Report on two problems in Arabic script handling in MS Word that severely hamper more sophisticated Arabic script OpenType fonts. Prepared by John Hudson, Tiro Typeworks, October 2009.

Problem 1 : diacritic display and colouring.

Word provides an option to turn off display of diacritics in Arabic text or to apply a distinct colour to diacritic marks. Since colouring of diacritics at the character level would disrupt OpenType Layout and hence Arabic script shaping, it is applied after layout and diacritics are identified not by their character encoding but by their identification as marks in the font GDEF table. The problem with this approach is that there is no way to distinguish different kinds of marks in a font, and hence all marks in the GDEF table are treated as if they are diacritics. In some Arabic fonts, particularly those in *nastaliq* and related styles, the dots that form part of the identity of the letter are decomposed and positioned as marks; this is necessary because the position of the dots relative to the letter shape is not consistent but varies contextually. Since the decomposed dots function as marks, they need to be classed as marks in the GDEF table, which means that Word ends up treating them as if they were diacritics: the dots disappear if the display diacritics option is disabled by the user, or assume a colour different from the letter to which they belong if diacritics are distinctively coloured.

The block of text below is set in Microsoft's new Urdu Typesetting *nastaliq* font, which is one of the fonts that use the dot decomposition method.

ان چیزوں کو شیکاگو میں ہوئی ایک نیلامی میں رکھا گیا تھا جس میں ان کے پسینے کے دھبے والے
سکارف ، نہانے والا گاؤن اور کچھ ریکارڈز شامل تھے۔ اس کے علاوہ ان کی ایک شرٹ بھی اس
نیلامی میں باؤن ہزار امریکی ڈالر میں فروخت ہوئی ہے۔ یہ تمام چیزیں ایلوس پریسلی کے ایک
دوست گیری پیپر کے ہیں۔ ان کا انتقال ہو چکا ہے اور وہ ایلوس پریسلی کے پرستاروں کا کلب
چلایا کرتے تھے۔ پیپر انیس سو اسّی میں انتقال کر گئے تھے۔

Here it is again with the diacritic display option disabled.

ان چیزوں کو شیکاگو میں ہوئی ایک نیلامی میں رکھا گیا تھا جس میں ان کے پسینے کے دھبے والے
سکارف ، نہانے والا گاؤن اور کچھ ریکارڈز شامل تھے۔ اس کے علاوہ ان کی ایک شرٹ بھی اس
نیلامی میں باؤن ہزار امریکی ڈالر میں فروخت ہوئی ہے۔ یہ تمام چیزیں ایلوس پریسلی کے ایک
دوست گیری پیپر کے ہیں۔ ان کا انتقال ہو چکا ہے اور وہ ایلوس پریسلی کے پرستاروں کا کلب
چلایا کرتے تھے۔ پیپر انیس سو اسّی میں انتقال کر گئے تھے۔

And here with diacritics distinctively coloured.

ان چیزوں کو شیکاگو میں ہوئی ایک نیلامی میں رکھا گیا تھا جس میں ان کے پسینے کے دھبے والے
سکارف ، نہانے والا گاؤن اور کچھ ریکارڈز شامل تھے۔ اس کے علاوہ ان کی ایک شرٹ بھی اس
نیلامی میں باؤن ہزار امریکی ڈالر میں فروخت ہوئی ہے۔ یہ تمام چیزیں ایلوس پریسلی کے ایک
دوست گیری پیپر کے ہیں۔ ان کا انتقال ہو چکا ہے اور وہ ایلوس پریسلی کے پرستاروں کا کلب
چلایا کرتے تھے۔ پیپر انیس سو اسّی میں انتقال کر گئے تھے۔

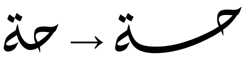
[I'll also note that it seems to me an error that diacritic display and colouring is a global option. It should be possible to display or not display Arabic diacritics, or to distinctively colour them, at the text selection level or, failing that, at the paragraph level. Creating this document required three different applications.]

