The Developing Orthographic Conventions of Roman Hindi-Urdu

Saiya Karamali
Outline

1. Some background on Hindi and Urdu and why I’m treating them together
2. Background: How technology has impacted written language and what this tells us about written language broadly
3. A (brief) linguistic framework for analyzing written language
4. What I’ve found about the orthographic conventions of Hindi-Urdu
5. What these findings tell us about how a writing system might develop organically
Hindi and Urdu

- Indo-Aryan languages spoken in India and Pakistan
- Sociocultural divisions: Hindus overwhelmingly identify as Hindi speakers and Muslims normally identify as Urdu speakers
- Closely related; mainly differentiated by script
  - Devanagari (Hindi) vs. Perso-Arabic (Urdu)
- Official languages in India and Pakistan

A multilingual street sign in India
A brief history of script choice in South Asia

• Both scripts have been in use in South Asia since the eleventh century, when Perso-Arabic was introduced
• Roman script was introduced by the British in the 19th century, but was largely rejected by the native population (Ahmad, 2011)
• 1800s, British India: Religious divisions grow and Hindi/Urdu script choice becomes a controversial issue
  • The British tried to introduce the Roman script, but it was largely rejected by the native population (Ahmad, 2011)
• 1900s: India/Pakistan become independent and adopt separate languages/scripts
• 1990s: English comes to be associated with economic prosperity and becomes the “language of the youth” (Nema & Chawla, 2018).
Early examples of Roman Hindi-Urdu

• Early use in the 1800s by European missionaries and British Indians

• Early pedagogical materials in Roman Hindi-Urdu: Rahman, 1923; Sharma, 1937

Gilchrist (1803): Fable I, The Trees and the Bramble

Kipling (1890): Gunga Din

Of all them blackfaced crew
The finest man I knew
Was our regimental bhisti, Gunga Din,
‘He was ’Din! Din! Din!
‘You limpin’ lump o’ brick-dust, Gunga Din!
‘Hi! Slippy hitherao
‘Water, get it! Panee lao,
‘You squidgy-nosed old idol, Gunga Din.’
• Technology has led several languages to adopt the Roman (Latin) script

• Some languages have done so reluctantly (i.e. Greek; Mouresioti & Terkourafi, 2021), and have taken advantage of increased tools for typing in traditional scripts

• Hindi-Urdu, by contrast, seems to have embraced the Roman script
  • Bali et al., 2014: 84% of Hindi Facebook posts were written in the Roman script
Why is Roman Hindi-Urdu so popular?

• Some possibilities:
  • Increased prestige of English in South Asia
  • The use of English as a *lingua franca*
  • “Hybrid Identity” of South Asians as a result of colonization and Western influence (Atta, 2021)
  • New ways of expressing linguistic identity outside of script choice?
When will he play?

Seriously, we’ve been wondering the same here in Lucknow.

Watching a movie takes time too. The kids need to go to school. Watch it quickly so neither your nor the nation’s time is wasted.
Roman Hindi-Urdu today: linguistic landscapes

Advertisement in Rawalpindi, Pakistan
Photo: Atta, 2021
Roman Hindi-Urdu today: linguistic landscapes

Photo from Lucknow, India, 2014
Terminology

• I follow Meletis & Dürscheid (2022):
  • *Grapheme*: the smallest unit of a writing system.
    • For alphabets, equivalent to a letter
  • A *script* is the set of graphemes used for a language
  • *Orthography*: The prescriptive or descriptive rules which govern how graphemes combine to form words
  • *Writing system*: a combination of script and orthography
  • `<a>`: the grapheme “a”
How does a community select a writing system?

- Meletis (2018): four major factors determine which writing system a language will adopt:
  - **Linguistic fit**: Does each sound have a unique orthographic representation? Does each grapheme represent a single sound?
  - **Psychological/Cognitive fit**: How easy is the writing system for readers to process?
  - **Sociocultural fit**: How well does a writing system match users’ identities? Do they wish to associate themselves with or distance themselves from users of particular scripts?
  - **Technological fit**: How easily is the writing system used on computers and mobile devices?
Some factors which could affect orthographic conventions

- Avoiding ambiguity (*linguistic fit*)
- Similarity to English orthography to increase ease of learning (*psychological fit*)
- Similarity to Hindi and Urdu orthographies (*psychological/sociocultural fit*)
- Avoiding diacritics and complex letter combinations (*technological/psychological fit*)
- Expressing identity as a Hindi/Urdu speaker (*sociocultural fit*)
- Reflecting phonological variation (*sociocultural fit*)
Research questions

• How is each phoneme represented in Roman Hindi-Urdu?
• How do linguistic fit, psychological/cognitive fit, sociocultural fit, and technological fit seem to shape these orthographic conventions?
• What does this data tell us about Hindi-Urdu speakers perceptions of sounds?
• Does the data reflect phonological variation?
Methods

