The Developing Orthographic Conventions of Roman Hindi-Urdu

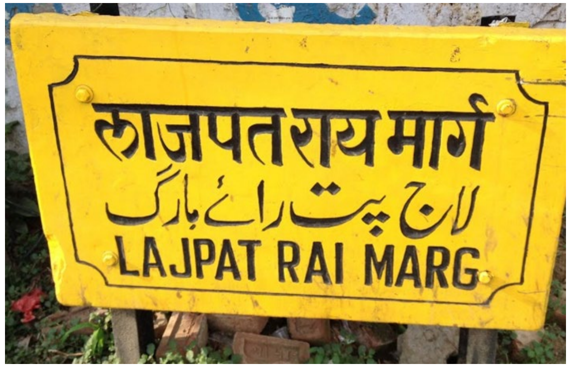
Saiya Karamali

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



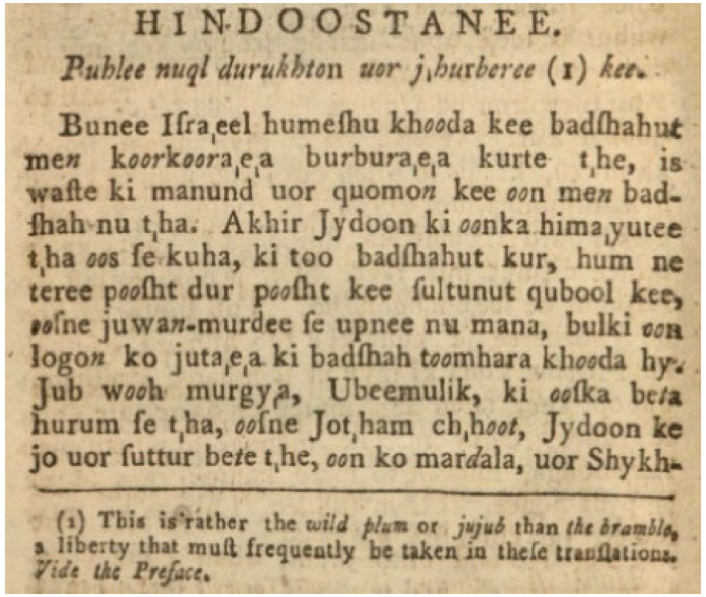
A multilingual street sign in India

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



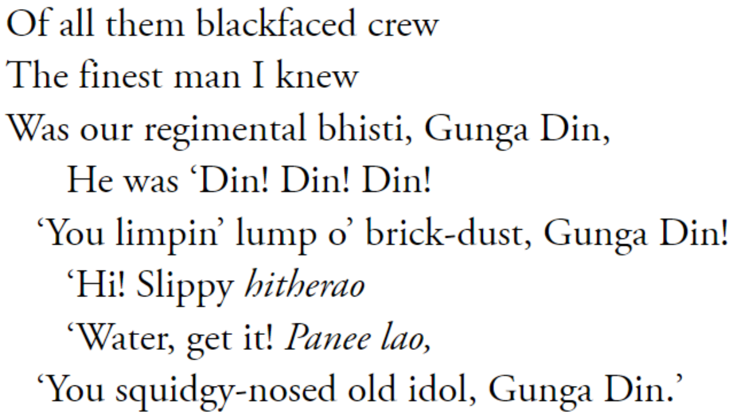
HINDOOSTANEE.