There seem to me a couple of ways to fix this problem. One possibility would be to define a new version of the OT GDEF table that provides for different categories of mark glyphs, such that glyphs representing diacritic characters might be distinguished from marks like the Arabic dots. Along the same lines, the GDEF table could contain specific colour class values for glyphs that should be coloured alike; this solution might benefit scripts such as Ethiopic which have bichromatic traditions that are not limited to marks. Obviously this would require updates to the OT spec, to fonts, and also to Word and other applications wanting to handle display and colouring of diacritics.

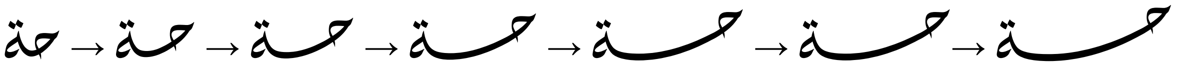
Another option would be an application level fix such that, rather than relying on GDEF glyph classification, Word would trace glyphs back through OTL lookups to find out with what character codes they are associated, so as to be able to accurately identify actual diacritics. Note, however, that some encoded Arabic characters might be decomposed into letter+diacritic mark for display purposes, so care would need to be taken not to presume that any decomposition represents a letter plus non-diacritic mark.

Problem 2 : *kashida* justification.

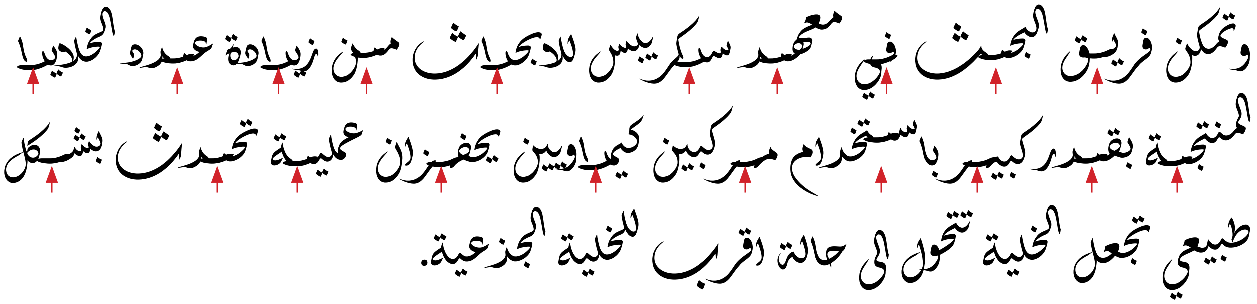
Word provides three levels of Arabic text justification using *kashida* (aka *tatweel*) insertion. It is helpful to understand what the *kashida* represents, which is a stroke elongation permissible in Arabic writing. Beginning with metal typesetting and carried over into digital typography, this elongation has typically been handled by insertion of one or more separate *kashida* glyphs, a mechanism that only works without difficulties in purely horizontal styles of Arabic type, *i.e.* not in any of the traditional styles of Arabic script. It is possible, using OpenType GSUB to implement elongations via either letter variants or insertion of non-flat connecting strokes, combined with GPOS cursive attachment positioning, as shown in this example from new Shallaal typeface in development for the Advanced Reading Technologies group:



The *kashida* may be manually inserted by the user as U+0640, and generally this seems to cause no problem in Word as this character is processed like any other Arabic character, going through full Unicode layout to arrive correct form and positioning. Multiple *kashida* characters may be entered by the user, and a font can intelligently combine these into connecting strokes of different lengths:

