Unicode Introduction

Ken Zook

November, 2006

Unicode properties

0041;LATIN CAPITAL LETTER A;Lu;0;L;;;;;N;;;;0061;

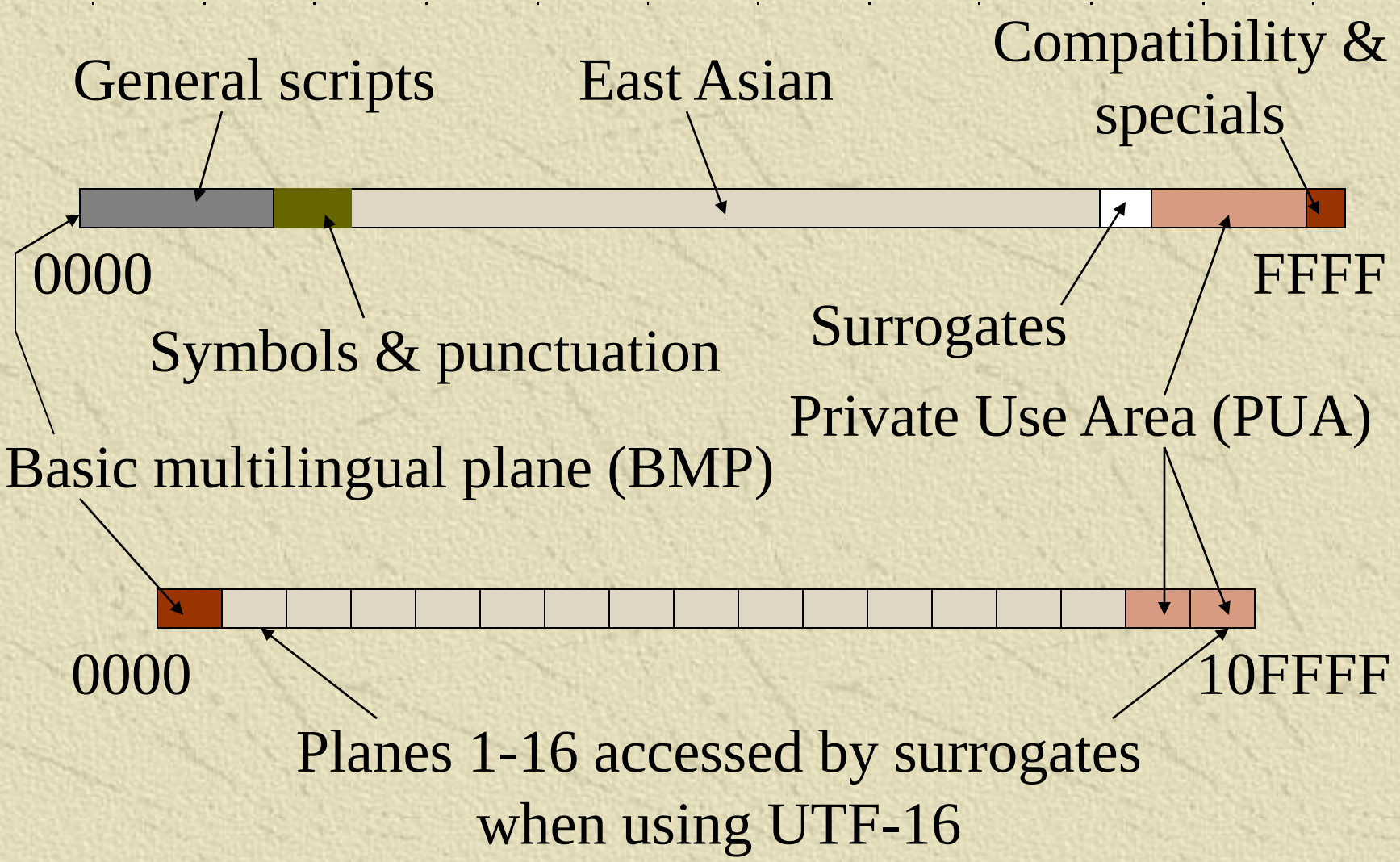
Representative glyph

A

Semantic properties

Code point: 0041
Name: LATIN CAPITAL LETTER A
General category: Uppercase letter (Lu)
Canonical combining class: Standard spacing (0)
Bidirectional category: Left-to-right (L)
Mirrored: no (N)
Lowercase mapping: 0061

