Grapholinguistics in the 21st Century (G21C), Paris, 8th of June, 2022

Graphemic Normalization of the Perso-Arabic Script

Raiomond Doctor, Alexander Gutkin, Cibu Johny, Brian Roark and Richard Sproat

Google Research

Agenda

01 Perso-Arabic: 10K foot view

02 Script Diaspora

03 Normalization Details

04 Basic Experiments

05 Conclusion

Google Research

01

Perso-Arabic: Brief Overview

01 - Perso-Arabic

Popularity and spread

- Perso-Arabic is used by over 600M people (Wikipedia).
- 308 living and historical writing systems (ScriptSource).
- Diverse language families: Afro-Asiatic, Indo-European, Niger-Congo, Turkic, Sino-Tibetan, etc.
- Sample of attested (historical) adaptations:
  - Southern Africa: Afrikaans & Malagasy
  - East Asia: Chinese & Japanese
  - Europe: Bosnian & Aljamiado for Romance languages.
- Modern adaptations: Dardic languages of North Pakistan (e.g., Torwali).

01 - Perso-Arabic

Popularity and spread (cont.)

Multiple reasons for the spread of Perso-Arabic and its popularity:

- Socio-political: conquest
- Economic: trading
- Liturgical: Arabic is the liturgical language of Islam, the Holy Qur'an and the Hadīth being written in Classical Arabic and cultures accepting Islam also accepted the script.

Another very important factor is the *script itself* ...

01 - Perso-Arabic

Origins

Pure consonantal *abjad*

Nabataean Aramaic (Turkmaniyyah, 50 C.E.)

Source: http://www.proel.org/index.php?pagina=alfabetos/nabateo

Pre-Islamic Arabic Jazm (Ḥarrān, Syria, 568 C.E.)

Source: S. D. Abulhab (2007)

... many North Arabian scripts

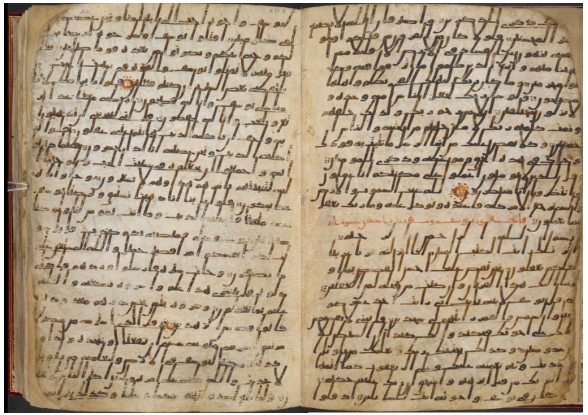
Qur'an in Hijazi style (goat skin, 568-645 C.E)

Source: Birmingham University

Google Research

01 - Perso-Arabic

Core properties



The Ma’il Qur’an (Hijazi, British Library)

The Blue Qur’an (Kufic, Metropolitan Museum)

- Cursivization aided by using vellum, or other soft materials.
- Consonant letters [associated shapes - *rasm* (“drawing”)].

Google Research

01 - Perso-Arabic

Core properties (cont.)

Arabic had more letters than Aramaic. Some letters had to do double duty.

Source: Early Kufic Qur'an (British Library)

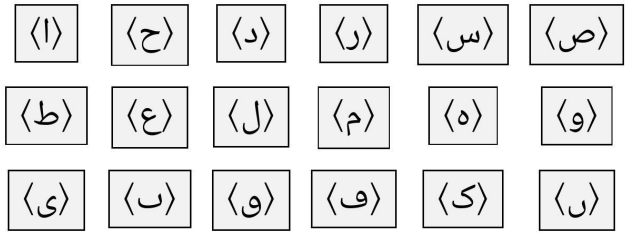
Evolution of diacritics for disambiguation:

- Red dots: *i'jām*
- Green dots: Glottal stop (*hamza*)
- Diagonal strokes: *harakat* (short vowels, etc.)