Publee nuql durukhton uor jˌhurberee (1) *kee.*

Bunee Iſraˌeel humeſhu kh*oo*da kee badſhahut me*n* k*oo*rk*oo*raˌeˌa bur*b*uraˌeˌa kurte tˌhe, is waſte ki manund uor quomo*n* kee *oo*n me*n* badſhah nu tˌha. Akhir Jydoon ki *oo*nka himaˌyutee tˌha *oo*s ſe kuha, ki too badſhahut kur, hum ne teree p*oo*ſht dur p*oo*ſht kee ſultunut qubool kee, *oo*ſne juwa*n*-murdee ſe upnee nu mana, bulki *oo*n logo*n* ko jutaˌeˌa ki badſhah t*oo*mhara kh*oo*da hy. Jub w*oo*h murgyˌa, Ubeemulik, ki *oo*ſka be*t*a hurum ſe tˌha, *oo*ſne JotˌHam chˌh*oo*t, Jydoon ke jo uor ſuttur be*t*e tˌhe, *oo*n ko mar*d*ala, uor Shykh-

(1) This is rather the *wild plum* or *jujub* than *the bramble*, a liberty that muſt frequently be taken in theſe tranſlations. *Vide the Preface.*

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



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.'

Kipling (1890): *Gunga Din*

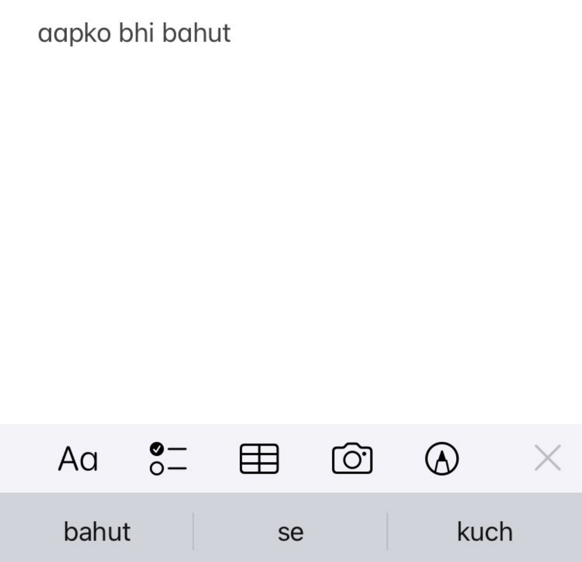
Background: written language and technology

- 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?

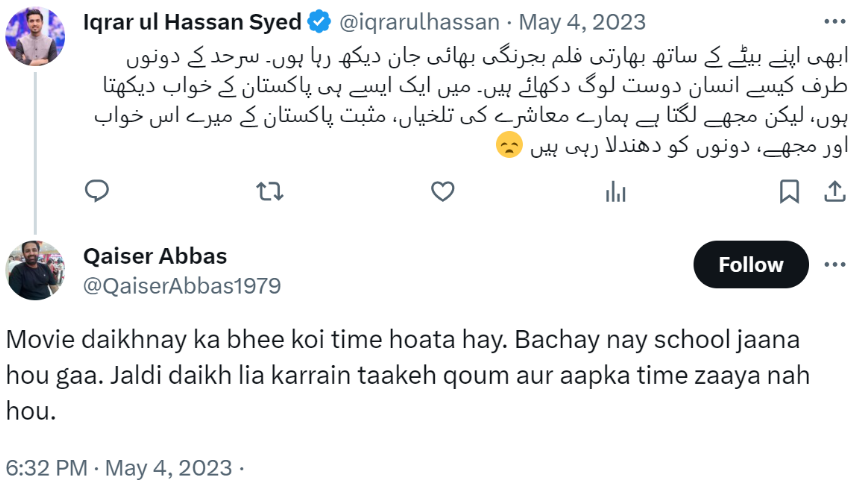
- Some possibilities:
  - Increased prestige of English in South Asia
  - The use of English as a *lingua franca*
  - “Hybrid Identity” of South Asians as a result of colonization and Western influence (Atta, 2021)
  - New ways of expressing linguistic identity outside of script choice?

