Arabic Variant Typology

Sarmad Hussain
Center for Language Engineering
Al-Khawarizmi Institute of Computer Science
Lahore, Pakistan

Variant Definition

- **Arabic Script Character Variant**
  - A Label Valid Character which is replaceable with another Label Valid Character within a label, as defined by a Label Generation Policy.
  - The relationship is generally symmetric in Arabic script.
- **Variant Character Set**
  - The set of code points consisting of a Valid Code Point and all of its variants.

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 tell: 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 پاکسٹان پاکسٹان پاکستان پاکسٹان پاکستان

Possible Types of Variants

- **Same**
  - Identical U+06CC (ی)/U+0649 (ى)
  - In Context U+06A9 (کبک)/U+0643 (كبك)
  - Normalization U+0632 (ز) /U+0631+U+06EC (ر ◌ٜ)
- **Similar**
  - Character U+06AA (ے) / U+06A9 (ک)
  - Diacritic U+062A (ت)/U+067A (ٹ)
- **Different**
  - Shape U+0629 (ۃ)/U+06C3 (ہۃ)
  - Character U+0629 (ة)/U+06C1 (ہ)

Status of Labels

- **Allocation**
  - In a DNS context, the first step on the way to Delegation. A registry (the parent side) is managing a zone. The registry makes an administrative association between a string and some entity that requests the string, making the string a label inside the zone, and a candidate for delegation. Allocation does not affect the DNS itself at all.
- **Delegation**
  - In a DNS context, the act of entering parent-side NS (nameserver) records in a zone, thereby creating a subordinate namespace with its own SOA (start of authority) record. See RFC 1034 for detailed discussion of how the DNS name space is broken up into zones.
- **Activation**
  - The process of making a domain name resolvable.
- **Reservation**
  - In Arabic Script IDN variants context, this is the process of having an unallocated variant label which relates to a Fundamental label that is allocated.
- **Blocking**
  - In Arabic Script IDN variants context, this is the process of having a variant label not allowed for allocation to anyone as long as its Fundamental label is allocated.

Variant Sets and Subsets

- **Variant Label Set**
  - A set of U-labels consisting of one Fundamental Label and its zero or more Variant Labels.
- **Activated Variant Label Subset**
  - The subset of Variant Label Set that is activated, or alternatively, the set containing the Fundamental Label and all its Activated Variants.
- **Allocated Variant Label Subset**
  - The subset of Variant Label Set that is allocated, or alternatively, the set containing the Fundamental Label and all its Allocated Variants.
- **Reserved Variant Label Subset**
  - The subset of Variant Label Set that is reserved, or alternatively, the set containing all the Reserved Variants of a Fundamental Label (and the Fundamental Label, if it is not activated).
- **Blocked Variant Label Subset**
  - The subset of Variant Label Set that is blocked, or alternatively, the set containing all the Blocked Variants of a Fundamental Label

Variant Challenges (and Solutions?)

- Security
  - Variant allocated to a different entity
    - Have comprehensive variant sets
- Usability
  - Label (variant) not resolving for user
    - Have allocatable variants – e.g. KB differences
- Manageability
  - Too many allocated variants to configure
  - Explosion as multiple levels accumulated
    - Have blockable variants – e.g. dot orientation

What is a variant in Arabic?

- **Same**
  - Identical U+06CC (ی)/U+0649 (ى)
  - In Context U+06A9 (كبک)/U+0643 (كبك)
  - Normalization U+0632 (ز) /U+0631+U+06EC (ر ٜ)
- **Similar**
  - Character U+06AA (ے) / U+06A9 (ک)
  - Diacritic U+062A (ت)/U+067A (ٹ)
- **Different**
  - Shape U+0629 (ۃ)/U+06C3 (ۃ)
  - Character U+0629 (ة)/U+06C1 (ہ)

Thank You!