17. Standards for Visual Sequence of characters for Indic Scripts

Mr. Sandeep Rao, Staff Scientist, Mr. Vinod Kumar, Software Specialist, CDAC,
E-mail : {sandeep, vinod}@ncst.ernet.in

I. INTRODUCTION

For Indic scripts, the Unicode standard [1] has addressed the issues related to the coding of characters, the logical representation of text and the rendering of the text. While the coding of characters and the logical sequence are formally defined, the specifications for the rendering of the text is driven by examples. informal, and incomplete for many scripts.

For example, Unicode formally specifies that the code in hexadecimal for the Tamil letter Ka is 0b95 and the code for Tamil vowel sign Ai is 0bc8. It specifies that the logical sequence of characters for the Tamil text KAi is <0b95 0bc8 >. The logical sequence in terms of glyphs would be < ">.In "Table 9-8 Vowel Reordering" (Section 9.6) the standard [1] shows that the text should be displayed as the visual sequence ". The vowel reordering is in the glyph domain and not in the character domain. This is clear from the paragraph that describes how split matras are displayed. The single glyph that represents the split matra character is transformed into two equivalent glyphs and then the glyphs are subject to vowel reordering. One of our suggestions is to formalize such vowel reordering in the character domain itself. The new table entry for the Tamil sequence is shown in Table I. We refer to characters with their hexadecimal codes (eg. 0b95) specified in the Unicode standard. For clarity, the shortened Unicode names (eg. Ka instead of Tamil Letter Ka) may sometimes be used.

Some specifications related to the visual form can only be shown graphically. But many other features of the visual sequence, like the vowel reordering, can be expressed formally in terms of the Unicode characters. In practice, the initial section of the Indic shaping pipeline does precisely such character level reordering. Microsoft Typography [4] has come out with character level reordering rules for most Indic scripts based on Unicode standards. However, the informality of the Unicode specifications for the visual sequence, and interference of the lower level OpenType font on the reordering, make the visual sequence recommended by Microsoft Typography amenable for much improvement.

Logical sequence	Visual sequence
0b95 0bc8	0bc8 0b95

TABLE I

TAMIL VOWEL REORDERING IN TERMS OF CHARACTERS

We plan to base the characters to glyphs transformation of Indic text on the Intelligent font layout and presentation model of ISO/IEC Technical Report 15285 [3]. The logical sequence of characters for Indic scripts is phonetic. The visual sequence, on the other hand, need not be phonetic in general as shown in the Tamil example KAi. The Unicode standard has attempted to explain the formation of visual sequence from the logical sequence by a set of convoluted rules. For example, see the Consonant RA Rules R2 - R8

in Section 9.1 of the Unicode standard [1]. Since the Unicode standard started with Devanagari, its treatment of other Indic scripts appears stretched. An incisive analysis in [5], albeit with respect to ZWJ, reinforces the idea that a logical sequence of characters should better first be transformed to a visual sequence of characters before the final visual sequence of glyphs is generated. Baggages associated with phonetic order can be dropped when the visual order is discussed.

II. TEXT RENDERING BASED ON ISO/IEC 15285

The most basic capability needed in text processing is the one to convert a logical sequence of characters to a renderable sequence of glyphs from a font. Rendering Latin text is straight forward, as each character can be transformed to a glyph using the CharMap in the font. For Latin, the transformation from characters to glyphs is one-to- one. The unit of orthography in Indic scripts is the syllable. An Indic syllable consists of one or more characters that can be rendered by a sequence of glyphs. The interactions between characters will be confined to be within the syllable. A character within a syllable will not move to or change the shape of the preceding or succeeding syllable. Thus _ Indic characters map to _ glyphs. This is the most general mapping from characters to glyphs advocated in ISO/IEC TR 15285 [3].