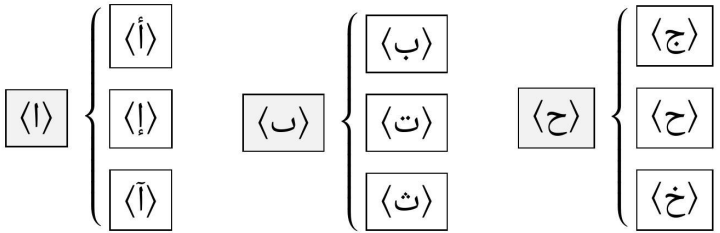
Google Research

01 - Perso-Arabic

Core properties (cont.)



Core 18 *rasm* shapes of Arabic



Example derivations using *i'jām*

The diacritics system evolved into a *productive* way to make new consonant symbols when the script was adapted to new languages.

01 - Perso-Arabic

Adaptations: Examples

- From *abjad* to full alphabet: Sorani Kurdish & Uyghur (Kaye, 1996). Similar to transition from Hebrew to Yiddish (Aronson, 1996).
- Even more *i'jam* dots (up to four).
- Additional diacritics (e.g., new *tashkīl*).
- New *rasm* shapes, e.g. *bari yeh* and *heh do chashmee* to handle aspiration (Urdu).
- ... and much more!

01 - Perso-Arabic

Digital Medium & Ambiguities

- Similar to Brahmic scripts, the script allows for more than one way to compose a character on the digital medium. Unicode:
  - *alef with madda above* = U+0622, or
  - *alef with madda above* = *alef* (U+0627) + *madda above* (U+0653)
- This results in *presentation ambiguity* and a special *normalization process* has been set in place by the Unicode to handle ***some*** simple cases: Normalization Form C (NFC).

01 - Perso-Arabic

Arabic Code Pages in Unicode

- 1991: Original appearance:
  - 169 atomic characters
- Unicode 14.0 (2021):
  - 10 code pages: letters, diacritics, ligatures, punctuation, cardinals, historic, regional
  - over 440 atomic characters of interest to us
  - ... and growing! e.g., ARABIC LIGATURE RAHIMAHUM ALLAAH [U+FD4F]
- Arabic case studies by teams formed by Internet Corporation for Assigned Names and Numbers (ICANN): Domain names, etc.

02

Script Diaspora

Sample of six languages

02 - Script Diaspora

Urdu

- Indo-Aryan language. ~230M L1&L2 speakers (Ethnologue).
- Styles: Nastaliq & Naskh. Nastaliq (preferred) evolved from Naskh in Persia (13th century C.E.). Naskh preferred on digital media & (older) hard types.
- Urdu *abjad* evolved from Persian:
  - From Arabic (28 letters)
  - 32 letters: (+ ⟨پ⟩ ,⟨ژ⟩ ,⟨گ⟩ ,⟨چ⟩), *yeh* (ي) → *farsi yeh* (ی), *kaf* (ك) → *keheh* (ک).

02 - Script Diaspora

Urdu (cont.)

The original 32-letter Persian inventory augmented as follows (simplified):

- Retroflexes: *rreh* (ڑ), *dal* (ڈ), *tteh* (ٹ). Derivation: *tteh* /ʈ/ = *teh* (ت) + *tah* (ط).
- Nasalization: *noon ghunna* (ں)
- Aspiration on preceding consonant: *two-eyed heh* or *heh doachashmee* (ھ).
- Word final /eː/ or /ɛː/: *yeh barree* or *greater yeh* (ے). From final form of (*farsi*) *yeh?*
- Voiced glottal fricative /ɦ/: *round he* or *gol he* or *heh goal* (ہ).
- ... and much more! (e.g., more diacritics for vowels, teh marbuta (ۃ) for grammatical marking (feminine gender) on nouns & adjectives)

Derivations highly influenced by the Nastaliq style.

Officially – 56 letters, but this is debated.

Google Research

02 - Script Diaspora

Punjabi (Shahmukhi)

Similar to Hindi/Urdu, script divide: Western vs. Eastern Punjabi (Gurkmukhi)

Shahmukhi mostly shares letter inventory with Urdu, + ...

