

# Blending Content for South Asian Language Pedagogy Part 2: South Asian Languages on the Internet

A. Sean Pue

South Asia Language Resource Center Pre-SASLI Workshop  
6/7/09

# Objectives

- To understand how South Asian languages work on the internet.
- To understand what Unicode is.
- To understand how to create webpages in South Asian languages.

# What is a digital text?

- Any text that you see on a computer—on a webpage, in a word processor, and so on—is treated by the computer program as a series of numbers.

# Text as Numbers

- For example, the word “Text” is a sequence consisting of the following numbers:
  - T: 84,
  - e: 101,
  - x: 120,
  - t: 116.
- The number a character refers to is called the character’s *codepoint*.
- The codepoint of a character is usually represented using hexadecimal (base 16) notation instead of decimal (base 10).

# Numbers in hexadecimal

- Hexadecimal (base 16) : uses the numbers 0-9 plus the letters A-F [coined in the 1960s to replace sexadecimal]
- Decimal 5 = Hexadecimal 5  
Decimal 9 = Hexadecimal 9  
Decimal 10 = Hexadecimal A  
Decimal 11 = Hexadecimal B  
Decimal 12 = Hexadecimal C  
Decimal 15 = Hexadecimal F  
Decimal 16 = Hexadecimal 10
- Why is it used? It's easy to read and computers like the number sixteen (it's  $2 \times 2 \times 2 \times 2$  after all!).

# Text in Hexadecimal

- T= 0054
  - e = 0065
  - x = 0078
  - T = 0074
- 
- Note that the codepoints are normally 4 characters long

# What is Character Encoding?

- Remember all characters are treated as numbers.
- Character Encoding refers to the underlying standard or code that governs which character will be represented as what number. An example is the Indian Script Code for Information Interchange (ISCII), which provides a character encoding for many South Asia languages.
- The most frequently used and versatile contemporary standard for character encoding is Unicode.

# What is Unicode?

- Unicode is an evolving industry standard for character encoding that maps the characters of nearly all languages and scripts as a specific number. This unique number provides a universal reference for storing, sorting, and manipulating texts.
- Unicode is the current cross-platform standard for character encoding. All contemporary operating systems (i.e. OS X, Windows, Linux) and most contemporary computer programs process text following this standard by default.

# What is in Unicode?

- With Unicode, there is no limit to how many characters can be added to the standard. There are currently over 100,000 characters mapped in Unicode from over thirty writing systems.
- Unicode codepoints are usually referred to in their hexadecimal form, often preceded by “U+” as in U+200D. Each codepoint also has a character name.

# Activity

- Unicode is divided based on script and not language. Explore the Unicode website and find the chart that describes your language's script:

<http://www.unicode.org/charts/>

# Unicode is a standard

- Unicode is only a standard for what number refers to what character. The actual display of characters involves a font.

# What is a font?

- A computer font is a data file containing rules for how to display certain characters in a particular typeface. These characters are referred to by numbers.

# Types of Fonts

- There are a number of different font formats:
  - OpenType (Microsoft + Adobe): The most versatile, supported by the three major operating systems (Windows, Linux, and OS X)
    - Unfortunately, complex scripts using OpenType is not as complete on Apple's OS X as in Windows and Linux
  - Apple Advanced Typography (AAT): Apple only. Good but only works on OS X
  - TrueType: Predecessor of OpenType; can't do complex scripts

# Fonts Before Unicode

- Before Unicode, most character encodings could only deal with a character map of 256 entries. To render complex scripts, specific fonts were developed to display specific languages. For a number that would normally display a letter like “T” for example they would put in a character or series of characters from their font.
- There are a lot of websites that still require you to “download a font”. That’s why.
- There is an extension to the Firefox browser called Padma that can convert a number of fonts to Unicode:  
<http://padma.mozdev.org/>

# What is a Unicode Font?

- A 'Unicode font' is a font that conforms to the Unicode standard.
- Like any other font, it provides rules for how to display characters, which are mapped as specific numbers. In a Unicode font, these numbers follow the Unicode standard.
- A Unicode font may contain the rules for displaying numerous scripts, or maybe just one. For example, the Tahoma font contains rules for how to display the Latin alphabet, as well as the scripts used by Urdu and Thai.

# Why use Unicode?

- If you write documents using Unicode, your text can be displayed in numerous Unicode fonts. If you send a document or put it up on the web and your readers do not have the same font you used when writing the text, they can still read it, provided they have some other Unicode font that can display those characters.

# How do you type in Unicode?

- In order to type in Unicode, one needs to install or enable an appropriate keyboard. What a keyboard does is map a specific character or sequence of characters to a unicode codepoint.
- The latest version of most current operating systems (Windows, OS X, Linux) already contain keyboard layouts for most South Asian scripts.
- Moreover, there are often different layouts available for the different scripts, such as a phonetic option or one closer to the layout used in the type writer of the native script.

# Where do I find more information?

- Most operating systems also have an onscreen keyboard feature, which is helpful while learning to type.
- The South Asia Language Resource Center lists a number of keyboard options on its fonts website.
- There is a lot of information online.
- Bring in your computers tomorrow.

# How Does Unicode Handle South Asian Languages?

- As a standard, Unicode is more concerned with characters and scripts than with languages. The characters used by different South Asian languages are therefore organized by script.
- Most often, characters in the same script have numbers that are next to each other in a character block .

# Where is my character?

- Currently (in Unicode 5.1) there are twelve “Indic script” character blocks in Unicode. They are: Bengali, Devanagari, Gujarati, Gurmukhi, Kannada, Limbu, Malayalam, Oriya, Sinhala, Syloti Nagri, Tamil, and Telugu.
- The characters used by right-to-left languages like Urdu, Pashto, and Sindhi are found on the Arabic character code chart.
- Tibetan has its own code chart.