The object model of the ISO/IEC TR 15285 is shown in Fig. 1. A **Syllable** can be **Logical** syllable or a **Renderable** syllable. Further, a **Logical** syllable can be an ordered sequence of Character, or an ordered sequence of Glyph. A Renderable on the other hand is an ordered sequence of **Glyph**.
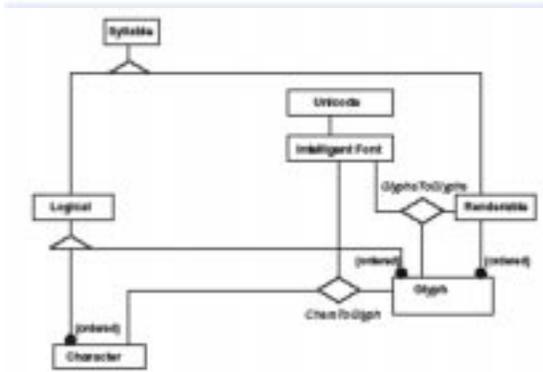
Fig. 1. Object Model for characters and glyphs. ISO TR 15285

An Intelligent Font with additional information on how a sequence of coded characters is transformed into a sequence of glyph identifiers, with associated position information implements the many-to-many CharsToGlyph and GlyphsToGlyphs relations. These transformations are carried out by a pipeline. The pipeline will have to be presented with nothing short of a syllable.

Operationally, the **Logical** syllable of characters is first transformed, character by character, to a **Logical** syllable of glyphs, using the CharMap in the font. The **Logical** syllable of glyphs is then transformed into a **Renderable** syllable of glyphs by reordering the glyphs, and substituting them, using substituting and positioning tables from the font.

III. DRAWBACKS OF ISO/IEC 15285 FOR INDIC SCRIPTS

Operation like reordering of glyphs is difficult with an Intelligent Font like OpenType. It is easier done in the character domain. So we enhance the ISO/IEC 15285 with a **Visual** syllable as a stepping stone between the **Logical** and **Renderable** syllables. Further, characters are grouped into subclusters. These two enhancements are shown in Fig. 2. The **Subcluster** of characters is introduced to treat certain character sequences as a single unit. The **Visual** syllable is for formalizing the reordering rules on basis of the subclusters. For example, consider the Devanagari logical sequence of characters <930 94d 915> = < Ra Halant Ka >. Here the character sequence <930 94d> forms a **SubCluster**. The **Logical** syllable of subclusters is <<930 94d> 915> . The Visual syllable of subclusters is <915 <930 94d>>
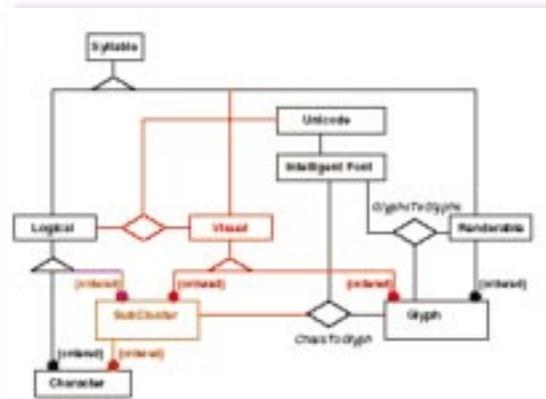

Fig. 2. Enhanced Object Model for characters and glyphs.

IV. TWO ENHANCEMENTS TO THE ISO/IEC 15285 MODEL - SUBCLUSTER AND VISUAL SYLLABLE

In Indic scripts, there are many cases when a group of characters are represented visually as a single glyph. Consonants separated by halant character may form a single ligature as with akhands. Or, a consonant may combine with the succeeding halant sign to a half form as in Devanagari. In scripts like Kannada, a consonant combines with the preceding halant to take a subscript form. A consonant and a vowel sign may combine. Lastly, some consonants combine with the halant signs to form consonant marks. The consonant marks should be distinguished from half or subscript forms. They are more akin to vowel signs.

In Indic scripts, the vowel signs and the consonant marks occasionally need reordering. For example, in Devanagari, the IMatra has to be relocated to the left of the first consonant. The **Visual** syllable specifies how the reordered character sequence looks like. The **Subcluster** is an artifact introduced to facilitate the proper departure and homing of these relocating marks. A relocating vowel sign should not squeeze in between a halant and a consonant that plan take a subscript form. Similarly all the component characters of a consonant mark should relocate together as a single unit.

V. SPECIFICATIONS FOR VISUAL SYLLABLE

The specification is embodied in several auxiliary tables and a main table. The important auxiliary tables are:

1) A **Split Matras** table that maps between a single character for a vowel sign and the characters for its constituent signs.

2) A **Akhands Proxy** table that maps between a sequence of characters for an akhand symbol and a single proxy character for the sequence.

3) A **Consonant Matra Proxy** table that maps between a consonant halant sequence and a single consonant-sign proxy character.

