Developing Orthographic Conventions of Roman Hindi-Urdu

Saiya Karamali

Abstract. The Roman script has become ubiquitous over the past few decades for writing Hindi-Urdu online. Not only have the orthographic conventions of Roman Hindi-Urdu not been studied in detail, but the lack of prescriptive rules gives us an opportunity to examine how a writing system might develop organically. After presenting a brief history of Roman Hindi-Urdu, and discussing possible reasons for its popularity, I use data from X (formerly Twitter) to identify the most common orthographic representations for each Hindi-Urdu phoneme. I then investigate how consistent these conventions are and which psychological, sociocultural, and technological factors appear to have shaped their development. I find a remarkable tendency toward a one-to-one relationship between orthographic representations and phonemes, but I also find some considerations which appear to override this tendency. By documenting the orthographic conventions of Roman Hindi-Urdu, this paper provides a foundation for future work on Roman Hindi-Urdu as well as a possible example of what an organically developing writing system might look like.

1. Introduction

Written language is full of the same kinds of complexities and intrigue found in spoken language. In our modern era of frequent internet communication, written language is a medium for users of vastly different backgrounds to interact with each other, and users find ways of conveying social meaning through their choice of words, script, and orthographic conventions. The study of written language has been largely neglected in linguistics, and this paper is an attempt to fill that gap.

Advances in technology over the last two to three decades have accelerated the evolution of many writing systems throughout the world. This includes shorthand used for instant messaging, visual characters such as emojis, the development of orthographies for previously unwritten lan-

Saiya Karamali 0009-0007-8598-2769
University of Washington
E-mail: karamali@uw.edu

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 361–391. https://doi.org/10.36824/2024-graf-kara
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

guages, and the application of novel scripts to languages in which they were not previously used (Meletis, 2018; Meletis and Dürscheid, 2022).

Two languages that have been significantly affected by changing writing systems are Hindi and Urdu. These closely related Indo-Aryan languages are extremely similar in informal registers. In formal speech, lexical differences grow, but syntax and phonology remain very similar (Masica, 1991). Yet, Hindi and Urdu are generally considered separate languages for historical, political, and sociocultural reasons (Ahmad, 2011), and one of the most visible differentiators between Urdu and Hindi is their use of separate writing systems. However, in recent years, speakers of both languages have begun to use the Roman (Latin) script in online and mobile communication as well as on billboards and storefronts (Atta, 2021; Bali, Sharma, Choudhury, and Vyas, 2014). Since the languages are so similar and are often indistinguishable in this form, I consider them together, and will refer to them as Hindi-Urdu.

Although Roman Hindi-Urdu dates to British colonization of South Asia, it lacks prescribed spelling rules and, in the technological age, has undergone rapid adoption and evolution. In this paper, I use a dataset harvested from X (formerly Twitter) to determine how the various sounds of Hindi-Urdu are represented in Roman orthography, and to perform an analysis of what this tells us about Hindi-Urdu and about how writing systems tend to develop when given the chance to do so organically.

In Section 1, I summarize the background literature on writing systems, Roman Hindi-Urdu, the history of Hindi and Urdu, and Hindi-Urdu phonology. In Section 2, I describe the goals, dataset, and methods of this study. In Section 3, I examine the orthographic conventions for writing each phoneme in Hindi-Urdu. In Section 4, I present the entire list of correspondences between phonemes and graphemes and identify the considerations which appear to have shaped the orthographic conventions of Roman Hindi-Urdu. And in Section 5, I present several important avenues for future work.

1.1. Literature Review

1.1.1. *Terminology for Discussing Writing Systems*

Following Meletis and Dürscheid (2022), I consider a *grapheme* to be the smallest unit of a writing system. For an alphabet, that is equivalent to a letter. A *script* is the set of graphemes used for a language, and an *orthography* consists of a set of rules that govern how graphemes are put together to form words. A *writing system*, then, consists of a script and, optionally, an orthography.

1.1.2. *The History of Hindi and Urdu*

Several authors have posited that Hindi and Urdu originated from among many social dialects spoken in Delhi in the 1600s (Masica, 1991; Schmidt, 2003). The *Devanagari* script, derived from older scripts which developed on the Indian subcontinent (Masica, 1991), and Perso-Arabic script, likely brought to India by Persian invaders in the eleventh century, were both in use by then (Masica, 1991; T. Rahman, 2011; Salomon, 2003, pp. 58, 89). Starting in the 18th century, sociocultural divisions between Hindus and Muslims grew—often due to deliberate actions by British colonizers (Faruqi, 2003). By the twentieth century, the name 'Urdu' and the Perso-Arabic script had become associated with Muslims; and the name 'Hindi' and the Devanagari script had become associated with Hindus (T. Rahman, 2011, p. 99). This divide continued to grow after India and Pakistan became independent, with Pakistan adopting Urdu as its national language and India adopting Hindi.

The difference in writing systems remains the most salient across-the-board difference between Urdu and Hindi (Ahmad, 2011; Rao, Vaid, Srinivasan, and Chen, 2011). In spoken language, the two are mutually intelligible, and regional differences are often greater than the differences between Urdu and Hindi (Schmidt, 2003). But although both languages include Perso-Arabic and Sanskrit loans, Urdu can be characterized by the use of more Perso-Arabic loans, and Hindi by the use of more Sanskrit loans (Kachru, 2006; Shapiro, 2003). This is true of sounds, too, where Urdu speakers are more likely to have acquired the Perso-Arabic loans [q x ɣ z f ʃ], and many Hindi speakers have acquired the Sanskrit loans [ɳ ʃ ʂ] (Schmidt, 2003). These differences are probably not entirely sociocultural; for instance, many Hindi speakers also may distinguish /f ʃ z/ (ibid.). Additionally, many self-identified Urdu speakers do not use the Perso-Arabic loans, or use them less frequently, and the same is true for self-identified Hindi speakers with Sanskrit loans (Fatima, 2011; Mangrio and Ali, 2014; Ohala, 1983). Still, these sounds are another possible way that speakers might distinguish Urdu from Hindi.

1.1.3. *History and uses of Roman Hindi-Urdu*

Roman Hindi-Urdu was used as early as the nineteenth century, particularly by European missionaries and British Indians (Ahmad, 2011; Nema and Chawla, 2018). The earliest Roman Hindi-Urdu publication with which I am familiar is Gilchrist's (1803) multilingual translation of Aesop's Fables, which includes Persian, Arabic, Hindi-Urdu, and several other South Asian languages, all written in the Roman script. Many Hindi-Urdu pedagogical materials since that time have used the Roman script as well, which may be because they were geared towards British learners or due to lower cost of publishing (e.g., Khan, 2000; A. Rah-

man, 1923; R. N. Sharma, 1937). Early Roman Hindi-Urdu publications generally devised their own systems of one-to-one correspondences between Roman characters and Hindi-Urdu sounds, and used a number of diacritics to do so. Some conventions from this period, such as the use of ⟨kh⟩ and ⟨gh⟩ to represent /x/ and /ɣ/, appear to have carried over to the present day, but many others—such as the representations for the vowels—have changed.

In general, Indians and Pakistanis under British rule did not adopt the Roman script (Faruqi, 2003), and upon independence in 1947, India and Pakistan officially adopted the traditional Hindi and Urdu scripts, respectively. Why, then, has Roman Hindi-Urdu exploded in popularity in recent years? Certainly, the emergence of technology has introduced the need to type, which was not always possible in scripts other than the Latin script. But with the introduction of physical and virtual keyboards for many more scripts, facility of use probably does not tell the whole story. Recent statistical analyses of Romanization are sparse, but languages appear to differ in whether the Roman script has continued to be used: for instance, Androutsopoulos (2009) and Mouresioti and Terkourafi (2021) both found that Romanization is unpopular among Greek speakers, but Bali, Sharma, Choudhury, and Vyas (2014) found that 84 percent of a sampling of Hindi Facebook posts were written in the Roman script.