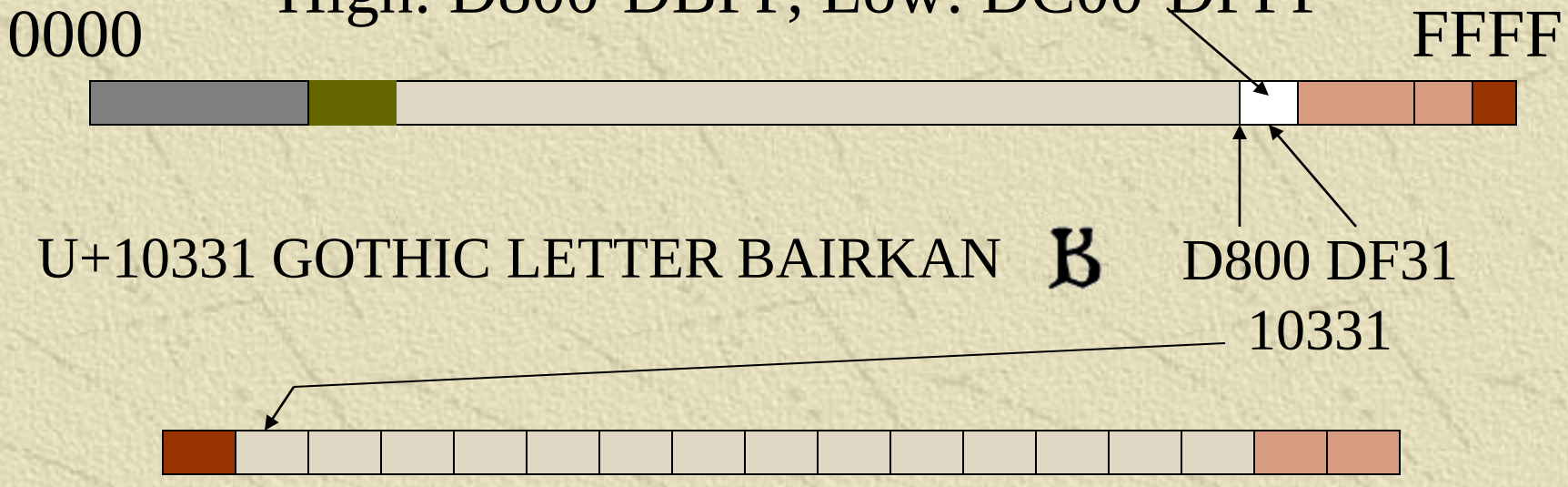
Unicode code space



Encoding Unicode

UTF-32 = 10331 (1 32-bit value / code point)
UTF-16 = D800 DF31 (FW/Win) (1-2 16-bit values / code point)
UTF-8 = F0 90 8C B1 (XML) (1-4 8-bit values / code point)

UTF-16 Surrogates: D800-DFFF
High: D800-DBFF, Low: DC00-DFFF



Surrogates used to access 10000-10FFFF in UTF-16

Private Use Area (SIL)

International PUA: F100-F8FF (2,047)

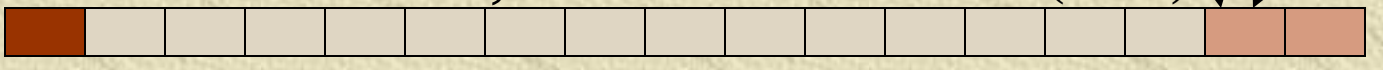
Entity PUA: E000-EFFF (4,095)