- Voiced retroflex lateral approximant /ɭ/: *lam with small tah above* (ﻟٜ),
- Voiced retroflex nasal /ɳ/: *noon with small tah* (ںٜ).

Overall, Shahmukhi & Urdu are mutually intelligible as scripts.

Similar to Urdu, the precise number of letters is a matter of debate:

- Some scholars mention the following four letters: *gueh* (ﮒ), *dal with two dots vertically below and small tah* (ڈٟ), *dyeh* (ﭽ), *beeh* (ﭖ).
- ... but we don't see them in our mined data. Noise from Sindhi & Dardic languages?

Google Research

02 - Script Diaspora

Sindhi

Another Indo-Aryan language recorded in Perso-Arabic & Devanagari.

- Unlike Urdu and Shahmukhi, does *not* use Nastaliq, Naskh only.
- First standard inventory of 52 letters goes back to Colonial Era (1853).
- Modern inventory has 64 (!) letters. *Some* unique features:
- Unlike Urdu & Shahmukhi, high *hamza* is part of atomic letters *yeh with hamza above* (ئ) and *waw with hamza above* (ؤ), somewhat similar to Pashto.
- Four implosives: */ɠ/ gueh* (ڳ), */ʃ/ dyeh* (ڃ), */ɗ/ dal with three dots above downwards* (ڈ), */ɓ/ beeh* (ٻ).
- Unlike Urdu, only uses *heh doachashmee* (ھ) to mark three aspirates, rest use standalone letters.
- Vowel diacritics not normally used, only to mark short vowels.

Google Research

02 - Script Diaspora

Kashmiri

Indo-Aryan language from Dardic group. Also recorded in Devanagari:

- The only Dardic language with surviving literary tradition.
- Nastaliq & Naskh styles. Nastaliq traditionally, but Naskh preferred digitally.
- Closer to alphabets, e.g., Sorani & Uyghur (vowels are regularly marked).
- *Some* unique features:
  - All vowels clearly represented in 8 pairs (short/long).
  - Palatalization: Letter *yeh with a ring below* (□),
  - “Long” schwa /ə:/: alef + *wavy hamza above*,
  - Long close/high central unrounded /ɨ:/: alef + *wavy hamza below*,
  - Use of *sukun* (or *jazm*) for marking consonant clusters not unique, but the shape is: inverted ⟨v⟩.

Google Research

02 - Script Diaspora

Malay (Jawi)

Jawi is used to record Malay language (& some others) from Austronesian family.

- Prefers Naskh style, Nastaliq not used.
- Comparatively small inventory (37 letters). *Abjad*, vowels usually unmarked.
- 28 letters from Arabic + some Persian, e.g. *tcheh* (چ).
- Some unique features:
  - Script-level marking of morphological process of full reduplication:
    - "Dogs": Arabic numeral ⟨٢⟩ ("2") as in انجيغ٢ "*anjeng*", contrasted with:
    - [Rumi] "Persons": "*orang-orang*".
  - Special *waw with dot above* (ۏ) for representing foreign loanwords (/v/ or /f/).

02 - Script Diaspora

Uyghur

Language from a Turkic family, Perso-Arabic officially reinstated in 1982.

- Proper alphabet, similar to Sorani Kurdish: vowels are explicitly marked.
- “Usual” 28 Arabic letters + 4 Persian letters, e.g., *gaf* (گ) for /g/.
- Some unique features:
  - Velar nasal /ŋ/ letter *ng* (ڭ) = (Arabic) *kaf* (ك) + three dots above,
  - Rich vocalic system, Turkic-specific, e.g., *waw*-based derivations:
    - Letter *yu* (ۇ) for /ü/ = *waw* + superscript *alef*,
    - Letter *ve* (ۋ) for semivowel /w/ = *waw* + three dots on top,
    - Letter *oe* (ۆ) for front rounded vowel /ø/ = *waw* + small v on top.