A possible reason for the popularity of the Roman script in South Asia today is the prestige of English—which uses the Roman Script—in that region. English was introduced by the British, and was taught in elite schools dating back to the 19th century (T. Rahman, 2020). And despite rising opposition to English at the time of independence, it still became an official language of both India and Pakistan (National Assembly of Pakistan, 1973; Parliament of India, 1963) as well as a *lingua franca* in multilingual regions of both countries (Mukherjee, 2010; Zaidi and Zaki, 2017). English-language education remains widespread in private schools in both India and Pakistan, leading many in the middle and upper classes to be comfortable using both the English language and the Roman script in formal settings (T. Rahman, 2020; Ramanathan, 2016).

Starting in the 1990s, economic development and globalization have also contributed to increased prestige of English in South Asia, to the point where it has become "'the' language of Indian youth" (Nema and Chawla, 2018, emphasis in original). It is reasonable to suppose, then, that the script of English has gained prestige along with the language. Previous linguistic work has found concrete uses for Roman Hindi-Urdu, as well. Rosowsky (2010) observed a Pakistani immigrant community in the UK and found that the Roman script is often used for teaching Urdu and for teaching the Quran, and that the youth often use the Roman script to transcribe Urdu songs and poetry. For these youth, the Roman script is a tool to help preserve cultural, linguistic, and re-

ligious identity. Atta (2021) found that Roman Urdu is commonly used in billboards and storefronts in Pakistani cities, citing citizens' 'hybrid identity' due to colonization and globalization.

Thus, today's Roman Hindi-Urdu appears to be neither a colonial imposition nor a direct reaction to technology. Instead, it is a response to the modern needs and attitudes of many Hindi-Urdu speakers, and appears to be an increasingly important part of their linguistic repertoire.

1.1.4. *Writing Systems*

Previous studies have examined how people adapt new scripts to their own languages. In some cases, users have clearly articulated philosophies on how they do this: Androutsopoulos (2009) found that although Greek has no prescriptive standards for Romanization, most users used one of three loose 'systems': correspondence between Roman letters and Greek sounds; correspondence between the shapes of Roman and Greek letters; or correspondence between English and Greek keyboard locations. The second and third of these would then restrict possible readers to those with knowledge of Greek script. In other cases, spelling conventions are much less standardized: Akbari (2013) studied Romanized Persian text messages by Iranian users and found a wide variety of spelling techniques used, including multiple graphemes being used to represent the same phoneme, variation in whether to use digits to represent a sequence of characters, and variation in whether to omit vowels.

Sociocultural factors have also led to the use of new scripts and orthographies. Sebba (2007) describes cases of oppressed groups developing their own writing conventions to distinguish their writing from that of an oppressing group, as well as of graffiti artists using intentional "misspellings" to invoke humor or cause their work to stand out. And QDJY is an alternative orthography for Bahasa Indonesia which was deliberately created by an online community (Putra and Triyono, 2018). QDJY uses the same script as the language's official writing system, but its highly innovative spelling conventions serve to identify QDJY users as part of that community.

Meletis (2018) identifies four factors which may affect which script and orthographic conventions eventually become widespread for a particular language. The first of these is *linguistic fit*, or how well a writing system reflects the language. Generally, this is determined by how close to one-to-one the correspondence between sounds and their representations is. The second factor is *psychological* or *cognitive* fit, or the relationship between a writing system and our psychological, physical, and cognitive faculties. Previous studies such as Rao, Vaid, Srinivasan, and Chen (2011) have shown that some writing systems are more easily processed cognitively than others. Higher linguistic fit will often lead to higher psychological fit, but psychological fit can be measured independently

through empirical experiments of reading and writing ability. The third factor is *sociocultural fit*, or what attitudes users have about a script and how well it fits their identities. The fourth possible factor is *technological fit*, or how easily the script can be used on computers and mobile devices. Meletis posits that these factors shape how likely a script is to be used for a particular language.

The Devanagari script has the highest linguistic fit of the writing systems used for Hindi-Urdu. That script has close to a one-to-one correspondence between sounds and their representations, whereas the Urdu script has sounds which are represented by multiple symbols, symbols which represent multiple sounds, and at least one silent letter; and vowels are generally omitted. The Hindi writing system has a higher psychological and cognitive fit, as well: Rao, Vaid, Srinivasan, and Chen (2011) found that Hindi-Urdu biliterate readers were consistently able to read more quickly in Devanagari than in Perso-Arabic.

With regard to sociocultural fit, Unseth (2005) identifies three sociocultural aims that script choice can achieve: identification with a particular group; distanciation from a particular group; and ease of interaction with outside communities. Indeed, the Perso-Arabic script has historically been so closely identified with South Asian Muslims that it has made them resistant to using both the Devanagari and Roman scripts, to the point where Urdu speakers sometimes use Perso-Arabic writing conventions even when writing in the Devanagari script, where such conventions are of no linguistic use (Ahmad, 2011). Additionally, the Roman script was historically associated with colonization, from which South Asians may have wished to distance themselves (ibid.). Recently, however, the Roman script has become associated with economic prosperity in South Asia (Nema and Chawla, 2018), which may have increased the sociocultural fit of the Roman script in both Hindi and Urdu. And finally, the Roman script may have the highest technological fit, because it was among the first scripts used on computers and mobile devices; but perhaps not overwhelmingly so, given that Urdu and Hindi keyboards are also available, particularly on mobile devices.

2. Methods

2.1. Dataset

I obtained my dataset by taking a random sampling of posts using the X streaming API[1], which returns a random sample of posts which are created while the stream is running. I collected posts between May 1

1. https://developer.x.com/en/docs/tutorials/stream-tweets-in-real-time

and May 9, 2023, resulting in a total of 56.5 GB of posts being collected. I filtered the posts to include only those in ASCII characters which were classified by X as either Hindi or Urdu. Web addresses and mentions were removed from posts, and duplicate reposts and quote posts were removed from the dataset. The resulting dataset consists of 8909 posts, only a small fraction of the posts collected.

More posts could not be collected because X removed access to its APIs for academic research as of June 8, 2023. X and other online fora are key sources of emerging linguistic data and unfortunately, none of these are currently available for reasonable pricing. X, for instance, now charges $5000 per month to perform the data collection carried out in this study. However, the 8909 posts provide a sample large enough to perform a linguistic analysis of what spelling conventions are used, and which I supplement with more fine-grained qualitative analysis.

2.2. Methods

From the 8909 posts I retained from X, I extracted the 2000 most frequent orthographic "words", where a word is a group of characters separated by whitespace. No effort was made at this stage to group together alternate orthographic representations which correspond to the same semantic word. I did, however, remove words which I judged to be proper names or loanwords from English, since proper names may have more standardized spellings and the spellings of English loanwords could be influenced by their English orthography.

In order to analyze each phoneme, I first used my own knowledge as well as manual examination of posts to identify the possible orthographic representations of each phoneme. I then filtered the set of orthographic words to include only words with these representations, and then manually inspected the resulting data and removed words which appeared not to include the target phoneme. I did this using my own knowledge as an Urdu speaker to determine the most likely phonetic form for each word, with assistance from Wiktionary (2023), where needed. Wiktionary has been previously used as a source of community-built transcriptions for a given community's language variety (see Meyer and Gurevych, 2012; Noruzi, 2023), and appeared to be the most reliable source available. I then counted the occurrences of each orthographic representation for the given phoneme.

A few phonemes were too rare to exhibit meaningful patterns within the top 2000 words. In these cases, as noted in Section 3, I expanded the dataset by searching directly for possible representations of words which contain the target phoneme. I identified these words by using wordfreq (Speer, 2022) to list the most frequent Hindi words containing a given Devanagari character.

In Section 3, I present my findings for each phoneme in the Hindi-Urdu phonemic inventory. I discuss patterns in how various natural classes of phonemes are written and explore what these tell us about Roman Hindi-Urdu writing conventions. I then explore the possible effects of phonological variation, sociocultural variation, and perception of the various sounds by Hindi-Urdu speakers.