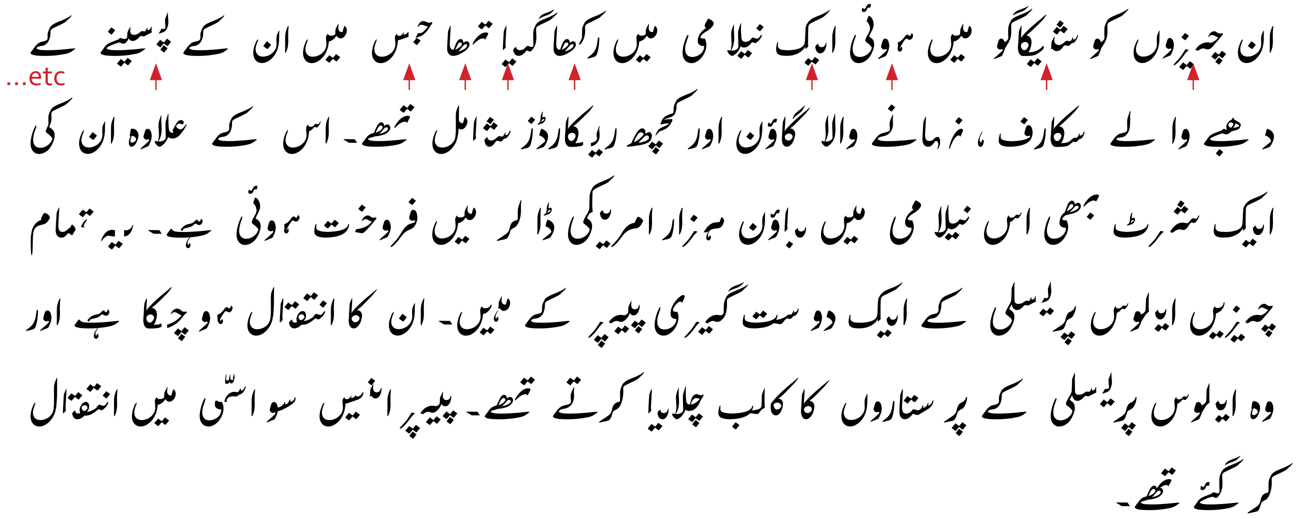


The *kashida* may also be inserted automatically as part of an application's text justification algorithm, and this is where Word runs into trouble with non-flat Arabic script styles, because the *kashida* insertion takes place after OpenType Layout, which means that neither GSUB nor GPOS are applied to these *kashidas* inserted in this way. The result is a serious mess. This is what happens with the Shallaal font, using Word's 'Justify Medium' setting:


وتمكن فريق البحث في معهد سكريبس للابحاث من زيادة عدد الخلايا
المنتجة بقدر كبير باستخدام مركبين كيماويين يحفزان عملية تحدث بشكل
طبيعي تجعل الخلية تتحول الى حالة اقرب للخلية الجزعية.

All the correct OTL shaping is taking place for unjustified text, and then the *kashida* glyphs are being inserted after the fact. The *kashida* glyphs and/or adjacent letters are not being changed to their appropriate forms for elongation using the contextual GSUB lookups, and they are not being properly connected using the cursive attachment GPOS lookups.

In the Urdu Typesetting font, which doesn't seem to include a *kashida* glyph, the proper text shaping is broken in numerous places, indicating that extra spacing is being applied as if there were a *kashida*:


ان چیزوں کو شکاگو میں ہوئی ایک نیلامی میں رکھا گیا تھا جس میں ان کے پسینے کے
...etc
دھبے والے سکارف ، نہانے والا گاؤن اور کچھ ریکارڈز شامل تھے۔ اس کے علاوہ ان کی
ایک شرٹ بھی اس نیلامی میں باؤن ہزار امریکی ڈالر میں فروخت ہوئی ہے۔ یہ تمام
چیزیں ایلوس پریسلی کے ایک دوست گیری پیپر کے ہیں۔ ان کا انتقال ہو چکا ہے اور
وہ ایلوس پریسلی کے پرستاروں کا کلب چلایا کرتے تھے۔ پیپر انیس سو اسّی میں انتقال
کر گئے تھے۔

I believe the only way to resolve this problem would be for Word to change the way in which it performs *kashida* justification. In effect, I think this will mean applying OTL features twice to justified Arabic text: once for unjustified text to get correct Arabic shaping and to determine where and how many *kashidas* need to be inserted to justify the text, and then again after *kashida* insertion to correctly shape and connect the *kashidas* to the letters.