PUA: E000-F8FF (6,400)

E010 (Philippines) maps to F2010

E010 (Russia) maps to F1010

PUA: F0000-FFFFD, 100000-10FFFD (131K)

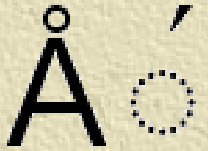


Unique entity mappings in upper PUA

Canonical equivalence

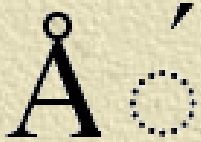
01FA

LATIN CAPITAL LETTER A WITH RING ABOVE AND ACUTE



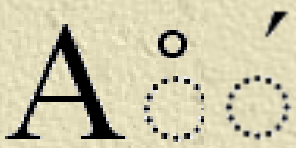
212B 0301

ANGSTROM SIGN
COMBINING ACUTE ACCENT



00C5 0301

LATIN CAPITAL LETTER A WITH RING ABOVE
COMBINING ACUTE ACCENT



0041 030A 0301

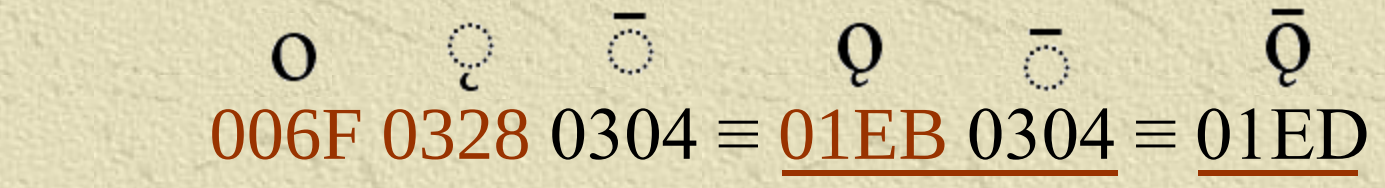
LATIN CAPITAL LETTER A
COMBINING RING ABOVE
COMBINING ACUTE ACCENT

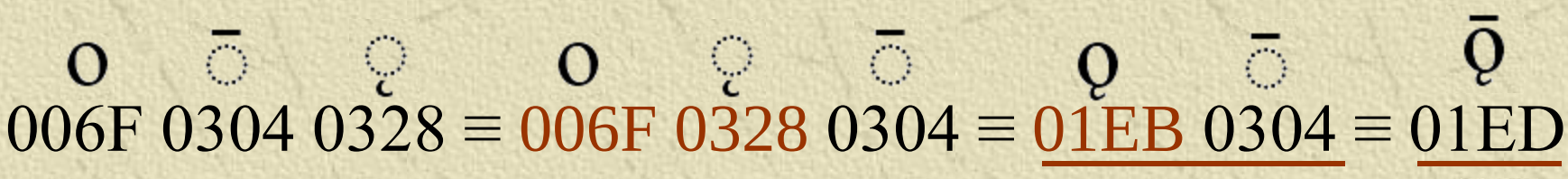
Normalization (NFD)

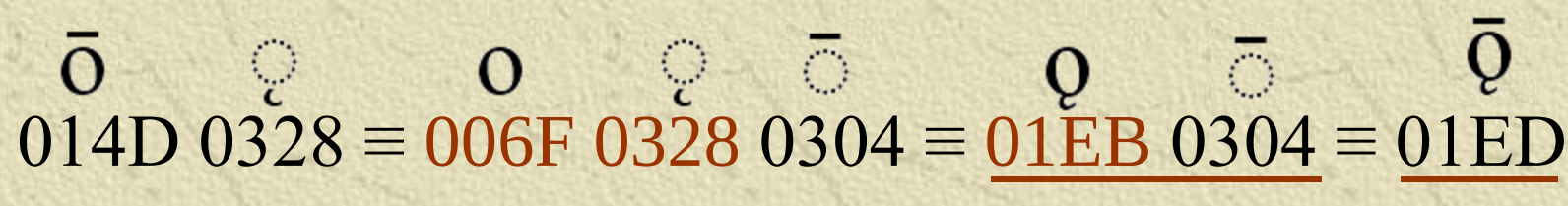
014D;LATIN SMALL LETTER O WITH MACRON;;0;;006F 0304…
01ED;LATIN SMALL LETTER O WITH OGONEK AND MACRON;;0;;01EB 0304…
01EB;LATIN SMALL LETTER O WITH OGONEK;;0;;006F 0328…
0304;COMBINING MACRON;;230…
0328;COMBINING OGONEK;;202…