Roman Hindi-Urdu today: digital forms



aapko bhi bahut

Screenshot from iOS Hindi (Latin) keyboard



Iqrar ul Hassan Syed @iqrarulhassan · May 4, 2023

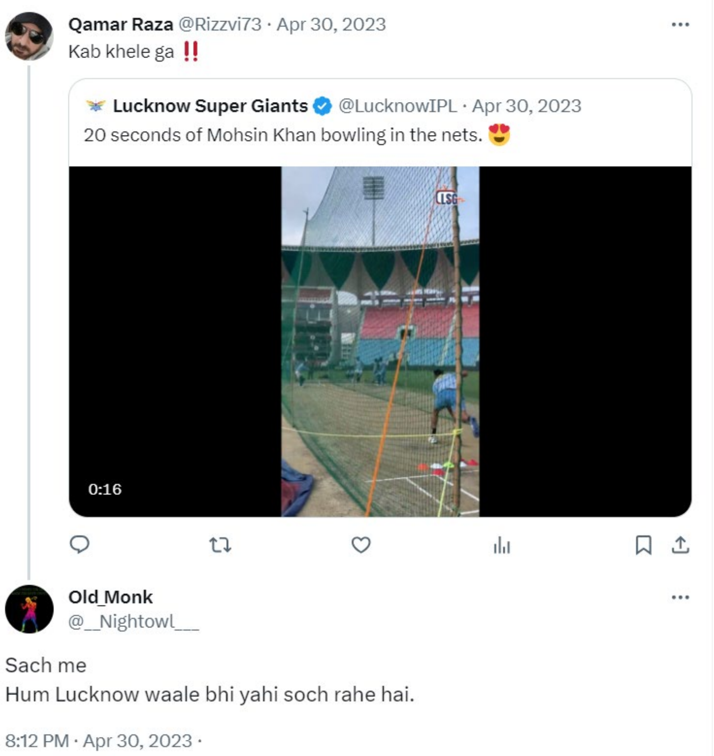
ابھی اپنے بیٹے کے ساتھ بھارتی فلم بجرنگی بھائی جان دیکھ رہا ہوں۔ سرحد کے دونوں
طرف کیسے انسان دوست لوگ دکھائے ہیں۔ میں ایک ایسے ہی پاکستان کے خواب دیکھتا
ہوں، لیکن مجھے لگتا ہے ہمارے معاشرے کی تلخیاں، مثبت پاکستان کے میرے اس خواب
اور مجھے، دونوں کو دھندلا رہی ہیں 😞

Qaiser Abbas
@QaiserAbbas1979

Movie daikhnay ka bhee koi time hoata hay. Bachay nay school jaana hou gaa. Jaldi daikh lia karrain taakeh qoum aur aapka time zaaya nah hou.

6:32 PM · May 4, 2023 ·

“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.”



Qamar Raza @Rizzvi73 · Apr 30, 2023
Kab khele ga ‼️

Lucknow Super Giants @LucknowIPL · Apr 30, 2023
20 seconds of Mohsin Khan bowling in the nets. 😍