3. Analysis

In this section, I begin with a discussion of the Hindi-Urdu phonemic inventory and of which phonemes are likely to be too rare to draw conclusions on. I then discuss the orthographic representations of the various classes of phonemes, starting with the consonants. For each phoneme for which I have sufficient data, I determine the possible orthographic representations and discuss what these tell us about linguistic, psychological, sociocultural, and technological fit in Roman Hindi-Urdu.

3.1. Phonemic Inventory

Table 1 displays the generally accepted phonemic consonant inventory of Hindi-Urdu (Ohala, 1994; Schmidt, 2003; Shapiro, 2003). Perso-Arabic and Sanskrit loans are noted.

Hindi-Urdu has a four way stop contrast for the four main places of articulation, and some speakers additionally have a voiceless uvular stop. Palatal affricates also exhibit a four way voicing contrast, and most native consonants also exhibit a phonemic length contrast word-medially before lax vowels. The geminate nasals, fricatives, and approximants do not occur frequently enough in my data to analyze in detail, but I do analyze the geminate stops as a possible model for how other geminates could potentially be written. /ɖ̤/ and /ʈ̤/ are in complementary distribution in native words, but /ɖ̤/ can occur in additional environments in English loanwords and thus is generally considered its own phoneme (Schmidt, 2003). Among nasal stops, only /m mː n nː/ are widespread across speakers and across phonemic environments; /ɳ ɲ ŋ/ occur contrastively only in formal Hindi registers, and only in homorganic nasal-stop clusters (Ohala, 1983, p. 5). /f z ʂ ʃ x ɣ h/ also occur only among a subset of speakers. Several authors consider /v/ to be an approximant because it is often realized as [w] (Kachru, 2006; Schmidt, 2003; Shapiro, 2003); and some authors transcribe /v/ with /ʋ/, as well (Ohala, 1994). But Pierrehumbert and Nair (1996) find clear frication in [v], and identify a pattern of complementary variation with [w]. Additionally, /vː/ is never pronounced as an approximant (Ohala, 1983, pp. 6-7). Thus, I analyze /v/ with the fricatives in my analysis.

	Bilabial	Labio-dental	Dental	Alve-olar	Retroflex	Palatal	Velar	Uvu-lar	Glot-tal
Stops/ Affricates	p pː pʰ pʰː b bː bʱ		t̪ t̪ː t̪ʰ t̪ʰː d̪ː d̪ː d̪ʱ d̪ʱː		ʈ ʈː ʈʰ ʈʰː ɖː ɖː ɖʱ ɖʱː	t͡ʃ t͡ʃː t͡ʃʰ t͡ʃʰː d͡ʒː d͡ʒː d͡ʒʱ d͡ʒʱː	k kː kʰ kʰː g gː gʱ gʱː	q[1]	
Nasals	m mː			n nː	ɳ[2]	ɲ	ŋ		
Fricatives		f[1] v vː		s sː z[1]	ʂ	ʃ[3]	x[1] ɣ[1]		h[2] ɦ
Approx-imants						j			
Lateral Approx-imants				l lː					
Taps				ɾ ɾː	ɽ ɽʱ				

TABLE 1. The phonemic consonant inventory of Hindi-Urdu

[1] Loaned from Perso-Arabic.
[2] Loaned from Sanskrit.
[3] Loaned from Sanskrit and Perso-Arabic.

A few different analyses of the Hindi-Urdu vowel system have been proposed. The first of two main points of contention is whether Hindi-Urdu has phonemic vowel length or a phonemic tenseness contrast. According to Schmidt (2003), /iː eː uː oː/ and /ɪ ɛ ʊ ɔ/ do differ in quality, but have nonetheless generally been analyzed as differing in length. The second point of contention is whether Hindi-Urdu has any diphthongs, and if so, how many. Most phonological grammars of Hindi Urdu posit that there are no diphthongs (Kachru, 2006; Ohala, 1983; 1994; Schmidt, 2003; Shapiro, 2003); by contrast, several recent studies have found between two and eighteen diphthongs in Hindi-Urdu, but these seem to differ significantly by dialect (Khurshid, Usman, and Javaid, 2003; Mishra and Bali, 2011; B. D. Sharma, 2018).

Assuming a tenseness contrast and no diphthongs results in the phonemic inventory shown in Figure 1. All but /æ/ occur in native words and exhibit a phonemic nasalization contrast. /æ/ is generally considered to be an additional phoneme, but it mainly occurs in loans from English, so I will not discuss it explicitly except where relevant to the orthographic representations of other vowels.

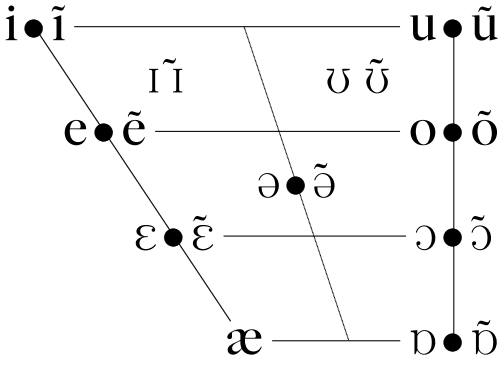

FIG. 1. An eleven vowel phonemic inventory of Hindi-Urdu (Ohala, 1983)

3.2. Oral Stops, Velar Fricatives, and Affricates

I begin by discussing the orthographic representations of the oral stops. Because most speakers have at least 30 oral stop phonemes in Hindi-Urdu, and the Roman script only has six standard letters for writing stops without diacritics, the stops are a good case for assessing the importance of linguistic fit. Yet, Roman Hindi-Urdu has developed remarkably consistent ways of marking aspirated, breathy voiced, and geminate

consonants. Place of articulation is represented with ⟨b d g p t k⟩, similarly to English: ⟨p b⟩ are bilabial stops; ⟨t d⟩ are dental stops; and ⟨k g⟩ are velar stops. ⟨c⟩ is in general not used: the only word in which it appears outside of ⟨ch⟩ is ⟨crore⟩ 'ten million', a spelling which dates back to 19th century British India (Oxford Dictionary Press, 1893). To represent gemination and voice quality contrasts, a second character is used. To represent an aspirated stop, ⟨h⟩ follows a voiceless stop symbol, and to represent a breathy voiced stop, ⟨h⟩ follows a voiced stop symbol. To represent a geminate stop, the stop symbol is repeated. The graphemic representations resulting from this system are shown in Table 2. Retroflex stops do not use unique symbols, and are instead represented using the same symbols as dental stops. I did not identify any aspirated retroflex stops, geminate retroflex stops, or aspirated geminate stops in the top 2000 words, although these are possible according to Ohala (1983, p. 2). If these follow the same pattern as the other stops, aspirated and geminate retroflex stops would be represented in the same way as their dental counterparts, and aspirated geminate stops could be represented using three-character graphemes with the stop character appearing twice, followed by ⟨h⟩. Further study will be required in order to determine whether this is the case.

Table 2. Representations of the various stop contrasts for voiced dental and retroflex stops

Place of Articulation	Bilabial	Dental	Velar	Retroflex
Singleton Unaspirated	⟨p⟩ ⟨b⟩	⟨t⟩ ⟨d⟩	⟨k⟩ ⟨g⟩	⟨t⟩ ⟨d⟩
Singleton Aspirated/Breathy	⟨ph⟩ ⟨bh⟩	⟨th⟩ ⟨dh⟩	⟨kh⟩ ⟨gh⟩	No data
Geminate Unaspirated	⟨pp⟩ ⟨bb⟩	⟨tt⟩ ⟨dd⟩	⟨kk⟩ ⟨gg⟩	No data

Although retroflex Hindi-Urdu stops occur in the same phonetic environments as dental stops, they are much rarer in native words. At the same time, retroflex stops occur almost exclusively in loanwords from English (Center for Language Engineering, n.d.). In the top 2000 words in my dataset, which excludes English loanwords, /ʈ/ occurs 5 times and /ɖ/occurs twice, in all cases represented by ⟨t⟩ and ⟨d⟩. Some of these occurrences are listed in Table 3.

