Inputting of Arabic inline characters in Unicode

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. *Hamzaẖ*
2. *ʾAlif k̲hanjariyyaẖ* "Dagger Alef"

Hamzaẖ

The case for inline *hamzaẖ* has been discussed in some detail but, it appears, without much official resolution. https://unicode.org/L2/L2014/14109-inline-chars.pdf https://unicode.org/L2/L2013/13226-koran-ortho.pdf Section IV

We will attempt to summarize the problem here.

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

1. Above a base letter, known as the "seat of the *hamzaẖ*": In writing the Arabic language *hamzaẖ* 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 heh: ۀ.
2. After (to the left of) a letter in the basic skeletal text. This has two sub-cases:
    a. Standalone *hamzaẖ*: After a non-joining base letter (ا د ذ ر ز و) or at the end of a word. In this case, *hamzaẖ* is written at the baseline. Examples: دعاء, سوء.
    b. Inline *hamzaẖ*: After a joining base letter in the middle of a word. In this case *hamzaẖ* is written between the two joining base letters, above the baseline, and without affecting the connection of the two base letters. Examples: خطيئة, شيئًا, بريئين

For a set of rules to determine the seat of the *hamzaẖ* and whether *hamzaẖ* is to be written inline, see this article.

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

For case 2a, Unicode provides 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 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, the vast majority of fonts consider U+0621 ARABIC LETTER HAMZA a breaking character.

In order to display *hamzaḧ* according to case 2b, most users are accustomed to approximating it with ء which between two joining letters looks similar to an inline *hamzaḧ* .ـئـ In fact, due to the prevalence of this approximation, many users will consider it the correct way of writing, even in handwritten text.

In order to follow the proper rules without approximating it with ,ء some fonts propose the use of tatweel with arabic *hamzaḧ* above. Examples:

This solution is not ideal because it introduces a character in the basic skeletal text (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 *hamzaḧ* comes between lam and alef. In this case *hamzaḧ* properly ought not to break the mandatory lam-alef ligature. Examples:

Such words are common in some writing styles of Quranic text. Even in non-Quranic text there are some words which require inline *hamzaḧ* between lam and alef. Examples: Dagger alef Dagger alef can also be considered an inline character. Depending on the writing style and intended pronunciation, it can appear either: Above a letter: ـٰـ ـٰـ ـٰـ ,ـٰـ or After a letter. This two has two sub-cases: After a non-joining letter: ـا ,ـا ـا ـا ـا ـا After a joining letter: ـٰـ ـٰـ Unicode provides U+0670 Arabic superscript alef to encode dagger alef. Being a combining character, it is sufficient to encode case 1 above.

Cases 2a and 2b do not have a standard specification for being input. Many fonts propose using tatweel + 0670 for case 2a, and NBSP/NNBSP + 0670 for case 2b.

However, this solution is problematic for some of the same reasons as for *hamzaḧ*: It introduces characters in the text which are not part of the word's spelling: tatweel and NBSP/NNBSP. 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. It may adversely affect computerized search and indexing.

Other fonts will automatically offset the dagger alef to the left if a fatha is encoded before it. Thus they can display ـَٰـ and ـٰـ but (problematically) not .ـَٰـ Summary We may condense the problem statement thus:

Arabic has the need to display inline characters as diacritics that: 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. Must not affect the joining of base letters. Solution We propose specifying the use of U+034F Combining Grapheme Joiner to determine whether a diacritic appears above or after a base letter.

Also that U+0654 Arabic *hamzaẗ* 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 intermediate CGJ.

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

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

We recommend to font developers and typographers that CGJ + *hamzaẗ* above after a non-joining letter be rendered identical to U+0621.

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