Arabic IDN Variants

Sarmad Hussain

Center for Language Engineering

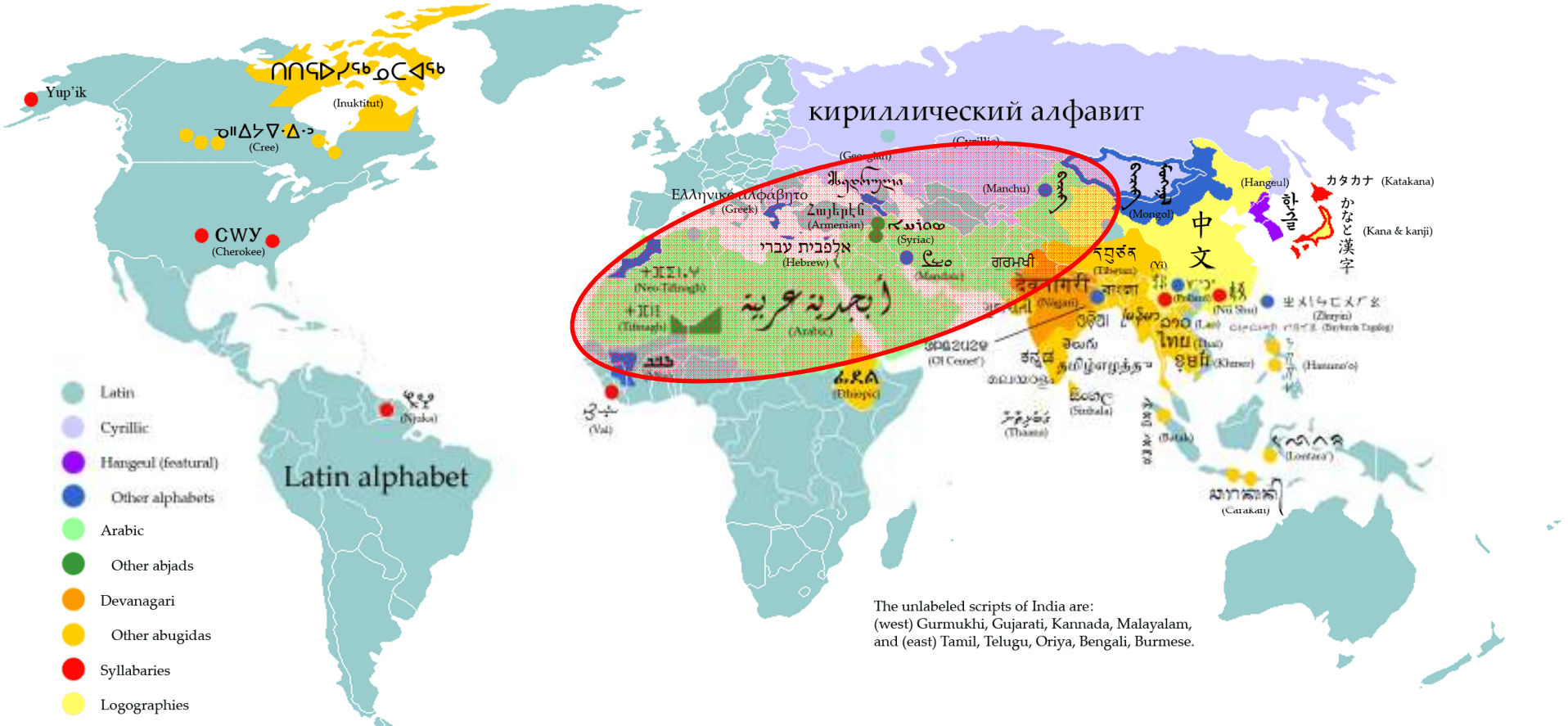
Al-Khawarizmi Institute of Computer Science

University of Engineering and Technology, Lahore

www.cle.org.pk

sarmad@cantab.net

Arabic Script



http://en.wikipedia.org/wiki/File:WritingSystemsoftheWorld4.png

Arabic Script

- Writing system extended to represent multiple languages spoken in:
  - Middle East: *Arabic, Kurdish, Azerbaijani*
  - Africa: *Arabic, Bedawi, Huasa, …*
  - Central Asia: *Kazakh, Uighur, Kirghiz, Azerbaijani*
  - South Asia: *Urdu, Pashto, Balochi, Sindhi, Kashmiri, Torwali, Burshuski, …*
  - South East Asia: *Jawi*

Arabic Script

- Consonantal (abjad)
  - Consonants written explicitly
  - Short vowels represented by optional vowel marks
  - Long vowels are represented by optional short vowel marks plus one of the three consonantal letters: ا و ى
- Bidirectional
  - Letters written from right to left
  - Digits written from left to right

ابجد ١٢١٢٣٤

Arabic Script

- Cursive
  - Characters join with others before and after them
  - Four shapes: initial, medial, final and isolated for most characters