English loanwords often feature different phonotactic constraints from native words and they may also borrow English orthographic conventions. Thus, the distribution of retroflex stops in Hindi-Urdu may itself lead to relatively few cases of complete ambiguity. Additionally, ⟨t⟩ and ⟨d⟩ are convenient to type and may already be associated with the retroflex stops as a result of English orthography. If ambiguity between the retroflex and dental stops is indeed low, I hypothesize that using ⟨t d⟩ for both sets could maximize technological fit at low cost of psycho-

TABLE 3. Miscellaneous examples of ⟨t⟩ and ⟨d⟩

Roman spelling	⟨ghanta⟩	⟨theek⟩	⟨dal⟩	⟨tha⟩	⟨tum⟩	⟨desh⟩
Phonemic representation	[gʱənʈə]	[ʈʰiːk]	[ɖɑːl]	[ʈ̪ʰɑ]	[ʈ̪ʊm]	[deːʃ]
Meaning	'hour'	'fine'	'put'	'was'	'you'	'country'
Occurrences	10	28	6	301	288	166

logical fit. The velar fricatives /x ɣ/ also reuse the stop representations in some cases. These sounds have no clear analogue in English orthography, and in Hindi-Urdu pedagogical literature and early colonial literature, these are typically represented by ⟨kh gh⟩ and differentiated from /kʰ gʱ/ only by diacritics. Complicating the situation is phonological variation: /x/ is frequently pronounced as /kh/ by some speakers, but /ɣ/ is often pronounced as [g] (Fatima, 2011). If phonological variation is represented in writing, then, one might expect /gʱ/ to be represented by ⟨g⟩, but /kʰ/ not to be represented by ⟨k⟩. Indeed, /ɣ/ is represented by ⟨g⟩ more frequently than by ⟨gh⟩, but /kh/ is exclusively represented by ⟨kh⟩. Full frequency counts for each orthographic representation are listed in Tables 4 and 5. Thus, I conclude that the representations of /x/ and /ɣ/ do reflect phonological variation.

TABLE 4. Frequency counts by phoneme represented by ⟨k⟩ and ⟨kh⟩

	⟨k⟩	⟨kh⟩
/kʰ/	0	1011
/x/	0	366
/k/	15558	0

TABLE 5. Frequency counts by phoneme represented by ⟨g⟩ and ⟨gh⟩

	⟨g⟩	⟨gh⟩
/gʱ/	0	178
/ɣ/	100	27
/g/	3326	0

Finally, I consider the voiceless uvular stop /q/. /q/ also has no obvious analogue in English orthography, and has historically been represented by ⟨q⟩ in early literature (Gilchrist, 1803; A. Rahman, 1923; R. N. Sharma, 1937). In similar fashion to /x ɣ/, /q/ surfaces as [k] for many speakers (Fatima, 2011; Shapiro, 2003). Like the retroflex stops, /q/ is relatively rare underlyingly, and I only identified 7 instances of it in the top 2000 most common words in the dataset. Table 6 shows the frequency counts of the letters used to represent /q/ in these words. On the one hand, the usage of ⟨k⟩ could reflect phonological variation, as was the case for /x ɣ/. But ⟨k⟩ could simply be another representation of /q/, which also seems plausible because of the rarity of /q/. Like the

retroflex stops, this would then be another possible case of orthographic representations being reused for rare phonemes.

Table 6. Orthographic representations of /q/ in the data

	⟨q⟩	⟨k⟩
/q/	129	66

Finally, the affricates /t͡ʃ/ and /d͡ʒ/ also exhibit an aspiration contrast as well as a length contrast, and in their singleton unaspirated forms are represented by ⟨ch⟩ and ⟨j⟩ respectively, which seem to be straightforwardly derived from English orthography. /d͡ʒ/ shows the same tendencies as the stops, with ⟨jh⟩ and ⟨jj⟩ representing the breathy voiced and geminated forms, respectively. The fact that ⟨ch⟩ is two characters long complicates things, however. If /t͡ʃ/ matched the behavior of the stops, ⟨chh⟩ would represent /t͡ʃʰ/ and ⟨cch⟩ or ⟨chch⟩ would likely represent /t͡ʃː/. Instead, this pattern seems to break down, with /t͡ʃ/ and /t͡ʃʰ/ both being predominantly represented by ⟨ch⟩, and the rarer forms /t͡ʃː/ and /t͡ʃʰː/ having all four possible representations. Table 7 shows the full frequency counts. The instability of /t͡ʃː/ and /t͡ʃʰː/ indicates that three- and four-character representations may be harder to distinguish from each other; future work can test this hypothesis. If this is true, the rarity of longer orthographic representations could represent the maximization of psychological fit at the cost of linguistic fit.

Table 7. Orthographic representations of /t͡ʃ t͡ʃʰ t͡ʃː t͡ʃʰː/

	⟨cch⟩	⟨chh⟩	⟨ch⟩	⟨chch⟩
/t͡ʃ/	0	0	600	0
/t͡ʃʰ/	0	78	559	0
/t͡ʃː/	18	9	35	11
/t͡ʃʰː/	32	52	11	24

3.3. Nasal Stops

As with the oral stops, the orthographic representations of nasal stops are relatively straightforward, with possible ambiguities occurring only in particular phonetic environments among a relatively rare set of words. /n/ and /m/—the only nasal stops to occur in a wide range of

phonetic environments—are exclusively represented by ⟨n⟩ and ⟨m⟩, respectively. /ɳ ŋ ɲ/ are all rare in my data, and I selected words from the entire dataset in order to study these. /ɳ/ and /ŋ/ are represented by ⟨n⟩ in all of the examples in my data, and /ɲ/, which occurs rarely in my data, is written using ⟨y⟩ or ⟨n⟩. All three of these occur only before homorganic stops and affricates, and many speakers exhibit no contrast even in those cases. So ⟨n⟩ is ambiguous only in particular consonant clusters and only for certain speakers. Thus, as with retroflex stops, existing orthographic representations are again re-used to represent rare phonemes.

3.4. Sibilants and Fricatives

Next, I consider the sibilants. The main point of note here is that although the core Hindi-Urdu alveolar sibilants are /s d͡ʒ/, /ʃ/ is present in Sanskrit and Perso-Arabic loans and /z/ is present in Perso-Arabic loans only. According to Ohala (1983, p. 4), most Hindi speakers pronounce /ʃ/ as [s] and /z/ as [d͡ʒ]. In written form, I found several examples of /z/ being written with ⟨j⟩, but none of /ʃ/ being written with ⟨s⟩. The frequency counts for each grapheme used to represent /s z ʃ d͡ʒ/ are shown in Tables 8 and 9. And finally, the Sanskrit loan /ʂ/ is rare in my data but appears to be represented by ⟨sh⟩ when it does appear. Again, rare phonemes seem more likely to re-use orthographic representations.

Table 8. Orthographic representations of /z/ and /d͡ʒ/

	⟨z⟩	⟨j⟩
/z/	465	192
/d͡ʒ/	0	3846

Table 9. Orthographic representations of /s/ and /ʃ/

	⟨s⟩	⟨sh⟩
/s/	5885	0
/ʃ/	0	1256

Shapiro (2003) indeed considers /ʃ/ to be a standard pronunciation, whereas /z/ almost certainly is not, so the use of ⟨sh⟩ to represent /ʃ/ may be a result of more speakers contrasting /s/ and /ʃ/ in speech. Additionally, perhaps the fact that /ʃ/ exists in both Sanskrit and Arabic loans neutralizes sociocultural factors which might make some Hindi speakers less likely to contrast /z/ and /d͡ʒ/ in writing even if they are aware of the contrast.

/f/ is another Perso-Arabic loan which is often pronounced as /p/ in speech. However, Shapiro (ibid.) and Mangrio and Ali (2014) also note the existence of hypercorrection in both Urdu and Hindi speakers, where /p/ may be pronounced as [f] even in native words. In the data, the only

such word I found was /pʰɪr/ 'then', which was spelled as ⟨fir⟩ and ⟨phir⟩ at nearly equal frequency. This word is the most cited example of this phenomenon and perhaps the most common one in speech, so this may well reflect actual pronunciation. Otherwise, ⟨p⟩ consistently represents /p/ and ⟨f⟩ consistently represents /f/.