03

Normalization Details

03 - Normalization Details

Script Normalization

Normalizing the text on the Web

[ALEF WITH MADDA ABOVE, HAMZA BELOW]

OR

[ALEF WITH HAMZA BELOW, MADDAH ABOVE]

IN

إئيكن ?

Google Research

03 - Normalization Details

Visual Normalization

The previous example demonstrates *visual invariance*. We identify several cases:

1. Canonical operations supported by Unicode (NFC/NFKC), e.g.:
   - [LETTER ALEF, MADDAH ABOVE] ⟶ LETTER ALEF WITH MADDA ABOVE
   - Combining marks: [SUKUN, SHADDA] ⟶ [SHADDA, SUKUN]
   - Presentation forms: ALEF WASLA ISOLATED FORM ⟶ ALEF WASLA
2. Extra *language-agnostic* rewrites, e.g.:
   - [LETTER WAW, DAMMA] (ۇ) ⟶ LETTER U (ۇ)
3. Language-specific rewrites, e.g. (Kashmiri):
   - [LETTER DAL, ROUNDED HIGH STOP WITH FILLED CENTRE] ⟶ LETTER THAL

We implement (1) + (2) + (3) in *visual normalization grammars*.

Google Research

03 - Normalization Details

Visual Normalization (cont.)

Visual normalization internal subgrammars as a sequence of compositions:

1. Rewrite of presentation forms.
2. NFC rewrites,
3. Position-independent rewrites,
4. Word non-final rewrites (initial and medial positions),
5. Word-final rewrites,
6. Isolated letter rewrites.

آزادئمذہب → (1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6) → آزادئمذہب

Google Research

03 - Normalization Details

Reading Normalization

Language-specific transformations that produce *different* output.

Source of errors:

- Imperfect input methods
- Inadequate script literacy
- Older Unicode versions

Examples:

- Sorani Kurdish, Punjabi: YEH ي → FARSI YEH ى
- Kashmiri: YEH WITH TAIL ى → KASHMIRI YEH
- Sindhi: ALEF MAKSURA ى → [YEH, SUPERSCRIPT ALEF] ٻ
- Pashto, Persian: KAF ك → KEHEH ک

Google Research

03 - Normalization Details

Reading Normalization (cont).

Reading normalization pipeline:

1. Visual normalization
    - ... includes the six subgrammars defined above
2. Reading normalization proper.

Urdu: [ALEF WITH] WAVY HAMZA [ABOVE] ⟶
[ALEF WITH] HAMZA [ABOVE]: حيثيأت ⟶ (1 ○ 2) ⟶ حيثيأت

No positional transformations yet.

Google Research

03 - Normalization Details

Nisaba Library

https://github.com/google-research/nisaba

- Based on *Pynini*: A Python library for weighted finite-state grammar compilation.
- Perso-Arabic Layer: Finite-state grammars for
  - (Reversible) Romanization (language-agnostic),
  - NFC Normalization (language-agnostic),
  - Visual (NFC++) Normalization (language-specific),
  - “Reading” Normalization (language-specific).
  - “Letter Typology”: Inverse index of letter → languages

03 - Normalization Details

Nisaba Library (current state)

- 8 languages using Perso-Arabic are supported
- 179 atomic letters documented for 59 languages based on ICANN recommendations, e.g.
    - SEEN WITH SMALL ARABIC LETTER TAH AND TWO DOTS [U+0770] (ݰ) → Khowar (کھووار) from Dardic group
    - BEEH [U+067B] (ﭒ) → { Saraiki, Sindhi }
        - ... as opposed to “common” BEH [U+0628] (ب) → {Arabic, Persian, ...}
- 697+ rewrites of Arabic Presentation Forms (initial/medial/final/isolated) from Unicode *Normalization Form Compatibility Composition* (NFKC)

Google Research

03 - Normalization Details

Why real data is complex

Observations from working with the data in the wild (e.g., CommonCrawl):