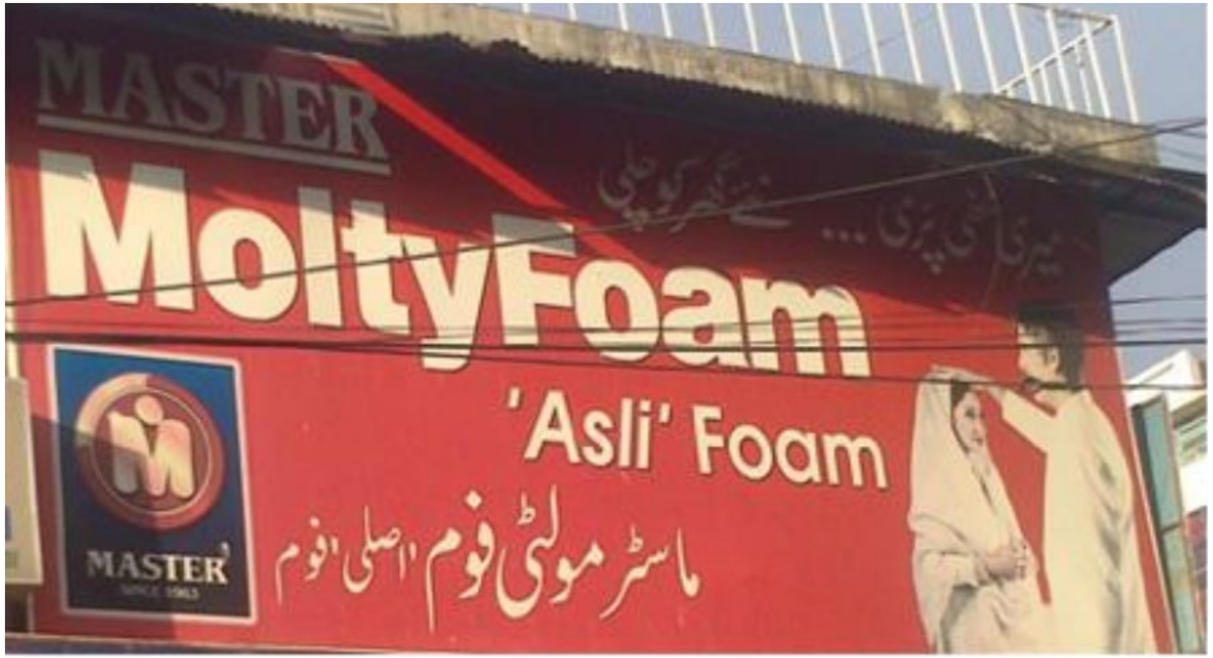
Old_Monk
@__Nightowl___

Sach me
Hum Lucknow waale bhi yahi soch rahe hai.

8:12 PM · Apr 30, 2023 ·

“When will he play?”
“Seriously, we’ve been wondering the same here in Lucknow”