The final fricative to consider is /v/, where a case of allophonic variation may give evidence for how a phoneme may be represented when the correspondence to the Roman script is not easily perceived. According to Pierrehumbert and Nair (1996), /v/ is realized as [w] in the onglide position of a syllable; that is, between a syllable-initial consonant and a vowel, or between an obstruent and a word boundary. They found [v] and [w] to be in free variation between a word-medial singleton consonant and a vowel; apparently because the singleton consonant can be analyzed either as a syllable onset or as the coda of the previous syllable. Furthermore, Grover's (2016) perceptual study found that native Hindi speakers could not reliably distinguish hearing [v] and [w]. If this is true, we might expect ⟨v⟩ and ⟨w⟩ both to be used to represent ⟨w⟩ regardless of phonetic environment.

In my data, ⟨w⟩ in fact occurs more frequently than ⟨v⟩ in all positions, including those that are realized as [v] in speech. The first two rows of Table 10 show the frequency counts of ⟨v⟩ and ⟨w⟩ for phonetic environments where [v] is the expected surface form, and the third row shows the scenario where [v] and [w] are in free variation. I have insufficient data for words where [w] is the expected surface form, but given that ⟨w⟩ is already the more likely orthographic representation, ⟨w⟩ appears to be the more likely representation across the board. I conclude, then, that the orthographic representation of /v/ does not reflect its phonetic realization, but given a perceived ambiguity in sounds represented by ⟨v⟩ and ⟨w⟩, ⟨w⟩ may be organically developing to be a standard form, possibly reflecting a tendency toward higher linguistic fit.

TABLE 10. Orthographic representations of /v/ by word position

Phonetic Environment	Expected Surface Form	⟨v⟩	⟨w⟩
Word-initial	[v]	200	984
Word-medial, following a vowel	[v]	76	233
Word-medial, following a singleton consonant	[v] or [w]	46	107

3.5. Approximants

Hindi-Urdu has the lateral approximant /l/ and four rhotic approximants/ɽ ɽʱ ɾ ɾː/. Previous work generally analyzes all of these as taps or flaps, (Kachru, 2006; Ohala, 1994; Schmidt, 2003; Shapiro, 2003), but Ohala (1994) notes that /ɾ/ is often realized as a trill, and /ɾː/ is always realized as a trill. The lateral approximant takes the straightforward representation ⟨l⟩ with little ambiguity, but the rhotic approximants are more complicated. I have already noted that retroflex *stops* are not distinguished orthographically from other apical stops. It is notable, then, that /ɽ/ is often written as ⟨d⟩. This distinguishes the retroflex and alveolar flaps from each other, but it means /d̪/, /ɖ/, and /ɽ/ are predominantly written the same way. Since /ɖ/ and /ɽ/ occur in complementary distribution except in English loanwords, there would be no ambiguity between these two phonemes for any of the words in my dataset. There could potentially be confusion in English loanwords, but since these often follow different phonotactic constraints, it is possible that they would be recognizable using other features.

The frequencies of all of the orthographic representations which represent rhotic approximants are shown in Table 11. A future perceptual study may shed light on whether distinguishing /ɽ/ from /r/ turns out to be more meaningful with regard to reducing ambiguity.

Table 11. Orthographic representations of rhotic approximants

	⟨r⟩	⟨rr⟩	⟨d⟩	⟨dd⟩	⟨rh⟩	⟨dh⟩
/ɾ/ or /ɾː/	6223	33	0	0	0	0
/ɽ/	63	0	129	0	0	0
/ɽʱ/	0	0	0	0	0	33
/d̪/ or /ɖ/	0	0	3470	0	0	0
/d̪ː/	0	0	0	95	0	0
/d̪ʱ/	0	0	0	0	0	462

It appears that /ɾ/ and /ɾː/ are both represented by ⟨r⟩. However, /ɾ/ in the interjection /əɾeː/ and the word /ɖəɾ/ 'fear' seems to be sometimes represented by ⟨rr⟩ (see Table 12), and /ɾː/ seems to never be represented by ⟨rr⟩. The reason for this is unclear, but it is possible that a larger dataset could show more uses of ⟨rr⟩. Overall, possible correspondences between the different taps and Roman script representations are much less obvious, and this seems to lessen the likelihood of achieving high linguistic fit for these phonemes.

TABLE 12. Orthographic representations of /əreː/ 'hey!' and /ɖər/ 'fear'

	/əreː/	/ɖər/
⟨rr⟩	23	10
⟨r⟩	32	16

3.6. /j/ and /ɦ/

The consonants that remain to be examined are /j/ and /ɦ/. Both of these are notable in my data because they allow me to consider whether Roman Hindi-Urdu inherits certain features from the historical Hindi and Urdu scripts.

/j/ often appears before or after a high or mid front vowel, as in [d̪ɪjɑ] 'give' and [gəji] 'go'-3pst (ibid.). These are generally not transcribed as vowel hiatus—and the glide is written in both the Hindi and Urdu orthographies—but vowel hiatus would presumably be a possible way to analyze these sequences. Table 13 shows the representations of various sequences that could potentially be perceived either as vowel hiatus or as an approximant between two vowels. In all of these sequences, it is more common to include the ⟨y⟩ than to omit it. I conclude that these sequences are typically not viewed as vowel hiatus, and this could be either because of how they are perceived in speech or because of how they have historically been written.

TABLE 13. Representations of high vowel-glide-vowel clusters

Representation	⟨iya⟩	⟨ia⟩	⟨iye⟩	⟨ie⟩	⟨iyo⟩
Phonetic Form	/ɪjɑ/	/ɪjɑ/	/ɪje/	/ɪje/	/ɪjo/
Occurrences	378	64	88	51	13

A notable feature of /ɦ/ is that both the Hindi and Urdu orthographies sometimes mark a historical word-final /ɦ/ which is now deleted (Oberlies, 2005). Of course, a word-final orthographic ⟨h⟩ is also common in English in words such as *ah* and *oh*. So, I consider whether an orthographic word-final ⟨h⟩ exists in Roman Hindi-Urdu and whether it is more likely in words where this mirrors the Hindi and Urdu representations. The vast majority of vowel-final words in the data are written with no final ⟨h⟩. Of the words that are indeed written with word-final ⟨h⟩, the majority end in low vowels, and the majority have a word-final ⟨h⟩ in Hindi orthography, as shown in Table 14. The representations in Urdu orthography often overlap but may not match exactly. Still, this is

potentially another case of features of historical writing systems being carried over to the Roman script.

Table 14. ⟨h⟩-final Roman Hindi-Urdu words by their Hindi representation

Word-final /ɦ/ represented in Hindi Orthography	No word-final /ɦ/ represented in Hindi
566	352

3.7. Oral Vowels

Of the eleven vowels in the Hindi-Urdu phonemic inventory, all of them are present in several dialects of English. Yet, English orthography does not represent these uniquely, which makes the vowels an interesting study of how the Roman script maps onto Hindi-Urdu phonemes. Devanagari does represent these uniquely (Shapiro, 2003), whereas Perso-Arabic represents the vowels with a total of only three graphemes. In his Roman-script translations of Aesop's Fables, Gilchrist (1803) devises his own system for representing the vowels: He adds ⟨uo⟩, ⟨ee⟩, ⟨oo⟩, and ⟨y⟩, but still appears to use ⟨oo⟩ to represent both /u/ and /ʊ/. Pedagogical materials published since have used a wide variety of systems: A. Rahman (1923) and R. N. Sharma (1937) use ⟨ī ē ū⟩ to represent /i e u/ and ⟨au⟩ to represent /ɔ/; and Khan (2000) uses capital letters to represent the lax vowels.