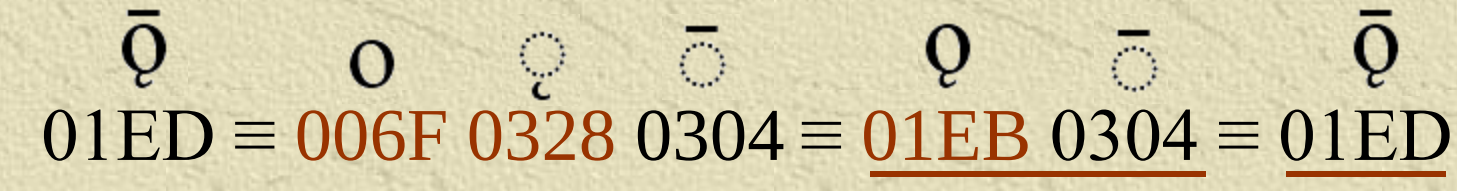
Normalization (NFC)

014D;LATIN SMALL LETTER O WITH MACRON;;0;;006F 0304…
01ED;LATIN SMALL LETTER O WITH OGONEK AND MACRON;;0;;01EB 0304…
01EB;LATIN SMALL LETTER O WITH OGONEK;;0;;006F 0328…
0304;COMBINING MACRON;;230…
0328;COMBINING OGONEK;;202…









Case mapping

- SpecialCasing.txt + UnicodeData.txt
- Unicode digraphs require title casing

DZ 01F1;LATIN CAPITAL LETTER DZ;Lu;;;;;;01F3;01F2

Dz 01F2;LATIN CAPITAL LETTER D WITH SMALL LETTER Z;Lt;;;;;;;;01F1;01F3;