Roman Hindi-Urdu today: linguistic landscapes



Advertisement in Rawalpindi, Pakistan
Photo: Atta, 2021

Roman Hindi-Urdu today: linguistic landscapes

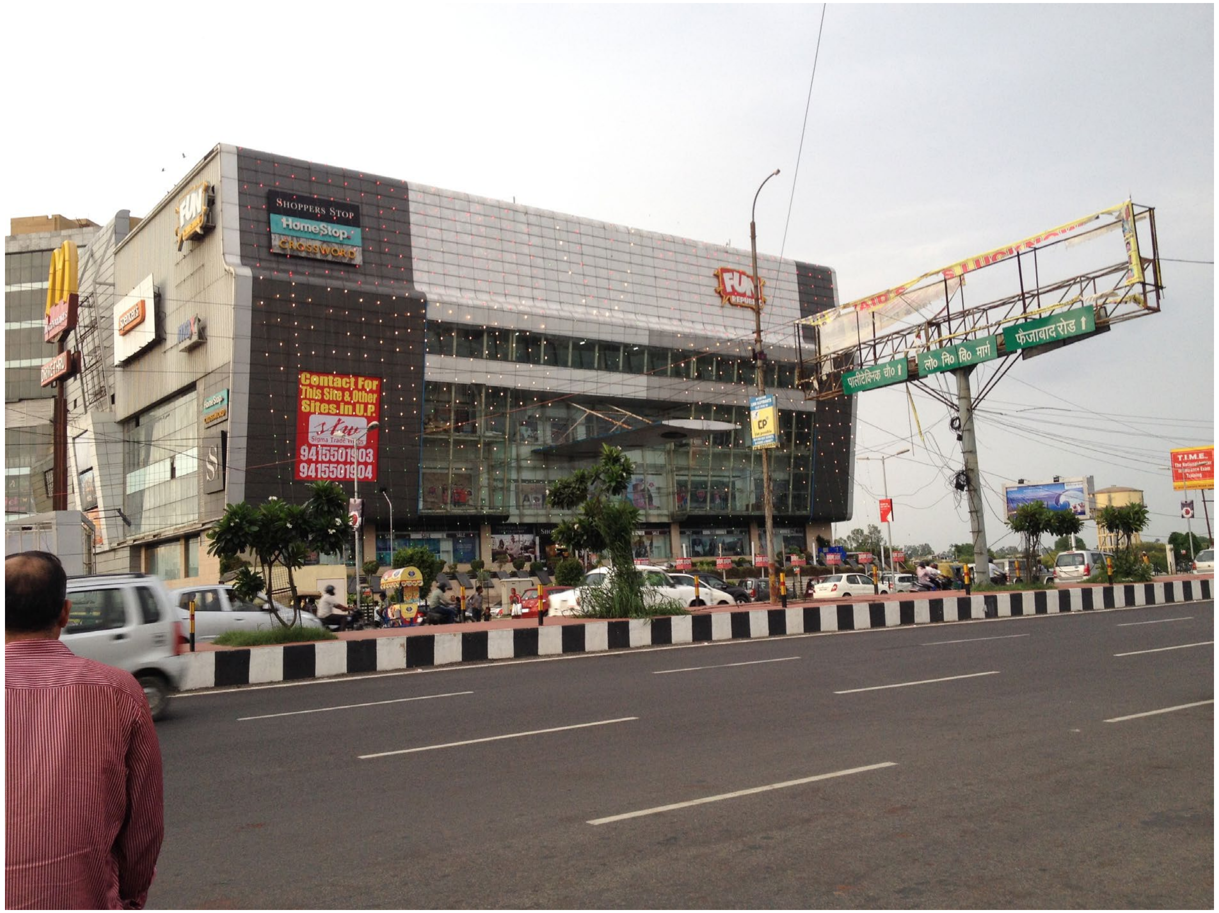


Photo from Lucknow, India, 2014

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
- W*riting 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*)

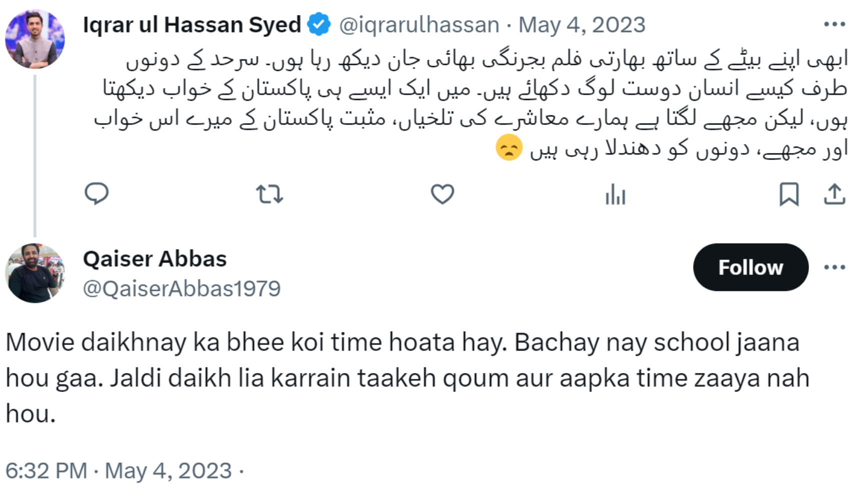
Research questions

- 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

ابھی اپنے بیٹے کے ساتھ بھارتی فلم بجرنگی بھائی جان دیکھ رہا ہوں۔ سرحد کے دونوں
طرف کیسے انسان دوست لوگ دکھائے ہیں۔ میں ایک ایسے ہی پاکستان کے خواب دیکھتا
ہوں، لیکن مجھے لگتا ہے ہمارے معاشرے کی تلخیاں، مثبت پاکستان کے میرے اس خواب
اور مجھے، دونوں کو دھندلا رہی ہیں 😞