I now analyze the orthographic representations of vowels in my data. I generally will focus on monophthongs in this analysis, but /ai oi au əu/—which may be diphthongs in some dialects, and cases of hiatus in others—appeared prominently in my data and their representations sometimes overlapped with the representations of monophthongs, so I include these as well in order to provide a fuller picture of the monophthongs.

First, I consider the front vowels as well as /ɑi/ and /oi/. Frequency counts of the graphemes and grapheme combinations used to represent these are shown in Table 15. ⟨i⟩ almost always represents both /i/ and /ɪ/. Using ⟨ee⟩ for /i/ is a possible way to draw that distinction, but is relatively uncommon. Interestingly, this is a departure from historical Roman-Urdu conventions at the expense of lower linguistic fit, and the reason for this is not entirely clear, given that ⟨ee⟩ is commonly used in English orthography to make exactly this distinction. A larger distinction is drawn between /e/ and /ɛ/, with ⟨e⟩ typically used to represent /e/, and ⟨ai⟩ typically used to represent /ɛ/. This, in turn, blurs the dis-

tinction between /ɑi/ and /ɛ/, but since /ɑi/ is not necessarily monosyllabic, it is possible that that distinction is clearer from context. /oi/ occurs in only one of the 2000 most common words, /koi/ 'somebody', and it is generally represented by ⟨oi⟩.

TABLE 15. Orthographic representations of front vowels and potential diphthongs involving front vowels

	⟨i⟩	⟨ai⟩	⟨oi⟩	⟨ii⟩	⟨e⟩	⟨ee⟩	⟨oyi⟩
/ɪ/	1246	0	0	0	2	0	0
/i/	2244	9	0	6	44	342	0
/ɛ/	0	806	0	0	163	0	0
/ɑi/	0	3986	0	0	0	0	0
/oi/	0	0	378	0	0	0	11
/e/	14	1320	0	0	13570	6	0
/ə/	0	0	0	0	137	0	0

I now turn my attention to sequences involving back vowels. The frequency counts of their representations are shown in Table 16. Analogous to my findings for front vowels, ⟨oo⟩ seems to have been largely abandoned in favor of using ⟨u⟩ to represent both /ʊ/ and /u/. However, ⟨o⟩ is used to represent /o/ and ⟨au⟩ generally represents /ɔ/. Thus, for both front and back vowels, high vowels are more ambiguous in Roman Hindi-Urdu than mid-vowels, and this may be evidence that tenseness distinctions between the mid vowels are more salient to speakers. /ɑ/ is represented by both ⟨a⟩ and ⟨aa⟩, but neither of these represents any other phonemes.

I observe considerable variation in the representation of /ə/. Shapiro (2003) considers this sound to be mid-central as in [ə] or [ʌ], whereas Ohala (1994) considers it to be lower, closer to [ɐ]. That would make ⟨a⟩, which represents other low vowels, a possible candidate. Orthographically, both the Perso-Arabic and Devanagari writing systems mark /ə/ only at the beginning of a word. Interestingly, Gilchrist (1803) uses ⟨u⟩

TABLE 16. Orthographic representations of back vowels

	⟨o⟩	⟨u⟩	⟨au⟩	⟨ou⟩	⟨oo⟩	⟨a⟩	⟨aa⟩
/ʊ/	0	3026	0	0	0	0	0
/u/	52	1368	0	0	225	0	0
/o/	6560	0	0	48	31	0	0
/ɔ/	239	0	927	0	0	0	0
/ɑu/	0	0	14	0	0	0	0
/əu/	0	0	13	0	0	0	0
/ɑ/	0	0	0	0	0	3787	2961

to represent /ə/. Additionally, /ə/ is often fronted to [ɛ] before /ɦ/, where the sound following /ɦ/ is not /ə/ (Pandey, 1991). /ə/, then, has several contenders for possible orthographic representations. And as with /v/, /ə/ gives us another opportunity to see whether an allophonic process is reflected in Roman orthography. The orthographic representations of /ə/ in my data are shown in Table 17.

TABLE 17. Orthographic representations of /ə/

Representation	Ø	⟨a⟩	⟨e⟩	⟨i⟩	⟨u⟩
Occurrences	1140	12327	126	12	276

/ə/ is most commonly represented by ⟨a⟩, while using no vowel is the second most common strategy, and the use of ⟨e⟩ and ⟨i⟩ is rare. ⟨u⟩ seems to appear in words involving the first-person pronoun morpheme /ɦəm/, but in no other words. Since /ɦəm/ is a common morpheme, this may be a convention that was carried over from early publications. That there is so much variation may indicate that phonemic /ə/ is not consistently perceived as any particular vowel. It is possible that perception of /ə/ depends on its phonological environment; this would resemble English, where /ə/ typically takes on acoustic properties of neighboring vowels (Cole, Linebaugh, Munson, and McMurray, 2010). Since /ɑ/ is the most common representation, it may be the most common way that /ə/ is perceived, or this may be an example of linguistic fit being arbitrarily maximized over time. Future work can further examine whether the representations of /ə/ differ based on their phonetic environment.

Another observation from Table 18 is that the omission of /ə/ in the orthography is more common than the use of ⟨e⟩, ⟨i⟩, and ⟨u⟩, but is not close to being the preferred strategy. It is possible that doing this more often would introduce ambiguity; in the data, the majority of omissions of /ə/ indeed do not create ambiguity. The most frequent examples are shown in Table 18.

TABLE 18. The four most common words where /ə/ is represented by the absence of a vowel

Orthographic representation	⟨nhi⟩	⟨kr⟩	⟨rha⟩	⟨pr⟩
Phonemic form	[nəhĩ]	[kər]	[rəɦɑ]	[pər]
Meaning	'no'	'do'	'stay'	'on'
Occurrences	348	154	93	76

3.8. Nasal Vowels

Nasal vowels are used in Hindi-Urdu morphology for inflecting both nouns and verbs. In both cases, nasalization of the final vowel, or the addition of a nasalized vowel to the end of the word, serves to mark pluralization of a noun or of a verb conjugation. Harbour (2021) points out that many languages do not mark such morphological processes phonetically, and even those that do often don't mark them orthographically. Both the Hindi and Urdu orthographies do mark nasalization however, and both do so by adding a separate character in addition to the vowel. In Perso-Arabic, this character is derived from the representation of ⟨n⟩, whereas Devanagari has a separate character specifically to indicate nasalization.

Outside of morphological processes, Ohala (1983) and Masica (1991) agree that Hindi-Urdu has phonemic nasal vowels word-finally and word medially before voiceless consonants. However, even when expanding my dataset to all words, I identified only four examples of word-medial nasal vowels. These are listed in Table 19 along with their frequencies. Although it is difficult to draw conclusions on such a small sample size, ⟨n⟩ is indeed used to distinguish word-medial nasal vowels from oral vowels. Note that use of the ⟨n⟩ matches the convention used in the Urdu writing system.

Table 19. Occurrences of underlying word-medial nasal vowels in the fully expanded dataset

Orthography	⟨saanp⟩	⟨ansu⟩/⟨aansu⟩	⟨hanso⟩
Meaning	'snake'	'tears'	'laugh'
Phonetic form	[sə̃ːp]	[ə̃ːsu]	[ɦə̃so]
Occurrences	2	6	1

Because phonemic nasal vowels are relatively rare other than in morphological inflections, I expand the dataset to the top 4000 most frequent words in my data. These include eight words with phonemic nasal vowels: /nəɦ̃i/ 'no', /ɦə̃ː/ 'yes', /kjʊ̃/ 'why', /jəɦ̃i/ 'here', /jɑɦə̃ː/ 'here', /kɑɦə̃ː/ 'where', /d͡ʒɑɦə̃ː/ 'where', and /vɑɦə̃ː/ 'there'. Again, nasalization can be indicated by inserting ⟨n⟩ after the vowel; but the nasal vowels in these words are much more frequently written without marking nasalization. Table 20 shows the frequency counts of each of the possible forms.