dz 01F3;LATIN SMALL LETTER DZ;Ll;;;;;;;;01F1;;01F2

- Case mapping is not reversible
  McConnel ⇨ mcconnel ⇨ MCCONNEL

Case mapping

- Case mapping may produce strings of different length

  J̌ J ◌̌

  01F0 ⇨ 004A 030C

- Case mapping may depend on the locale

	i	I
English	0069 ⇨	0049
	i	İ
Turkish/Azeri	0069 ⇨	0130

Case mapping

- Case mapping may depend on context

Σ → σ

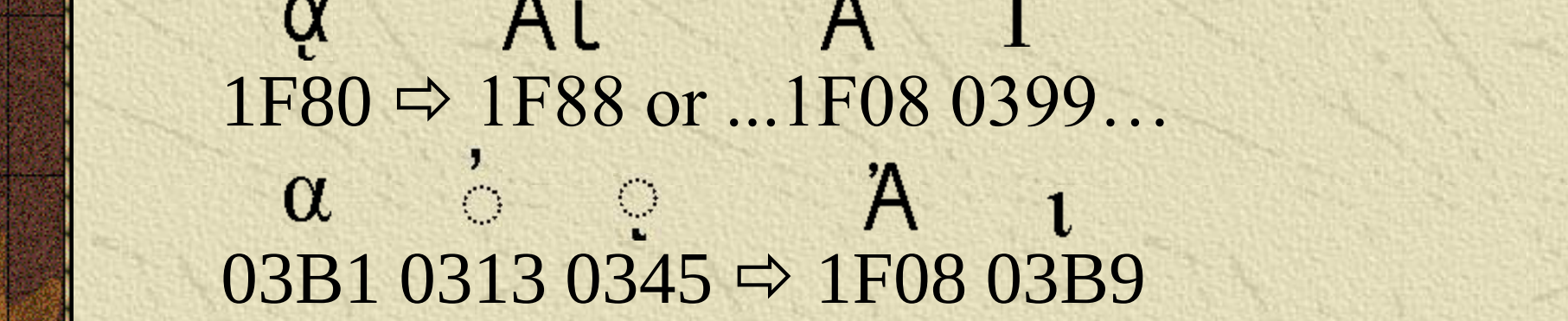
03A3 <letter> ⇨ 03C3

Σ → ς

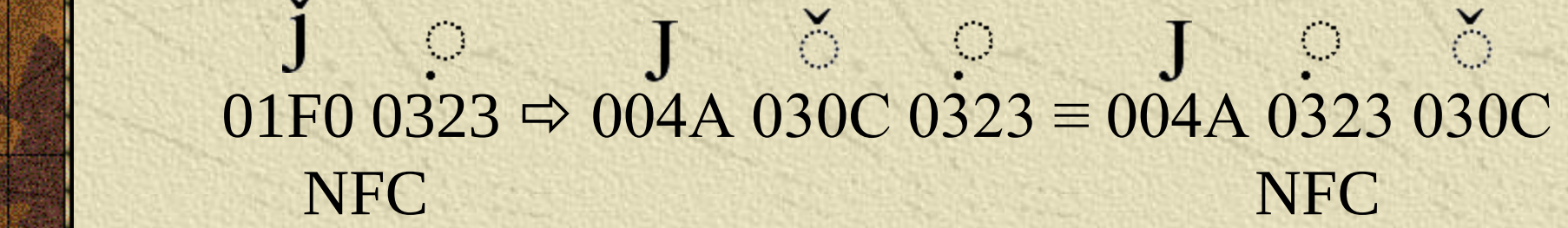
03A3 ⇨ 03C2

Case mapping

- Some characters require special handling



- Case mapping may not preserve normalization



Smart rendering: Arabic

Keyboard:

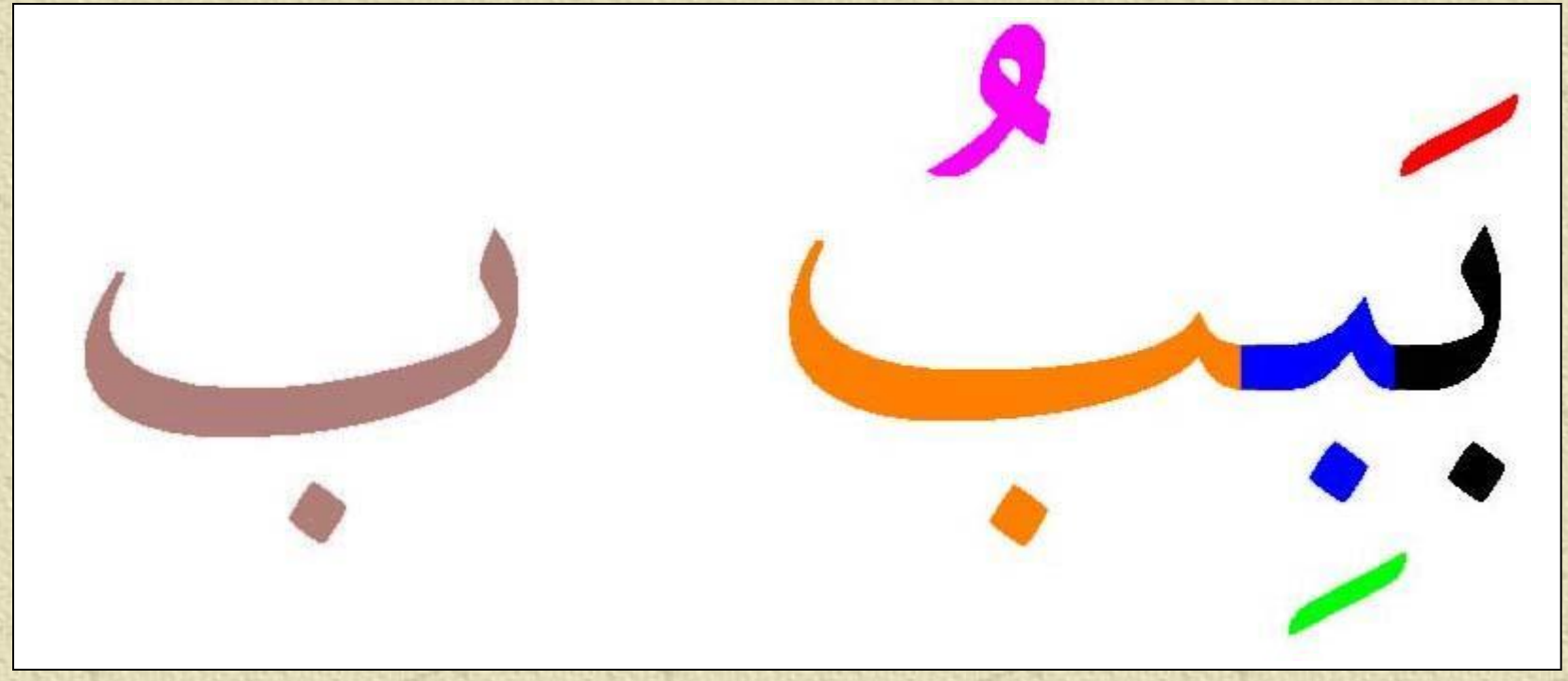
babibu b

Code points:

0628 064e 0628 0650

0628 064f 0020 0628

Screen:



Smart rendering: Burmese

Keyboard:

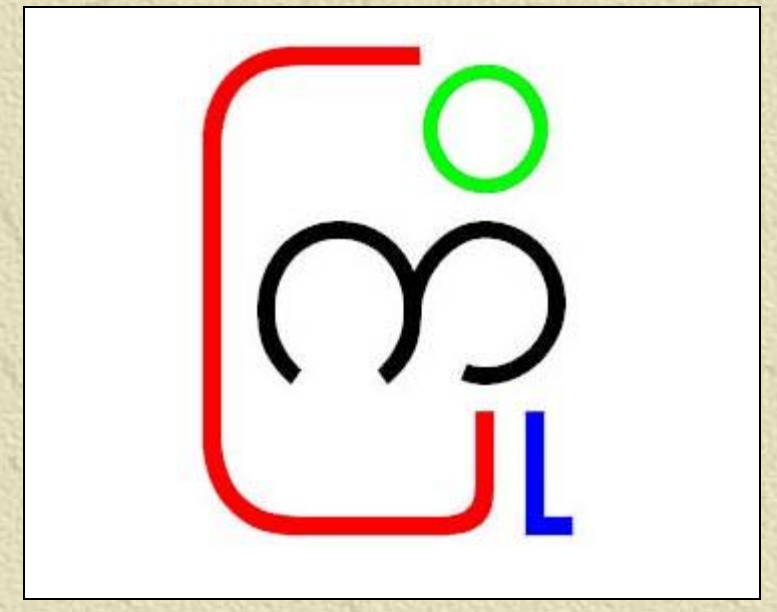
krui

Code points:

1000 1039 101b

102f 102d

Screen:



Smart rendering: Tamil

Keyboard: Ur rU yU NU mU kU jU

Code points: b8a bb0 bb0 bc2 baf bc2 ba3 bc2 bae bc2 b95 bc2 b9c bc2

Screen: