Status report on Script Extensions and Indic properties

Roozbeh Pournader, Google Inc.
August 5, 2014

Background

At the UTC meeting #139 (May 2014), the author was assigned the task of further studying some open issues ScriptExtensions for historic scripts and updates to Indic properties and report back to the UTC. This is a quick progress report.

Details

A publicly visible git repository has been created at https://github.com/roozbehp/unicode-data to develop the data files rapidly. Extended data files have been created for ScriptExtensions.txt, IndicSyllabicCategory.txt, and IndicPositionalCategory.txt (renamed from IndicMatraCategory.txt). TODO lists have been created with expected remaining tasks and are available in the same repository.

The following lists some of the changes applied in the above-mentioned repository:

1. In expectation of the expansion of the Indic Matra category to encode the visual position of Indic combining marks, a practice that partially started in Unicode 7.0 with providing Matra categories for various non-Matra characters in Khmer, Tai Tham, Lepcha, and Cham, the file and the property have been renamed “Indic Positional Category”.
2. Various characters, especially Vedic signs, have been provided Indic Positional categories:
   IPC=Right:
   U+1CE1 VEDIC TONE ATHARVAVEDIC INDEPENDENT SVARITA
   IPC=Top:
   U+0951 DEVANAGARI STRESS SIGN UDATTA
   U+0953 DEVANAGARI GRAVE ACCENT
   U+0954 DEVANAGARI ACUTE ACCENT
   U+1CD0 VEDIC TONE KARSHANA
   U+1CD1 VEDIC TONE SHARA
   U+1CD2 VEDIC TONE PRENKHA
   U+1CDA VEDIC TONE DOUBLE SVARITA
   U+1CDB VEDIC TONE TRIPLE SVARITA
   U+1CE0 VEDIC TONE RIGVEDIC KASHMIRI INDEPENDENT SVARITA
   U+1CF4 VEDIC TONE CANDRA ABOVE
   IPC=Bottom

U+0952 DEVANAGARI STRESS SIGN ANUDATTA
U+1CD5 VEDIC TONE YAJURVEDIC AGGRAVATED INDEPENDENT SVARITA
U+1CD6 VEDIC TONE YAJURVEDIC INDEPENDENT SVARITA
U+1CD7 VEDIC TONE YAJURVEDIC KATHAKA INDEPENDENT SVARITA
U+1CD8 VEDIC TONE CANDRA BELOW
U+1CD9 VEDIC TONE YAJURVEDIC KATHAKA INDEPENDENT SVARITA SCHROEDER
U+1CDC VEDIC TONE KATHAKA ANUDATTA
U+1CDD VEDIC TONE DOT BELOW
U+1CDE VEDIC TONE TWO DOTS BELOW
U+1CDF VEDIC TONE THREE DOTS BELOW
U+1CED VEDIC SIGN TIRYAK

IPC=Overstruck

U+1CD4 VEDIC SIGN YAJURVEDIC MIDLINE SVARITA
U+1CE2 VEDIC SIGN VISARGA SVARITA
U+1CE3 VEDIC SIGN VISARGA UDATTA
U+1CE4 VEDIC SIGN REVERSED VISARGA UDATTA
U+1CE5 VEDIC SIGN VISARGA ANUDATTA
U+1CE6 VEDIC SIGN REVERSED VISARGA ANUDATTA
U+1CE7 VEDIC SIGN VISARGA UDATTA WITH TAIL
U+1CE8 VEDIC SIGN VISARGA ANUDATTA WITH TAIL

3. After a better understanding of the interaction of Indic cantillation marks, the Indic Syllabic Categories for the following characters have been changed from Tone_Mark to Cantillation_Mark. These characters apply to the whole syllable, similar to the combining Devanagari digits and letters which already had the property value of Cantillation_Mark:

0951	DEVANAGARI STRESS SIGN UDATTA
0952	DEVANAGARI STRESS SIGN ANUDATTA
1CD0..1CD2	VEDIC TONE KARSHANA..VEDIC TONE PRENKHA
1CD4..1CE0	VEDIC SIGN YAJURVEDIC MIDLINE SVARITA..VEDIC TONE RIGVEDIC KASHMIRI INDEPENDENT SVARITA
1CE1	VEDIC TONE ATHARVAVEDIC INDEPENDENT SVARITA
1CF4	VEDIC TONE CANDRA ABOVE
1CF8	VEDIC TONE RING ABOVE
1CF9	VEDIC TONE DOUBLE RING ABOVE

4. Evidence for the usage of various Vedic signs in non-Devanagari scripts were found by the author by going through all the related proposals in the UTC and WG2 documents, and by browsing Vedic text in various scripts publicly available on the web. Based on that research, the ScriptExtension properties of the following characters have been extended to include scripts for which there is evidence of usage:
    a. Devanagari and Grantha: 1CD0, 1CD2, 1CD3, 1CF2, 1CF3, 1CF4, 1CF8, 1CF9

b. Devanagari and Kannada: 1CF5
c. Devanagari and Sharada: 1CD7, 1CD9, 1CDC, 1CDD, 1CE0
d. Devanagari and Telugu: 1CDA
e. Devanagari, Grantha, and Latin: 20F0
f. Devanagari, Grantha, Latin, and Telugu: 0952
g. Devanagari, Grantha, Latin, Sharada, and Telugu: 0951

The author is continuing the effort, but cannot guarantee that all the work would be finished by the time Unicode 8.0 goes to beta. He expects this to be an ongoing effort, especially since the understanding of the less frequent cases is limited and all potential contributors (including the author) appear to be very busy with other projects. Still, apart from the contributors to L2/14-126R (especially Behdad Esfahbod, Andrew Glass, and Anshuman Pandey) contacts have been made with Shriramana Sharma and Peter Scharf asking for their help to help improve the three above properties for Indic characters.

Bibliography

1. Roozbeh Pournader and Behdad Esfahbod. 2014. “A bag of suggested improvements to Unicode’s provisional Indic properties.” UTC Document Register L2/14-065, The Unicode Consortium. http://www.unicode.org/L2/L2014/14065-indic-properties.pdf
2. Roozbeh Pournader. 2014. “Improvements requested for Unicode Indic properties.” UTC Document Register L2/14-126R, The Unicode Consortium. http://www.unicode.org/L2/L2014/14126r-indic-properties.pdf