# What are the issues with Left-to-Right South Asian scripts? 1

- The twelve “Indic scripts” and Tibetan all function in a similar way in Unicode, as the writing systems themselves are similar.
- All of these scripts are segmented in a such a way that one or more consonants are written as a cluster followed by a specific vowel sign, and consonants in a cluster will change their shape—either assuming a half-form or a combined form—before a following vowel.
- Unicode is concerned only with the underlying letter and not with the shapes they assume in a word—that’s left up to the font.

# What are the issues with Left-to-Right South Asian scripts? 2

- Because Unicode is only concerned with letters and vowels, each script is assigned only up to 128 codepoints even though there are many more possible visual forms for these letters.
- For left-to-right South Asian scripts, Unicode contains codepoints for independent vowels (used when a vowel is not preceded by a consonant), consonants, and dependent vowel signs (used when vowels follow consonants), as well as digits and other “various signs.” Common punctuation marks are found in the Devanagari range.

# How to combine LTR letters in Unicode

- There are a few ways to indicate that consonants should be combined, and they all involve inserting certain characters in between the consonants. The first of these characters is actually a sign in the left-to-right scripts referred to as a virām, halant, or hasant. All of the left-to-right South Asian scripts in Unicode contain this character.

# How to combine LTR letters in Unicode 2

- Here are the combining characters:
- U+094D DEVANAGARI SIGN VIRAMA  
U+09CD BENGALI SIGN VIRAMA  
U+0A4D GURMUKHI SIGN VIRAMA  
U+0ACD GUJARATI SIGN VIRAMA  
U+0B4D ORIYA SIGN VIRAMA  
U+0BCD TAMIL SIGN VIRAMA  
U+0C4D TELUGU SIGN VIRAMA  
U+0CCD KANNADA SIGN VIRAMA  
U+0D4D MALAYALAM SIGN VIRAMA  
U+0F84 TIBETAN MARK HALANTA  
U+A806 SYLOTI NAGRI SIGN HASANTA

# How to combine LTR letters in Unicode 3

- When a VIRAMA Unicode character is added in between consonants, the consonants will be displayed in their conjoined form. For example, the typical way to write the Hindi Word *kyā* (क्या) is:

- U+0915 क DEVANAGARI LETTER KA
- U+094D DEVANAGARI SIGN VIRAMA
- U+092F य DEVANAGARI LETTER YA
- U+093E ट DEVANAGARI VOWEL SIGN AA
- = क्या

# Summary for LTR Languages

- In summary, the important aspect of the way Unicode deals with left-to-right South Asian scripts is that the conjunct forms are created by adding a VIRAMA character between the consonants. There is no “क्य” character. Instead, it’s क + virama + य.
- There are distinct codepoints for combining vowels and independent vowels, i.e. ठ and आ.

# What are the issues for Right-to-left languages?

- Right-to-left South Asian languages use scripts derived from the script used for Arabic. The special characters of South Asian languages have unique codepoints in Unicode.
- Unicode does not distinguish between variants of right-to-left scripts, such as nastaliq or naskh.
- It is only concerned with the underlying letters and signs, not with the shape in which they combine together.
- Therefore, there is no need to worry about initial, medial, or final forms of letters when entering text.

# What are the issues for Right-to-left languages? 2

- Example:
  - The fake word **بيب** consists only of  
U+0628 **ب** ARABIC LETTER BEH (3 times)

# RTL Ambiguities

- There appears to be a certain amount of ambiguity in the Arabic code chart because a number of different characters look the same.
- However, the mapping of South Asian right-to-left languages is now more or less standardized in practice.
- One reason is the exclusion from popular fonts and keyboards of certain codepoints deemed inappropriate for South Asian languages as they are more properly part of Arabic.

# RTL South Asia Standards

- While the letter ARABIC LETTER FARSI YEH (U+06CC) looks identical to ARABIC LETTER ALEF MAKSURA (U+0649), ALEF MAKSURA is often not included in current fonts. While ARABIC LETTER YEH (U+064A) looks the same as ARABIC LETTER FARSI YEH in initial and medial positions, it is best avoided, because many fonts for South Asian languages do not include it.
- Similarly, ARABIC LETTER HEH DOACHASHMEE (U+06BE) should be used in place of ARABIC LETTER HEH (U+0647), even though they look identical. For the regular heh, the preferred codepoint is ARABIC LETTER HEH GOAL (U+06C1).

# RTL Standards 2

- The Center for Research in Urdu Language Processing (<http://crulp.org>), which is leading the development of fonts and keyboard for Pakistani languages on a whole, has managed to curtail the use of numerous “Arabic” characters by not including them in their fonts or eliminating them from their keyboards. As a result, Urdu and other Pakistani languages on the web are more standard than Persian, for example.

# RTL Issue: Hamza

- The Unicode codepoint used for the “seated” hamza is ARABIC LETTER YEH WITH HAMZA ABOVE (U+0626). It will look a yeh with a hamza above it on the keyboard.
- The ARABIC LETTER HAMZA (U+0621) is used in the standalone, unconnected position.
- There is a third codepoint, ARABIC HAMZA ABOVE (U+0654), which renders a hamza above a letter, such as BAREE YEH, VAO, and so on.

# How to put Unicode on a webpage?

- For the purposes of material developed online for SASLI, you can just type in Unicode and it will be saved appropriately on the webpage.
- The reason is that the webpage is already setup for what is called the “utf-8” character set.
- As long as the header of the HTML or XHTML file shows that it will be encoded in the unicode character set, you are good to go.
- (Activity: Show how this done and demonstrate)