4) A **Character Class** table that maps each character to one or more classes.

The main **Visual Syllable** table specifies the structural relationship between the logical and visual syllables. The character in the **Character Class** and **Visual Syllable** tables may be a proxy character generated using the auxiliary tables. A proxy character uses reserved positions in the Unicode script range. Some of them have been recommended by TDIL [2].

The procedure to convert the **Logical** syllable of characters to a **Visual** syllable of characters has the following important steps.

[1] Substitute split matras with their constituents using **Split Matras** table.

[2] Replace each akhand sequence with single akhand character proxy using **Akhands Proxy** table.

[3] Replace each consonant halant sequence with a consonant matra proxy using **Consonant Matra Proxy** table.

[4] Scan the input logical syllable. Each character obtained during the scan is checked for relocatability in the **Visual Syllable** table. If relocatable, then the input syllable is parsed according to the logical syllable template in the table and converted into the visual form according to the visual syllable template.

[5] Replace each consonant matra proxy with its sequence.

[6] Replace each akhand character proxy with its sequence.

The proxy is a coding of some character sequences convenient in the reordering logic. It does not have the backing of the Unicode standard. That is why we replace all the proxies with their original characters. The procedure is entirely table driven and independent of scripts. All script dependent features are embodied in the tables. So a specification for forming the visual syllable for an Indic script involves stating these tables alone.

VI. VISUAL SYLLABLE FOR DEVANAGARI

There is no split matra in Devanagari. The first entry in the Devanagari Consonant Matra Table III means that if you find a sequence <94d, 930> after a consonant then it can be replaced with a Ra vattu consonant sign of code 97d.

In the Devanagari Character Class Table IV the proxy for Reph has been classified as a matra. The first entry in the Devanagari Visual syllable Table V implies that I Matra, shown in boldface, is to be relocated leftward before first consonant.

VII. CONCLUSION

For Indic scripts, the visual order of displayed glyphs often differs from the order obtained by displaying the glyphs of logically ordered characters. Specification of the visual order in the glyph domain is informal. Specification of the visual order in the character domain should depend only on the properties of the script and the minimal model of sequential rendering of glyphs. In this proposal, we have extended the Unicode specifications from logical order to visual order in the character domain. The visual order for Devanagari is shown.

VIII. REFERENCES

[1] The Unicode Standard, Version 4.0: The Unicode Consortium. Addison-Wesley, Reading, MA, USA, 2003.

[2] Unicode Standard for Indian Scripts. http://tdil.mit.gov.in/pchangeuni.htm, 2004.

[3] Edwin Hart and Alan Griffee. An operational model for characters and glyphs. Technical Report ISO/IEC TR 15285, 1998.

[4] Microsoft Corporation. Developing OpenType Fonts for Indic Scripts. http://www.microsoft.com/typography/SpecificationsOverview.mspx, 2004.

[5] Peter Constable, Microsoft. Proposal on Clarification and Consolidation of the Function of ZERO WIDTH JOINER in Indic Scripts. Technical report, http:unicode.org/, June 2004.

Akhand		Notes
proxy	sequence	
0972	0915, 094d, 0937	KSsa from TDIL reco
0973	091c, 094d, 091e	JNya from TDIL reco

TABLE II

DEVANAGARI AKHANDS PROXY

Context	ConsonantMatra		Notes
	proxy	sequence	
After Cons	097d	094d, 0930	Ra Vattu from CDAC-Mumbai reco
Before Cons	0975	0930, 094d	Reph from TDIL reco

TABLE III

DEVANAGARI CONSONANT MATRA PROXY

Character	Class	Notes
0905	Vowel	Deva Letter A
⋮		
0915	Cons	Deva Letter KA
⋮		
093f	Matra	Deva Vowel Matra I
0975	Matra	Reph Proxy
⋮		

TABLE IV

DEVANAGARI CHARACTER CLASS

Relocating character	093f (Vowel sign I)
Logical syllable	$< Cons_{first} \cdots 093f >$
Visual syllable	$< 093f \; Cons_{first} \cdots >$
Relocating character	0975 (Proxy Reph mark)
Logical syllable	$< 0975 \cdots Cons[Matra]_{last} >$
Visual syllable	$< \cdots Cons[Matra]_{last} \; 0975 >$

TABLE V

DEVANAGARI VISUAL SYLLABLE