining base letter be displayed identical to ۥ U+06E5 ARABIC SMALL WAW.

References

Milo, Thomas. 2013. "Arabic Amphibious Characters Phonetics, Phonology, Orthography, Calligraphy and Typography." *UTC Document L2/13-226*. https://unicode.org/L2/L2013/13226-koran-ortho.pdf.

———. 2014. "Arabic Inline Charactersfor Qur'ānic and Classic Orthographyin Unicode and Computer Typography." *UTC Document L2/14-109*. https://unicode.org/L2/L2014/14109-inline-chars.pdf.

Pournader, Roozbeh. 2009. "Discussion Document for Polishing Koranic Support in Unicode." *UTC Document L2/09-358R*. https://www.unicode.org/L2/L2009/09358r-koranic-status.pdf.