• Data from X collected between May 1 and May 9, 2023:
  • Selected ASCII tweets that were automatically classified as Urdu or Hindi
  • Eliminated duplicates resulting from retweets/quote tweets
  • Resulted in 8909 usable tweets

• Composed a dataset of each word in the data
  • Removed proper names, obvious English loanwords, non-Hindi-Urdu words, web addresses, and X usernames
  • Ignored case

• For most of the analysis, used the top 2000 most frequent words only
  • At least six occurrences
Example tweets from my data

*Iqrar ul Hassan Syed* @Iqrarulhassan · May 4, 2023

When will he play?

Seriously, we’ve been wondering the same here in Lucknow

*Qaiser Abbas* @QaiserAbbas1979

Watching a movie takes time too. The kids need to go to school. Watch it quickly so neither your nor the nation’s time is wasted.

*Old Monk* @_Nightowl___

When will he play?

“When will he play?”

“When will he play?”

“When will he play?”

“When will he play?”
Analysis

• Case 1 (stops): Reducing ambiguity
• Case 2 (velar fricatives): Is dialectal variation reflected in the orthography?
• Case 3 (/v/): What happens when there are two equally-plausible Roman-script equivalents?
• Case 4 (vowel tenseness/length): What can we learn about speakers’ auditory perception of phonemes?
Stops

• Four-way stop contrast, plus phonemic geminate consonants
• Stop/affricate equivalents to all of the English places of articulation, plus retroflex stops
• How do we represent all of these using the Roman script!? 

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stops</td>
<td>p pʰ b bʰ</td>
<td>t tʰ d dʰ</td>
<td>t tʰ d dʰ</td>
<td>k kʰ g gʰ</td>
<td>qʰ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Stops: observations

• When straightforward Roman script analogues are available, they are typically used
  • /p t k tʃ b dʒ/ \(\rightarrow\) <p t k ch b d g j>

• Retroflex and dental stops both represented with <t d>
  • But the rarity of retroflex stops probably makes these easier to process

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>/ʈ/</th>
<th>/ɖ/</th>
<th>/ʈ̪/</th>
<th>/ɖ̪/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrences</td>
<td>1,052</td>
<td>773</td>
<td>364</td>
<td>155</td>
</tr>
</tbody>
</table>

Occurrences of dental and retroflex stops, including rejected English loanwords
Aspirate/Geminate stops

• /h/ as an aspiration/breathiness marker
  • <th ph kh bh dh gh>

• Consonant repeated to indicate gemination
  • <pp tt kk bb dd gg>
Velar fricatives

• /x/ and /ɣ/ are generally represented with <kh> and <gh>
  • Overlap with /kʰ/ and /gʲ/.
  • Likely from early literature; used diacritics to distinguish

• Variation in representing /x/ and /ɣ/ may reflect phonological variation
  • /x ~ kʰ/ but /ɣ ~ g/

<table>
<thead>
<tr>
<th></th>
<th>/ɣ/</th>
<th>/g/</th>
<th>/x/</th>
<th>/k/</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;g&gt;</td>
<td>100</td>
<td>3326</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&lt;k&gt;</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15558</td>
</tr>
</tbody>
</table>

Frequency counts by phoneme represented by <g> and <k>
/v/: Examining linguistic fit

- Allophonic variation between [v] and [w]
- In perceptual studies, native Hindi speakers could not reliably distinguish [v] and [w] (Grover, 2016).
- Preference for <w> therefore seems arbitrary
- Evidence for tendency towards higher linguistic fit

<table>
<thead>
<tr>
<th>Expected surface form</th>
<th>&lt;v&gt;</th>
<th>&lt;w&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>[v]</td>
<td>276</td>
<td>1217</td>
</tr>
<tr>
<td>Free variation</td>
<td>46</td>
<td>107</td>
</tr>
</tbody>
</table>

Orthographic representations of /v/ by phonological enviroment
Vowels tenseness/length: examining auditory perception

