L2/18-046

Re: Virama + ZWJ + Vowel_Dependent
From: Mark Davis
Date: 2018-01-21

The following action was unclear to me, and the more I looked into it, the more I think it might touch more places. So I wrote up a strawman for discussion in the UTC.

> [153-A34] Action Item for Mark Davis, Editorial Committee: Document in Proposed Update UTS #39 that virama followed by ZWJ followed by IndicSyllabicCategory=Vowel Dependent, and describe how to check for that sequence, for Unicode 11.0.

The action was in the context of the discussion of Peter C's document http://www.unicode.org/L2/L2017/17365-visual-ambiguity-indic.pdf, which points out the visual ambiguity with cases like GA + virama + zwj + VOWEL SIGN AA.

I'm guessing that the action should be:

> [153-A34] Action Item for Mark Davis, Editorial Committee: Document in Proposed Update UTS #39 and Proposed Update UAX #31 that the following sequence should be disallowed: virama followed by ZWJ followed by a character with property value Indic_Syllabic_Category=Vowel_Dependent; and describe how to check for that sequence, for Unicode 11.0.

Background

UTS #39 doesn't currently talk about ZWJ; that is deferred to UAX #31. The latter has:

> **B. Allow ZWJ in the following context:**
> **In a conjunct context.** That is, a sequence of the form:
> ○ A Letter, followed by a Virama, followed by a ZWJ
> This corresponds to the following regular expression (in Perl-style syntax): **/$L $V ZWJ/** where:
> $L= [:General_Category=Letter:]
> $V = [:Canonical_Combining_Class=Virama:]

There is also the following having to do with Virama.

> **A2. Allow ZWNJ in the following context:**
>
> **In a conjunct context.** That is, a sequence of the form:
>
> ● A Letter, followed by a Virama, followed by a ZWNJ
>
> This corresponds to the following regular expression (in Perl-style syntax): **/$L $V ZWNJ**, where:

$L = [:General_Category=Letter:]
$V = [:Canonical_Combining_Class=Virama:]

Strawman proposal

I tried to make minimal changes to the text; clearly certain areas could benefit from further rewrite.

1. modify UAX #31 to change B to

B. Allow ZWJ in the following context:

- A Letter, followed by a Virama, followed by a ZWJ, not followed by a character of type *Indic_Syllabic_Category=Vowel_Dependent*

This corresponds to the following regular expression (in Perl-style syntax): **/$L $V ZWJ (?!$D)/**
where:

$L= [:General_Category=Letter:]

$V = [:Canonical_Combining_Class=Virama:]

$D = [:Indic_Syllabic_Category=Vowel_Dependent:]

2. Allow for combining marks.

This isn't directly related, but as discussed in other contexts, B and A2 need to allow combining marks between the $L and virama or virama and ZWJ/ZWNJ, so that they don't forbid nukta and others. (They have to be allowed on both sides of $V for canonical equivalence.)

So the formulations for B and A2 should

- Replace "followed by a Virama" by "followed by a Virama (optionally preceded or succeeded by certain non-spacing marks)
- Change $V to $C $V $C
- Add $C = [:General_Category=Mn:]

(Note: we could tighten $C by only allowing those with CCC≠0 between the Virama and ZWJ/ZWNJ. So the regex would be C_0$ $V C_1$)

3. Add a section 3.1.1 to
http://www.unicode.org/reports/tr39/#General_Security_Profile

3.1.1 Joining Controls

The determination of whether ZWJ and ZWNJ are allowed or restricted depends on the context, as described in [UAX31]. That is, they are only allowed in the following contexts; otherwise they are Restricted.

/$LJ $T* ZWNJ $T* $RJ/ or

/$L $V ZWNJ/ or
/$L $V ZWJ (?!$D)/

where
$T = [:Joining_Type=Transparent:]
$RJ = [[:Joining_Type=Dual_Joining:][:Joining_Type=Right_Joining:]]
$LJ = [[:Joining_Type=Dual_Joining:][:Joining_Type=Left_Joining:]]
$L = [:General_Category=Letter:]
$V = [:Canonical_Combining_Class=Virama:]
$D = [:Indic_Syllabic_Category=Vowel_Dependent:]

$V Contents

FYI, because of the discussion in discussion around http://www.unicode.org/reports/tr29/proposed.html, I checked the above formulation of $V against

$V' = [[:Indic_Syllabic_Category=Virama:][:Indic_Syllabic_Category=Invisible_Stacker:]]

As it turns out, $V' is a proper subset of $V. The following characters are in $V but not in $V'

U+0D3B	MALAYALAM SIGN VERTICAL BAR VIRAMA
U+0D3C	MALAYALAM SIGN CIRCULAR VIRAMA
U+0E3A	THAI CHARACTER PHINTHU
U+0F84	TIBETAN MARK HALANTA
U+103A	MYANMAR SIGN ASAT
U+1714	TAGALOG SIGN VIRAMA
U+1734	HANUNOO SIGN PAMUDPOD
U+1BAA	SUNDANESE SIGN PAMAAEH
U+1BF2	BATAK PANGOLAT
U+1BF3	BATAK PANONGONAN
U+2D7F	TIFINAGH CONSONANT JOINER
U+A806	SYLOTI NAGRI SIGN HASANTA
U+A953	REJANG VIRAMA
U+ABED	MEETEI MAYEK APUN IYEK
U+1107F	BRAHMI NUMBER JOINER
U+11134	CHAKMA MAAYYAA
U+112EA	KHUDAWADI SIGN VIRAMA

U+1172B	AHOM SIGN KILLER
U+11A34	ZANABAZAR SQUARE SIGN VIRAMA
U+11D44	MASARAM GONDI SIGN HALANTA

From Cibu:

Couple of thoughts from the Malayalam point of view:
Following contexts should be allowed for requesting reformed or traditional conjuncts as per Unicode10.0.0/ch12 page 505. This needs to be called out in the section 3 of the above proposal.
/$L ZWNJ $V $L/
/$L ZWJ $V $L/

Also, when we disallow /$L $V ZWJ $D/, it is disallowing the sequences involving legacy chillus. That is, for example, <CHILLU N, VOWEL SIGN E> is a valid sequence (Examples in Unicode10.0.0/ch12 Table 12.36). It's legacy equivalent would be <NA, VIRAMA, ZWJ, VOWEL SIGN E>. It might be OK to disallow this; but, we should be mindful of this side effect.