Finally, I examine word-final nasal vowels that do result from inflectional processes. The most frequent of these processes is pluralization of both noun and verb forms, where a word-final oral vowel is nasalized, or a nasal vowel is added to a word-final consonant. In both cases, the plural form often forms a minimal pair with the singular form, which makes it

Table 20. Frequencies of the different representations of nasal vowels in the top 4000 words in the dataset, excluding nasal vowels resulting from inflectional processes. V represents a vowel grapheme.

Orthographic Form	V⟨n⟩	V
Frequency	289	1476

difficult to quantify cases where the nasalized form is not marked. Chief among these is [ɦẽː], the third person plural form of 'to be', which often cannot be disambiguated from the singular form [ɦeː] without examining each post containing them. Table 21 lists the representations of these words in the data with their frequencies—the forms ending in ⟨n⟩ are unambiguously plural.

Table 21. Representations of the third person 'to be' forms

Representation	⟨hai⟩	⟨hain⟩	⟨h⟩	⟨hen⟩	⟨hein⟩	⟨hn⟩
Phonetic forms	[ɦeː] or [ɦẽː]	[ɦẽː]	[ɦeː] or [ɦẽː]	[ɦẽː]	[ɦẽː]	[ɦẽː]
Occurrences	3285	411	349	20	17	14

I now turn my attention to the remainder of the inflected forms. Once again expanding to the 4000 most frequent words, the word [betɪjɑ̃] 'daughters' is written with a word-final ⟨n⟩ 7 times, and without a word-final ⟨n⟩ 8 times. For this word, there is no ambiguity between the singular and plural forms. However, I identified 295 other inflected forms with word-final nasal vowels, and all of them were written with nasalization marked. The most frequent of these are shown with their frequencies in Table 22.

Table 22. Common morphological inflections with word-final nasal vowels

Representation	⟨logon⟩	⟨doston⟩	⟨walon⟩
Phonetic form	[loːgõ]	[doːstõ]	[vɑːlõ]
Meaning	'people'	'friends'	'ones'
Occurrences	75	28	21

I conclude, then, that nasalization is marked when it is a result of grammatical inflection, but is not marked otherwise. This contradicts Harbour's (2021) finding which indicates that inflection is less likely to be marked orthographically. One possible reason is that nasalization is marked in both the Devanagari and Perso-Arabic writing systems, which

itself may have originated because of the existence of phonemic nasal vowels outside of grammatical inflection. When it comes to explaining why nasalization is often not marked, the most common instances of this are in very common words. Such words would be easier to recognize, and may also be more likely to have nonstandard forms develop for them which are easier to type. For instance, words which lack a nasalization marker are also more likely to lack some vowel representations, and some of these are presented with their frequency counts in Table 23.

TABLE 23. Miscellaneous abbreviated forms among words with nasal vowels

Representation	⟨nhi⟩	⟨nh⟩	⟨nhe⟩	⟨h⟩	⟨hn⟩
Phonetic forms	/nəɦĩː/	/nəɦĩː/	/nəɦĩː/	/ɦẽ/ or /ɦẽː/	/ɦẽː/
Meaning	'no'	'no'	'no'	'be'-3	'be'-3pl
Occurrences	348	11	7	349	14

I have analyzed each of the phonemes in Hindi-Urdu which were available in my data. I have identified the degree of linguistic fit in each phoneme's representations and have attempted to identify the origin of each representation.

4. Discussion

In this section, I begin with a summary of each phoneme's possible orthographic representations in my dataset. I then discuss the broader patterns that seem to occur throughout my data, and summarize the factors that appear to affect orthographic representation as well as to what degree linguistic fit, psychological fit, and technological fit appear to be maximized.

Table 24 lists the orthographic representations I have identified for each phoneme which I analyzed. The phonemes are listed in the order in which they appear in Section 3, and the representations for each phoneme are listed in decreasing order of frequency. Representations that are markedly rare are marked as "(marginal)".

TABLE 24. Correspondences between orthographic representations and phonemes

IPA	Roman	IPA	Roman
/p/	⟨p⟩	/n/	⟨n⟩
/b/	⟨b⟩	/ɳ/	⟨n⟩

Continued on next page

Table 24 – continued from previous page

IPA	Roman	IPA	Roman
/t̪/	⟨t⟩	/ɳ/	⟨n⟩
/d̪/	⟨d⟩	/ɲ/	⟨n⟩ or ⟨y⟩
/k/	⟨k⟩	/s/	⟨s⟩
/g/	⟨g⟩	/z/ (/d͡ʒ/ for some speakers)	⟨z⟩ or ⟨j⟩
/ʈ/	⟨t⟩	/ʃ/ (/s/ for some speakers)	⟨sh⟩
/ɖ/	⟨d⟩	/ʂ/	⟨sh⟩
/pʰ/	⟨ph⟩ or ⟨f⟩	/f/	⟨f⟩
/bʱ/	⟨bh⟩	/v/	⟨w⟩ or ⟨v⟩
/t̪ʰ/	⟨th⟩	/l/	⟨l⟩
/d̪ʱ/	⟨dh⟩	/ɾ/	⟨r⟩ or ⟨rr⟩
/kʰ/	⟨kh⟩	/ɽ/	⟨r⟩ or ⟨d⟩
/gʱ/	⟨gh⟩	/ɽʱ/	⟨rh⟩
/pː/	⟨pp⟩	/ɦ/	⟨h⟩
/bː/	⟨bb⟩	/j/	⟨y⟩
/t̪ː/	⟨tt⟩	/ɪ/	⟨i⟩ or ⟨e⟩ (marginal)
/d̪ː/	⟨dd⟩	/i/	⟨i⟩, ⟨ee⟩, ⟨e⟩ (marginal), ⟨ai⟩ (marginal), or ⟨ii⟩ (marginal)
/kː/	⟨kk⟩	/ɛ/	⟨ai⟩ or ⟨e⟩
/gː/	⟨gg⟩	/ɑi/	⟨ai⟩
/x/ ([kʰ] for some speakers)	⟨kh⟩	/oi/	⟨oi⟩ or ⟨oyi⟩ (marginal)
/ɣ/ ([g] for some speakers)	⟨g⟩ or ⟨gh⟩	/e/	⟨e⟩, ⟨ai⟩, ⟨i⟩, or ⟨ee⟩ (marginal)
/q/	⟨q⟩ or ⟨k⟩	/ʊ/	⟨u⟩
/d͡ʒ/	⟨j⟩	/u/	⟨u⟩, ⟨oo⟩, or ⟨o⟩ (marginal)
/d͡ʒʱ/	⟨jh⟩	/o/	⟨o⟩, ⟨ou⟩ (marginal), or ⟨oo⟩ (marginal)
/d͡ʒː/	⟨jj⟩	/ɔ/	⟨au⟩ or ⟨o⟩
/t͡ʃ/	⟨ch⟩	/ɑu/	⟨au⟩
/t͡ʃʰ/	⟨ch⟩ or ⟨chh⟩	/əu/	⟨au⟩
/t͡ʃː/	⟨ch⟩, ⟨cch⟩, ⟨chch⟩, or ⟨chh⟩	/ɑ/	⟨a⟩ or ⟨aa⟩
/t͡ʃʰː/	⟨chh⟩, ⟨cch⟩, ⟨chch⟩, or ⟨ch⟩	/ə/	⟨a⟩, , ⟨u⟩, ⟨e⟩, or ⟨i⟩ (marginal)
/m/	⟨m⟩	Ṽ	V or V⟨n⟩

In Table 25, I categorize each phonemic consonant's representations as having likely originated from either English orthography, historical Hindi-Urdu writing systems, colonial and early pedagogical literature, or the closest Roman script analogue according to users' auditory perception. Where applicable, typical orthographic conventions of English and other Roman scripts are generally followed. Orthographic representations are most consistent for sounds with analogues in English; these are /b ɖ g p ʈ k t͡ʃ d͡ʒ m n f s ʃ/. The written forms of /ɦ/ and /j/ also fall into this category, but word-final ⟨h⟩ and intervocalic ⟨y⟩ reflect similar Devanagari and Perso-Arabic conventions. The use of ⟨h⟩ as an aspiration marker also reflects the way aspiration is represented in Perso-Arabic. The use of ⟨kh gh q⟩ to represent /x ɣ q/ can be traced back to early transliterations and pedagogical literature. The representations of /r ɽ v/ have no such history, so their use may reflect users' auditory perception, reduction of ambiguity with other phonemes, or simply an arbitrary development among comparably plausible phonemes. Neither Hindi and Urdu keyboard layouts nor the visual shape of the letters appear to have an influence on the orthographic conventions I observe.

