L2/15-062

Subject: Ramifications for UAX 31 of "use of ZWJ/ZWNJ ahead of virama"
From: Mark Davis
Date: 2015-02-02

I have the following action.

139	A086	Mark Davis	Add information about use of ZWJ/ZWNJ ahead of virama in section 2.3 of UAX #31, for Unicode 8.0.

However, the action is not abundantly clear what "information" needed to be added. So, poking around, I found that after the action in the minutes (http://www.unicode.org/L2/L2014/14100.htm), there was:

> **[139-M1] Motion:** Adopt the proposal restricted to only apply to Malayalam conjoining forms in PRI #263, and to not depend upon PRI #37, for Unicode 8.0.

The term "the proposal" seems to be referring to "Section 7 Proposal" of unicode.org/L2/L2004/04279-zwj-indic.pdf , which is:

The characters ZWJ and ZWNJ can constrain the possible renderings as follows:

- For all Indic scripts, ZWNJ can be used in a sequence < C1, virama, ZWNJ, C2 > to explicitly restrict the display to the level-3 alternative, the overt halant form. No other function for ZWNJ is defined.
- For C1 a C1-conjoining consonant, ZWJ can be used in a sequence < C1, VIRAMA, ZWJ, C2 > to restrict the display to level 2 or level 3. Specifically, this sequence requests the half form of C1, to be combined with the full form of C2. If C1 has no half form, then fallback to the level 3 display is used.
- For C2 a C2-conjoining consonant, ZWJ can be used in a sequence < C1, ZWJ, VIRAMA, C2 > to restrict the display to level 2 or level 3. Specifically, this sequence requests the sub- or post-base form of C2, to be combined with the full form of C1. If C2 has no sub- or post-base form, then fallback to the level 3 display is used.
- For a C1-conjoining consonant, the sequence < C, VIRAMA, ZWJ > can be used to display the half form of C in isolation.
- For a C2-conjoining consonant, the sequence < SPACE, ZWJ, VIRAMA, C > can be used to display the sub- or post-base form of C in isolation.

(I added linebreaks before the bullets in the following, since the copy from PDF has ugly formatting.)

It appears from that text, that the only allowed representations with ZWNJ/ZWJ would be:

1. Letter Virama ZWNJ Letter // corresponds to A2 in UAX31, **but with extra final Letter**
2. Letter Virama ZWJ Letter // corresponds to B in UAX31, **but with extra final Letter**
3. Letter ZWJ Virama Letter // **not in UAX31**
4. Letter Virama ZWJ // corresponds to B, **without final Letter**
5. Space ZWJ Virama Letter // not applicable to UAX31, since spaces aren't allowed in IDs.

So it appears that the action would be to add a B2 rule in UAX 31 to account for #3.

However, I have some additional questions.

1. The action says “ZWJ/ZWNJ ahead of virama”, but it appears that only ZWJ is needed in #3.
2. Allowing ZWJ/ZWNJ in general identifiers is problematic, because they are in most contexts invisible. So we restrict the contexts for it.
    a. If a Letter is really required after ZWNJ, it appears that we should amend A2 to be more restrictive.
3. I want to confirm again that a final ZWJ (#4) is required in identifers.
    a. If it really is, we should have an example that requires it.
    b. Otherwise, we should amend B to require a final Letter.

However, because of the sensitivity of this area and relations to IDNA, we might want to have a longer PRI for it to allow for more consideration by parties involved, and thus make no changes in v8.0, but rather post them for comment right afterwards, targeted at 9.0.