Unicode Implementation and Rakaransaya, Yansaya Rendering Issue

Harsha Wijayawardhana B.S. in Biochemistry (Miami), CITP (UK), FBCS (UK)

In collaboration with LK Domain Registry

Introduction

Although Sinhala was encoded in 1997 and published in 1999, implementation of Sinhala Unicode took some time. LLWG spearheaded the implementation phase of Sinhala Unicode, as mentioned before. The most important decision made during the implementation stage was the standardization of the input keyboard and how to store the Sinhala two-byte sequences in computer memory. It took some time for Sri Lankan experts to understand how a rendering engine worked in Unicode. Sinhala has three consonant conjuncts and many conjuncts. Two of the consonant conjuncts do not have non-conjunct forms however, other conjuncts have non-conjunct forms written with Hal Kirima or Halantha.

Before the independence of Sri Lanka from Britain, conjuncts form became the default form like in Devanagari. The default form would change after the independence (may be due to the use of the typewriter, which could not technically type all the conjunct forms) to their non-conjunct forms except for two consonant conjunct forms as mentioned before: Rakaaraansaya and Yansaya. Rakaaraansaya is Hal Kirima with Ra (ර්) and Yansaya is Hal kirima with Ya (ය්). Words with repaya, which is also consonant conjunct, have acceptable non-conjunct forms. In the Sinhala ISO/Unicode standard, the above sequences are known as Named sequences.

Current Rendering of Rakaaraansaya and Yansaya

Unicode consortium advocated in the early two thousand that consonant conjuncts such as Rakaaraansaya and yansaya should not be encoded in Unicode. Although all members of the LLWG were of the view that the best option for rendering the above consonant conjuncts would be to encode them in Unicode, the majority felt that pursuing encoding would stall the implementation of Unicode in Sinhala; therefore, they argued that the LLWG expedite the Unicode implementation. Whereas, I remained a strong advocate of encoding these two consonant conjuncts. The Unicode consortium strongly advised using Zero Width Joiner (ZWJ) and Zero Width Non-Joiner (ZWNJ), typesetting hidden characters, to render strings that could be written using, more than one form. In other words, if two forms are available, one form has to be represented in a string of default code points, and the other form with a string consisting of ZWJ or ZWNJ to make it unique. Prof. Gihan Dias proposed an elegant solution to the problem in 2003, where only ZWJ would be used. I also proposed a solution that was not as elegant as Prof. Gihan's. The first solution I proposed had the default form as the conjunct form, and to depict the current default form, I suggested using ZWNJ. Subsequently, I also agreed with the rest of the members that Prof. Gihan's solution was much better and more elegant.

න්ද : 0db1/0dca/0daf
Letter NA+Hal Kirima+Letter Da
න්ද : 0db1/0dca/200c/0daf
Letter Na+Hal Kirima+ ZWJ+Letter Da

Releasing of Unicode Sinhala Fonts

ICTA and LLWG released version 2 of SLS 1134 as the Sinhala Input standard standardizing Extended Wijesekera keyboard as the default keyboard for Sinhala input for digital devices. SLS 1134 ver. 2 standardized how to store Sinhala characters in computer memory using Consonants and Vowel modifier combinations like all other Indic scripts. SLS 1134 ver. 2 came out in 2004, and the third revision, which I authored, was published in 2011. The above release of the standard, SLS 1134 ver. 2, enabled, for the first time, digital content. It also made it possible to develop Software and OS developers to configure rendering engines to support Sinhala. I tested Sinhala Unicode font shapes rendering with Mr. Winnie Hettigoda's (famous cartoonist) Sinhala font shapes in my lab at the UCSC. I used the Microsoft Visual OpenType Layout tool (VOLT) to have font rules. Mr. Winnie Hettigoda finished his font later, and I also released Sarasavi Unicode, Dr. Nandasara's 7-bit font, as a Unicode font using his Sinhala glyphs, which were designed as a Serif font. In 2005, Microsoft released Pota and, later, Iskoola Pota as their default fonts for their Operating Systems and Applications by Microsoft. In 2009, Unicode pioneers came together to train font makers on how to create Unicode fonts. At the end of the workshop, ICTA released 16 different fonts using font rules created for Bashitha. LLWG advised ICTA to release Bashita font rules as Free and Open Source for font creators to make new fonts (https://www.language.lk/en/download/unicode-fonts/).

Rendering Issue of Rakaaraansaya and Yansaya

Google began striping off ZWJ on its search engines since it was a hidden character to prevent phishing attacks. Although the Unicode consortium recommended we use ZWJ and ZWNJ for rendering rules, most online applications began stripping off ZWJ and ZWNJ. The above resulted in breaking words with Rakaaraansaya and Yansaya, which include the name of our country, Sri Lanka, in Sinhala. Since other conjunct forms had non-conjunct forms, they did not look awkward.

ශ්‍ර (Correct Form) -> ශ+0dca (hal Kirima)+200c (ZWJ)+ර
when 200c is removed: ශ්‍ර
ශ්‍රී ලංකා (correct form): ශ්‍රී ලංකා (Country Name)
සත්‍ය (correct Form) : සත්‍ය (truth in Sinhala)

The LLWG wrote to Google and other online application providers to stop stripping off ZWJ. Most complied, however, the Desktop Application of Facebook still strips off the ZWJ.

Repaya (Consonant Conjunct)

In addition, ZWJ is used for rendering Repaya or Reph form and is one of the three consonant conjuncts Sinhala has. 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 හර්‍ෂ.The computer stores the Reph form in its persistent memory in the following code sequence:
Repaya form: 0dca 0dbb 200D (්+ ර + ZWJ)