ب ببب چ چچچ س ssss

  - Some characters do no join with characters after them

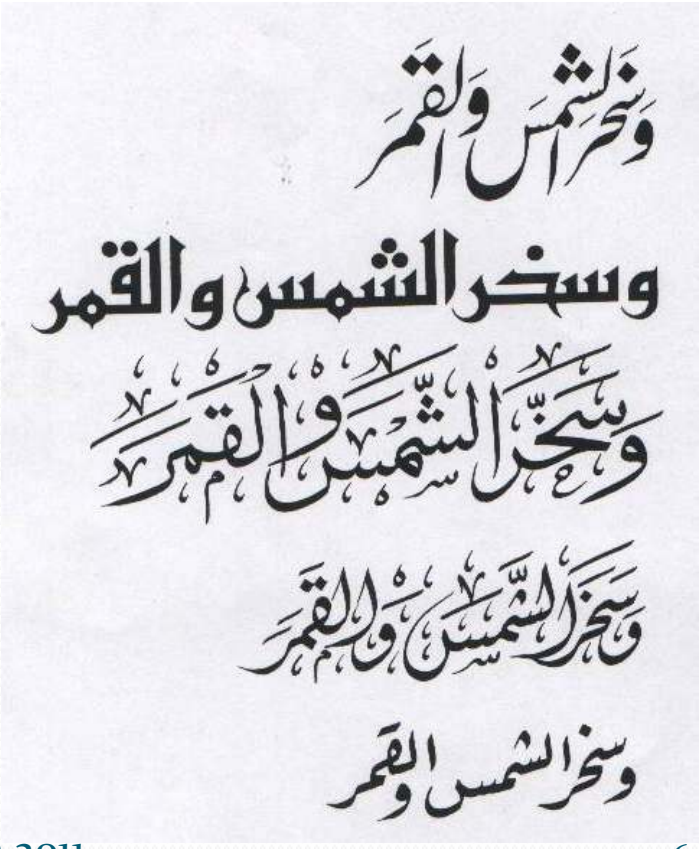
ا ر ڑ ز ژ د ڈ ذ و ...

- Unicode range U+0600-U+06FF and U+0750-U+77F

Arabic Script

- Multiple writing styles
  - Naskh – Arabic, Sindhi, etc.
  - Nastalique – Persian, Urdu, Pashto, etc.
  - Others used frequently, but as stylistic variations
    - Kufi
    - Thuluth
    - Diwani
    - Riqa

وَ سخرَالشَمسَ وَالقمرَ



Sources of Variants

- intrinsically, some strings are considered equivalent by a language or script community

- extrinsically, script encoding scheme (Unicode) is non-optimal, introducing additional ambiguity for end-users

Motivation

- Without Variant Management
  - security threats, allowing for easy phishing
  - perception of a broken internet experience by end-user

Issues

- Technical
  - Required combining marks
  - Optional combining marks
  - Same shape in a particular position
  - Similar shape in a particular position
  - Digits
  - Joining characters – ZWNJ
- User Interface
  - Input method variation
  - Bidirectional rendering issues in applications
  - General rendering issues in applications
- Policy
  - Specification of language table
  - Bundling, blocking, reserving

Required Combining Marks

Combining Mark	Composed Form	Decomposed Form	Unicode Normalized Form
◌ٓ U+0653	آ U+0622	ا◌ٓ U+0627 U+0653	Defined
◌ٹ U+0615	ڑ U+0691	ر◌ٹ U+0631 U+0615	Not Defined
◌ٰ U+065B	U+(ڈ)EE	د◌ٰ U+062F	Not Defined

Can you guess: ڑ ڑ ڑ

Optional Combining Marks

- Vowel Marks
- Honorifics
- Consonantal Gemination
- …

أحد = احد

أُحد = احد

أُحد ≠ أحد

ا◌̃ U+0627 U+0653	ا̃	ا U+0627	Not same
ا◌ُ U+0627 U+64F	اُ	ا U+0627	same

Same Shape

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
ک U+06A9	ک	ک	ک	ک
ك U+0643	ک	ک	ك	ك
ة U+0629	-	-	ة	ة
ۃ U+06C3	-	-	ۃ	ۃ

Can you guess: پاکستان پاکستان

Similar Shape

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
ک U+06A9	ک	ک	ک	ک
ک U+06AA	ک	ک	ک	ک
ت U+062A	ت	ت	ت	ت
ٹ U+067A	ٹ	ٹ	ٹ	ٹ

پاکستان پاکستان پاکستان پاکستان پاکستان پاکستان

Digits

ASCII	Arabic-Indic	Eastern Arabic-Indic
(U+0030) 0	(U+0660)٠	(U+06F0)۰
(U+0031) 1	(U+0661)١	(U+06F1)۱
(U+0032) 2	(U+0662)٢	(U+06F2)۲
(U+0033) 3	(U+0663)٣	(U+06F3)۳
(U+0034) 4	(U+0664)٤	(U+06F4)۴
(U+0035) 5	(U+0665)٥	(U+06F5)۵
(U+0036) 6	(U+0666)٦	(U+06F6)۶
(U+0037) 7	(U+0667)٧	(U+06F7)۷
(U+0038) 8	(U+0668)٨	(U+06F8)۸
(U+0039) 9	(U+0669)٩	(U+06F9)۹

Different or Same?

123abc123
١٢٣abc١٢٣
١٢٣abc١٢٣
١٢٣abc١٢٣
١٢٣abc١٢٣
١٢٣abc123
123abc١٢٣
١٢٣abc123
123abc١٢٣

ZWNJ

- Exact variants ?
  - طب
    - ط ب
  - طب
    - ط ZWNJ ب
- Non-exact variants?
  - خوبصورت
  - خوب‌صورت

User Interface Issues

- Typing a string in
  - application V
  - for language W
  - in country X
  - with keyboard Y
  - in operating system Z

Application Interfaces

Application Interfaces

Policy

- Complete and correct “Language” table
  - Mechanisms to record characters
  - Mechanisms to record variants
  - Mechanisms to record rules
    - ZWNJ
    - Script Mixing
    - Digit Mixing
    - …
  - Mechanisms to record differences in registration vs. resolution (e.g. kafs)
- Mechanisms to record registration differences
  - Variant bundling/blocking/reservation

One World One Internet

Variant handling essential

For secure unified internet experience to global users

using Internationalize Domain Names

Arabic scripts have significant variant issues

which must be addressed for this purpose


مركز تحقيقات لسانيات