• All Hindi-Urdu vowels are common across English dialects
  • Length or tenseness contrast for high- and mid-vowels
• Lots of variation in literature (repeated representations in red)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>e</th>
<th>ê</th>
<th>i</th>
<th>i</th>
<th>u</th>
<th>o</th>
<th>ô</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gilchrist (1803)</td>
<td>a&lt;a&gt;</td>
<td>e&lt;e&gt;</td>
<td>e&lt;e&gt;</td>
<td>i&lt;i&gt;</td>
<td>i&lt;i&gt;</td>
<td>u&lt;oo&gt;</td>
<td>o&lt;oo&gt;</td>
<td>o&lt;αι&gt;</td>
<td>e&lt;e&gt;</td>
</tr>
<tr>
<td>Rahman (1923)</td>
<td>a&lt;ā&gt;</td>
<td>e&lt;ē&gt;</td>
<td>e&lt;e&gt;</td>
<td>i&lt;i&gt;</td>
<td>i&lt;i&gt;</td>
<td>u&lt;ū&gt;</td>
<td>o&lt;o&gt;</td>
<td>o&lt;au&gt;</td>
<td>e&lt;e&gt;</td>
</tr>
<tr>
<td>Sharma (1937)</td>
<td>a&lt;ā&gt;</td>
<td>e&lt;e&gt;/&lt;ai&gt;</td>
<td>e&lt;e&gt;</td>
<td>i&lt;i&gt;</td>
<td>i&lt;i&gt;</td>
<td>u&lt;ū&gt;</td>
<td>o&lt;o&gt;</td>
<td>o&lt;au&gt;</td>
<td>e&lt;e&gt;</td>
</tr>
<tr>
<td>Khan (2000)</td>
<td>a&lt;a&gt;</td>
<td>e&lt;e&gt;</td>
<td>E&lt;E&gt;</td>
<td>i&lt;i&gt;</td>
<td>i&lt;i&gt;</td>
<td>u&lt;U&gt;</td>
<td>o&lt;o&gt;</td>
<td>o&lt;O&gt;</td>
<td>e&lt;e&gt;</td>
</tr>
</tbody>
</table>

• So, do users prioritize linguistic fit or convenience/ease of use?
• A system like Gilchrist’s (1803) might maximize both

An eleven vowel phonemic inventory of Hindi-Urdu (Ohala, 1994)
## Analysis: Vowels

### Orthographic representations of front vowels

<table>
<thead>
<tr>
<th></th>
<th>&lt;i&gt;</th>
<th>&lt;ai&gt;</th>
<th>&lt;e&gt;</th>
<th>&lt;ee&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ɪ/</td>
<td>1246</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>/i/</td>
<td>2244</td>
<td>9</td>
<td>44</td>
<td>342</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>0</td>
<td>806</td>
<td>163</td>
<td>0</td>
</tr>
<tr>
<td>/e/</td>
<td>14</td>
<td>1320</td>
<td>13570</td>
<td>6</td>
</tr>
</tbody>
</table>

### Orthographic representations of back high- and mid- vowels

<table>
<thead>
<tr>
<th></th>
<th>&lt;o&gt;</th>
<th>&lt;u&gt;</th>
<th>&lt;au&gt;</th>
<th>&lt;oo&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ɑ/</td>
<td>0</td>
<td>3026</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>/u/</td>
<td>52</td>
<td>1368</td>
<td>0</td>
<td>225</td>
</tr>
<tr>
<td>/o/</td>
<td>6560</td>
<td>0</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>/ɔ/</td>
<td>239</td>
<td>0</td>
<td>927</td>
<td>0</td>
</tr>
</tbody>
</table>
Analysis: Vowels

- `<u>` generally represents both /ʊ/ and /u/; `<i>` generally represents both /i/ and /ɪ/
  - Thus, both linguistic fit and possible ease of learning are rejected in favor of increased ambiguity
- But `<au>` generally represents /ɔ/ and `<ai>` generally represents /ɛ/.

Possible conclusions:
- Users inherit a comfort with vowel ambiguity from English and/or Perso-Arabic
- The tenseness/length contrast is more easily perceived for mid-vowels than high vowels
Discussion

• Users do maximize linguistic fit, but with the following considerations:
  • Ambiguity is more tolerable for rare phonemes (i.e. retroflex consonants)
  • Perception and dialect variation are often reflected (i.e. /gɦ/)
  • Longer orthographic representations for a single phoneme are generally not favored (i.e. /ʧʧ/)
  • Variation is more likely for phonemes with no clear English equivalent (i.e. retroflex stops)
  • Conventions from English and from early Roman Hindi-Urdu literature may also be carried over (i.e. /x γ/)
Possible next steps

• Sociocultural variation – does Roman Hindi-Urdu retain ways of distinguishing self-identified Hindi and Urdu speakers?
• Do users have perceptions about ‘proper’ ways to write Hindi-Urdu?
• Is Roman Hindi-Urdu popular among users?
• How does processing time for the Roman script compare with the Perso-Arabic and Devanagari scripts?
Conclusion

• Meletis’ (2018) framework reasonably explains what we see for Roman Hindi-Urdu

• Linguistic fit, familiar orthographic conventions, and users’ perception of phonemes all help to shape the orthographic conventions of Roman Hindi-Urdu

• Other organically developing writing systems may be shaped by similar factors
References