Rendering Issues due to "ZWJ" in Sinhala Domain Names and Impact of IDNA 2003 and IDNA 2008 on Sinhala Domain Names Implementation

Harsha Wijayawardhana B.Sc. (Miami), CITP (UK), FBCS (UK)

COO/CTO Theekshana, Chair, Local Language Working Group and UA Local Initiative

1. Introduction

Sinhala Script is classified under Abiguda Script and has evolved from Brahmi, like all other Indic Scripts. Sinhala Script is a circular-shaped unique script. Zero Width Joiner (ZWJ), a hidden character, has been introduced by Sinhala Unicode implementers for Sinhala rendering of words which can be written in two or three different ways using ligatures or conjuncts. The presence of ZWJ in the Sinhala Unicode sequence indicates to the rendering engine that it must render it as either a conjunct or a consonant conjunct or Repaya. Conjunct also has two types: touching letters, where two letters touch, in ecclesiastic writing in Pali using Sinhala Script, or tying of two letters forming a conjunct without Halantha or Hal kirima. Since Sri Lankan independence, the non-conjunct form (න්ද) with Halantha has become popular and is the default form from the conjunct given here (ඳ). As mentioned, the general public is slowly moving away from the traditional writing of Theravada Pali Buddhist cannon with touching Sinhala letters (ම්ම). Therefore, the breaking or splitting of conjunct forms due to stripping off the ZWJ has no bearing on the current users of Sinhala except for two Consonants Conjuncts, Rakaransaya and Yansaya.

2. Rakaransa and Yansaya Consonant conjuncts

Rakaransaya makes Ka (ක) Kra (ක්‍ර), and the composite glyph, Shri (ශ්‍රී) is made up of three components, the letter Sha (ශ) and bottom Rakar form and the top vowel modifier for long vowel i (ී). Computer stores Rakar form in its persistent memory

as the following Unicode Code Sequence and SLS 1134 version 3 Rakar and Yansa forms are given as the following sequences:

Rakar form: 0DCA 200D 0DBB (්+ ZWJ + ර),

Yansa form: 0DCA 200D 0DBA (් + ZWJ + ය)

In the Sinhala ISO/Unicode standard, the above sequences are known as Named sequences. In addition, ZWJ is used for rendering Repaya or Reph form. The above form replaces Halantha or Hal Kirima (්) and Ra (ර) with a special symbol or Repaya (ී) on top of the consonant, sitting on the right-hand side of Ra with Halantha or Hal Kirima. For instance, හර්ෂ will become හර්ෂී . Computer stores the Reph form in its persistent memory in the following code sequence:

Reph form: 0DBB 0DBB 200D (්+ ර + ZWJ)

3. Stripping Off of ZWJ

ZWJ became a fundamental issue when rendering Sinhala content on the Internet. Due to possible phishing attacks, most web applications began stripping off the ZWJ in a string on the Web. The above resulted in breaking Conjunct forms, both Conjunct Consonants and Reph forms. Conjunct forms and Reph forms did not have any visual impact since, as mentioned, forms with Halantha or Hal Kirima in Sinhala have become the norm or the default form. However, the breaking up of Consonant Conjunct forms became a controversial issue. Let us examine the glyph Shri (ශ්‍රී) to see what happens when ZWJ is stripped off. As mentioned earlier, this letter is made up of the following three components:

Sha(ශ)+ (්+ ZWJ + ර + +ී) = (ශ්‍රී)

When zwj is stripped off, Shri (ශ්‍රී) breaks up into two letters, one with "Hal Kirima" and the other with the "Pilla" on top of Ra (ර). Let us examine how it breaks into the above two forms. In Unicode rendering, a Shaping Engine or Rendering Engine such as Microsoft's Universal Shaping Engine (USE) looks from the first Consonant to the next Consonant to implement vowel modifiers since vowel modifiers are stored in the persistent memory after a consonant. The Named Sequence for Rakar

Form, ්+ ZWJ+ර, after the Consonant, Sha, is taken by the Rendering Engine as a modifier. Hence, with the Pilla for a long I (ී), Shri takes the default or the correct form, ශ්‍රී. However, in the absence or stripped off ZWJ, the Rendering Engine treats Sha (ශ) and Ra (ර) as two independent consonants. In between two consonants, Hal kirima remains, and in the second sequence after Ra (ර), Pilla stays intact. Hence, the string now renders as (ශ්රී).

Currently, most web applications do not strip off ZWJ in Sinhala content on the Internet, except for the desktop version of Facebook. Mobile versions of Facebook render strings with ZWJ accurately.

4. IDNA2003 and IDNA 2008

Internationalizing Domain Names in Application (IDNA) 2003 protocol enabled to have Internationalized Domain Names (IDNs) on the Internet. Dot Lanka and Dot Illangai of Sri Lanka became some of the first non-ASCII-based Domain Names delegated by ICANN in 2010. However, one major issue remained. IDNA 2003 barred or blocked possible strings with ZWJ as either Top Level Domains or from the second or third level. Due to the above, words with "Sri" automatically got barred from being registered as a generic Top Level Domain (gTLD). Therefore, Sri Lanka opted for its Country Code Top Level Domain in Sinhala as Dot Lanka (.ලංකා) without "Sri".

Several Other Scripts, such as Devnagri, Kannada, and Malayalam, use the ZWJ for rendering. Though, the most affected script is the Sinhala Script. For example, when checking a ten million word corpus for words with Rakar and Yansa forms, the following statistics were given:

Yansaya: 251,718 words

Rakaransaya: 503228 words

The total percentage of words with Yansa and Rakar forms came to be roughly 8% in the above corpus. These words originated from Sanskrit and are commonly used every day for formal writing on governance, democracy, media, international

affairs, etc. Therefore, the impact of barring these words from the Root Zone or the Top Level Domain is very high.

With the release of IDNA 2008, it is now allowed to have ZWJ in Domain Names. When Sri Lanka developed its Rules for Sinhala Labels for IDNs, words with ZWJ are barred from registration as gTLDs. However, in Sinhala Label Generation Rules for the Root Zone LGR, ZWJ is allowed in the second or third level. It was recently decided a string with ZWJ and without will be issued to the same customer for the second level as variant strings.

When checking browsers, except for Firefox, others still do not support IDNA 2008. Technically, most online Punycode generators generated the Punycode without ZWJ, only supporting IDNA 2003---a major issue for Sinhala and Sri Lanka.

Conclusion

Sinhala scribes use many words with the two Consonant Conjuncts. Therefore, these words must render correctly. It is more important that these words should render correctly in the second and third levels of a domain. It is the request of Sinhala users to implement IDNA 2008 in all Software web Applications until a permanent solution is found, such as encoding of both Rakaranaya and Yansaya in the Unicode.