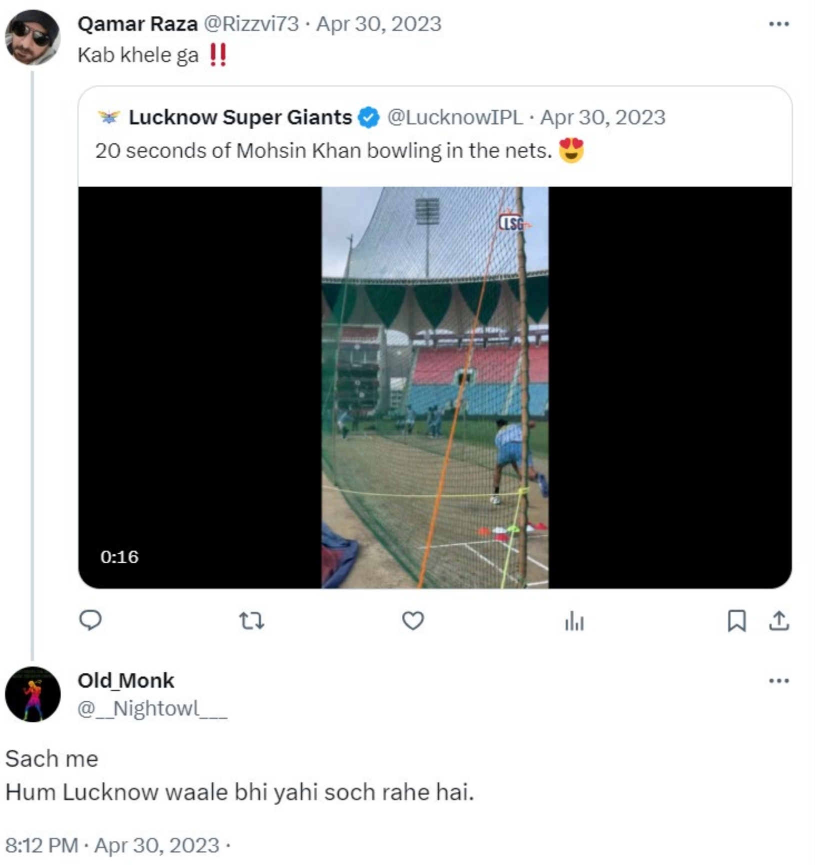
Qaiser Abbas
@QaiserAbbas1979

Follow

Movie daikhnay ka bhee koi time hoata hay. Bachay nay school jaana hou gaa. Jaldi daikh lia karrain taakeh qoum aur aapka time zaaya nah hou.

6:32 PM · May 4, 2023 ·

"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."



Qamar Raza @Rizzvi73 · Apr 30, 2023

Kab khele ga ‼️

Lucknow Super Giants @LucknowIPL · Apr 30, 2023

20 seconds of Mohsin Khan bowling in the nets. 😍

Old_Monk
@__Nightowl__

Sach me
Hum Lucknow waale bhi yahi soch rahe hai.

8:12 PM · Apr 30, 2023 ·

"When will he play?"
"Seriously, we've been wondering the same here in Lucknow"

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!?

	Bilabial	Labio-dental	Dental	Alveolar	Retroflex	Palatal	Velar	Uvular	Glottal
Stops	**p pʰ b bʱ**		**t̪ t̪ʰ d̪ d̪ʱ**		**ʈ ʈʰ ɖ ɖʱ**		**k kʰ g gʱ**	**q**[1]	

Stops: observations

- When straightforward Roman script analogues are available, they are typically used
  - /p t̪ k t͡ʃ b d̪ g d͡ʒ/ -> <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

Phoneme	/t̪/	/d̪/	/ʈ/	/ɖ/
Occurrences	1,052	773	364	155

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 /gʱ/
  - Likely from early literature; used diacritics to distinguish
- Variation in representing /x/ and /ɣ/ may reflect phonological variation
  - /x ~ kʰ / but /ɣ ~ g/

	/ɣ/	/g/	/x/	/k/
<g>	100	3326	0	0
<k>	0	0	0	15558

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

Expected surface form	<v>	<w>
[v]	276	1217
Free variation	46	107

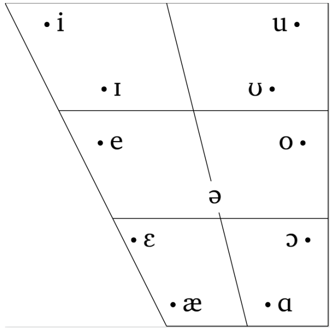
Orthographic representations of /v/ by phonological enviroment