- Deprecated letters, e.g. Kashmiri:
  - ALEF WITH WAVY HAMZA BELOW ⟶ [ALEF, WAVY HAMZA BELOW]
- Use of isolated (presentation) forms within words
- Use of Zero-Width Non-Joiners (ZWNJ) to guide the display
- Code-switching: Hard to apply language-specific token normalization without good short-segment LangID.
- Confusables with digits, e.g.:
  - ALEF WITH MADDA ABOVE “آ” vs. ["ARABIC-INDIC DIGIT ONE", "MADDAH ABOVE"] "١"

Google Research

04

Experiments

04 - Experiments

Motivation

- How “useful” is the resolution of representation ambiguities for different languages using Perso-Arabic script?
- How to quantify the “usefulness”?
- If the phenomena being normalized are rare, then the difference between the conditions will be small; and if the normalizations do not result in better text representations, then the normalized conditions may exhibit a lower quality in the validation.

04 - Experiments

Design

Simple procedure:

1. Given a text corpus $\mathfrak{T}$, run reading normalization on it obtaining $\mathfrak{T}^R$.
2. Repeat *k* times (*k*=100) for $\mathfrak{T}$ and $\mathfrak{T}^R$:
   2.1. Randomly shuffle the corpus.
   2.2. Split into 80% training and 20% test partitions. Make sure that the normalized sentences are confined to the training set.
   2.3. Train statistical model and evaluate it on the test partition obtaining metrics of interest $\mathfrak{Q}_k$ (or $\mathfrak{Q}_k^R$).
3. Statistically compare (using significance testing) the two samples $\mathfrak{Q}$ and $\mathfrak{Q}^R$ of accumulated results.

Google Research

04 - Experiments

Data

- We use available up-to-date Wikipedia dumps (all but Jawi).
- Wikipedia data is good for our purposes - good balance between reasonably clean data and comprehensive sample of written language.
- ... but it still requires significant cleanup:
  - markup removal
  - text in other scripts
  - noise & spam, etc.
- For Malay (https://arxiv.org/abs/2205.03983)

Google Research

04 - Experiments

Methods

Models:

- Character- and word-level statistical *n*-gram models (interpolated).
- Model orders: characters: n ∈ [3, 10]; words: n ∈ [2, 5]

Evaluation:

- Compute cross-entropy (bits per character or word): how well the model describes the test set (lower - better).
- Obtain sets of cross-entropies for normalized/unnormalized, each model order (*n*) and 100 folds. Then analyze the differences between the sets.

04 - Experiments

Results

Statistical significance testing:

- Welch-Satterthwaite *t*-test (WS)
- [non-parametric]: Mann-Whitney (MW)
- [non-parametric]: Brunner-Munzel (BM)

The tests provide the t statistic, the p-value and the estimated confidence interval (CI) $[L, H]$ for the 95% confidence level at the significance level of $\alpha = 0.05$.

04 - Experiments

Summary & Discussion

Sample of three groups of observed outcomes:

1. **Jawi:**
   - Normalization is *very effective* (up to 3% improvement).
2. **Sorani Kurdish, Shahmukhi, Sindhi, South Azerbaijani, Urdu & Uyghur:**
   - Small but *statistically significant* improvements across the board.
3. **Kashmiri:**
   - Very small dataset. But character models mostly do well.
   - Unable to train on words reliably. *Some* model configurations show improvements, but these are *not significant*.

05

Conclusion & Future Work

05 - Conclusions & Future Work

To summarize:

- Partially touched on some complexities of Perso-Arabic script universe.
- Essentially, each language has its own script.
- Low-level language-specific script normalization rules are effective in reducing the representation ambiguities.
- Downstream Natural Language Processing (NLP) applications will benefit from these techniques.

Future work:

- Expanding coverage for existing languages.
- New languages, especially lower-resource ones, e.g. Balochi.
- Experiments: Neural Machine Translation, Language Identification, Sentiment Analysis, etc.

Google Research

Thank You

Alexander Gutkin

Please see our upcoming paper for more details.

Google Research