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)

	ɑ	e	ɛ	ɪ	i	u	ʊ	o	ɔ	ə
Gilchrist (1803)	<a>	<e>	<e>	<i>	<ee>	<oo>	<oo>	<o>	<uo>	<u>
Rahman (1923)	<ā>	<ē>	<e>	<i>	<ī>	<ū>	<u>	<o>	<au>	<a>
Sharma (1937)	<ā>	<e>/<ai>	<e>	<i>	<ī>	<ū>	<u>	<o>	<au>	<a>
Khan (2000)	<a>	<e>	<E>	<I>	<i>	<u>	<U>	<o>	<O>	<A>

- 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

	<i>	<ai>	<e>	<ee>
/ɪ/	**1246**	0	2	0
/i/	**2244**	9	44	342
/ɛ/	0	**806**	163	0
/e/	14	1320	**13570**	6

Orthographic representations of front vowels

	<o>	<u>	<au>	<oo>
/ʊ/	0	**3026**	0	0
/u/	52	**1368**	0	225
/o/	**6560**	0	0	31
/ɔ/	239	0	**927**	0

Orthographic representations of back high- and mid- vowels

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. /g$^{\text{ɦ}}$/)
  - Longer orthographic representations for a single phoneme are generally not favored (i.e. /t͡ʃh t͡ʃː/)
  - 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 ɣ/)

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

- Ahmad, R. (2011). Urdu in Devanagari: Shifting orthographic practices and Muslim identity in Delhi. *Language in Society*, *40*(3), 259-284.
- Atta, A. (2021). Scripts on linguistic landscapes: A marker of hybrid identity in urban areas of Pakistan. *Journal of Nusantara Studies (JONUS)*, *6*(2), 58-96.
- Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the first workshop on computational approaches to code switching* (pp. 116-126).
- Faruqi, S. R. (2003). A Long History of Urdu Literary Culture, Part I. *Literary cultures in history: Reconstructions from South Asia*, 805-63.
- Gilchrist, J. B. (Ed.). (1803). Polyglot fables. Printed at the Hurkaru Office.
- Grover, V. (2016). Perception and Production of /v/ and /w/ in Hindi speakers (Doctoral Dissertation, City University of New York).
- Khan, M. Ahmad. (2000). Urdu phonology. Dept. of Linguistics, Aligarh Muslim University.
- Kipling, R (1890). Gunga Din. Retrieved October 22, 2022 from https://www.poetryfoundation.org/poems/46783/gunga-din
- Meletis, D. (2018). What is natural in writing?: Prolegomena to a Natural Grapholinguistics. *Written Language & Literacy*, *21*(1), 52-88.
- Meletis, D., & Dürscheid, C. (2022). Writing Systems and Their Use. In *Writing Systems and Their Use*. De Gruyter Mouton.
- Mouresioti, E., & Terkourafi, M. (2021). Καλημέρα, kalimera or kalhmera?: A mixed methods study of Greek native speakers' attitudes to using the Greek and Roman scripts in emails and SMS. Journal of Greek Linguistics, 21(2), 224-262.
- Nema, N., & Chawla, J. K. (2018). The Dialectics of Hinglish: A Perspective. *Applied Linguistics Papers*, (25/2), 37-51.
- Ohala, M. (1994). Hindi. *Journal of the International Phonetic Association*, *24*(1), 35-38.
- Pierrehumbert, J., & Nair, R. (1996). Implications of Hindi prosodic structure. *Current trends in phonology: Models and methods*, *2*, 549-584.
- Rahman, A. (1923). Urdu conversational exercises. Karachi: Azizi'z Oriental Book Depot.
- Rahman, T. (2011). From Hindi to Urdu : a social and political history. Oxford University Press.
- Sharma, R. N. (1937). Roman Urdu: A comprehensive study in Hindustani. R. N. Sharma.