Table 25. Hypothesized origins of the orthographic representations for each phonemic consonant

English orthography	Devanagari and Perso-Arabic writing systems	Early literature	Perception of sounds without clear equivalents
/b ɖ g p ʈ k ɖ ʈ d͡ʒ t͡ʃ m n ɳ ŋ f s ʂ z l/	/ɦ j/	/x ɣ q/	/r ɽ v ɲ/

Linguistic fit, technological fit, and psychological fit all appear to play a role in predicting preferred representations of phonemes. Technological fit is maximized by avoiding any diacritics, and the dispreference for three-character and longer representations may reflect such representations being harder to process. As for the linguistic fit of the various orthographic representations, Table 26 summarizes these for the consonants. Geminate consonants and aspirated stops are omitted for clarity, but the mechanisms to represent aspiration and gemination also lead to one-to-one correspondences. In general, higher linguistic fit is preferred. There is no starker example of this than the Roman representations of /k/ and /t͡ʃ/: in English, /k/ is represented by both ⟨c⟩ and ⟨k⟩; and in both historical and modern Roman Hindi-Urdu publications, /t͡ʃ/ is represented by both ⟨ch⟩ and ⟨c⟩. In my data, however, ⟨c⟩ virtually never appears outside of ⟨ch⟩. /v/ provides another example of the possible tendency towards a one-to-one correspondence, where ⟨w⟩ is

more commonly used than ⟨v⟩ even though speakers perceive /v/ as [w] and [v] at approximately equal rates. The representations of /ɔ/ and /ɛ/ display this tendency as well.

Table 26. Linguistic fit of the Roman orthographic representations of Hindi-Urdu consonants

One-to-one correspondence	Phonemes represented in multiple ways	Orthographic representations used for multiple phonemes
/b d̪ g p t̪ k d͡ʒ t͡ʃ m ɳ ŋ f s ʃ l/	/v z t͡ʃʰ q x ɣ /	⟨d dd g kh r d t n y sh⟩

A few factors do appear to result in lower linguistic fit. These are summarized in Table 27. The first of these is the absence of a phoneme in English, or the lack of an obvious choice in Roman script. Hindi-Urdu /x ɣ q/ do not have phonemic analogues in English, and /v/ has two equally likely phonemic analogues according to auditory perception. This helps to explain the instability of the representations of each. The second possible factor is dialectal variation in how various phonemes are pronounced in speech. That /ɣ q z/ all have multiple orthographic representations which represent known phonological variation—and /x/ does not have multiple representations, because it is coincidentally represented identically to its alternative surface form /kʰ/—is evidence for this. Third, the lack of overlap in the phonetic environments of two graphemes seems to lessen the cost with regard to linguistic and psychological fit of using single graphemes to represent multiple phonemes. The retroflex consonants, which are generally written the same way as their non-retroflex counterparts, are an example of this, where the convenience of a more intuitive representation may override the disruption of linguistic and psychological fit. /ŋ/ and /ɲ/ exhibit the same pattern. Fifth, linguistic fit seems to be harder to maintain for longer orthographic representations. /t͡ʃː/, /t͡ʃʰ/, and /t͡ʃʰː/, all of which are often represented by three and four character combinations, display more variation in orthographic form than those geminates and aspirated stops which are written with one- and two-character Roman representations.

The high degree of variation in phonetic realization of vowels makes it difficult to determine whether differences in representation signify dialectal differences or differences in orthographic convention. Additionally, several Hindi-Urdu vowels may have phonetic realizations that resemble other vowels (Schmidt, 2003), which further complicates the analysis. The biggest difference I observe in vowels is that in multiple cases, the orthography does not take advantage of every potential orthographic representation to maximize linguistic fit. The lax/tense dis-

Table 27. Considerations that lead linguistic fit to decrease

Consideration	Example phonemes
Lack of an obvious choice in Roman script	/x ɣ q v ɾ ɽ/
Dialectal variation	/p x ɣ q z/
Rarity of phonemes	/ʈ ɖ ʂ ɲ ɋ/
Perceptual similarity to other phonemes	/ʈ ɖ ɾ ɽ/
Length of the orthographic representation	/t͡ʃ: t͡ʃʰ/

tinction is generally not reflected in the orthography for high vowels, despite the availability of ⟨ee⟩ and ⟨oo⟩; yet, ⟨au⟩ and ⟨ai⟩ are used in order to make the lax/tense distinction visible for mid vowels. This may reflect differences in the salience of that contrast for different vowel heights. Finally, the multitude of possible orthographic representations of /ə/ may reflect differences in the perception of that vowel in different phonetic environments, as well as the apparent tendency not to omit vowels in Roman orthography.

I have shown that Roman Hindi-Urdu reflects phonological variation between speakers, and some of these cases of variation are similar to cases of sociocultural variation found in spoken language, such as the use of phonemes loaned from Perso-Arabic (ibid.). I do not have the demographic data to determine whether any variation is the result of sociocultural factors, but it is plausible that users could intentionally choose from among these forms to identify themselves with or distance themselves from regional or religious groups associated with particular forms. The representations of /q/, /x/, /ɣ/, /z/, and /p/ all could reflect this phenomenon. Another potential way of signaling sociocultural identity in the Roman script is to carry over written conventions from the Urdu and Hindi orthographies, since these are already associated with speakers' sociocultural identity. Orthographic representations of vowels are one arena in which this could appear. Further work which takes users' demographic information into account can test these hypotheses directly and also may shed light on other sources of sociocultural variation.

5. Future Directions

Many future studies could expand upon this work. With regard to determining the orthographic representations of each phoneme, the rarer phonemes could be studied much more thoroughly given more data. Additionally, more data could allow for more detailed study of how phonetic and orthographic environments might affect the representation of

phonemes; for instance, how particular consonant clusters are written, or whether the representations of vowels such as /ə/ depend on their phonetic environments.

Several findings in this study indicate that psychological fit affects the representations of some phonemes, and psycholinguistic studies could test whether graphemes which represent multiple phonemes or long representations take longer to read and write, and whether the rarity of a phoneme affects processing speed. Similarly, much work remains to be done on sociocultural fit: future work could compare the orthographic conventions used by self-identified Urdu and Hindi speakers, or of internet users in India and Pakistan. Finally, surveys could be conducted to determine whether users have a conscious philosophy for how to write Hindi-Urdu in the Roman script, as well as to learn about users' intuitions as to which representations look 'correct' to them.

6. Conclusion

This study uses data from X to document the orthographic conventions of Hindi-Urdu and discusses the possible factors which have shaped these conventions in the absence of prescriptive standardization. The Roman script has become one of the main ways of using Hindi-Urdu in written form, and this study can help us understand how it is used. Linguistic fit, historical orthographic conventions, and users' auditory perception of phonemes all appear to help shape the orthographic conventions of Roman Hindi-Urdu, and they all might be expected to play a role in other organically developing writing systems. Finally, I have discussed some of the questions which remain to be answered and how future work may address them.

References

Ahmad, R. (2011). "Urdu in Devanagari: Shifting orthographic practices and Muslim identity in Delhi". In: *Language in Society* 40.3, pp. 259–284.

Akbari, M. (2013). "A preliminary linguistic analysis of Romanized Persian SMS messages". In: *Journal of Novel Applied Sciences* 2.8, pp. 197–205.

Androutsopoulos, J. (2009). "'Greeklish': Transliteration practice and discourse in the context of computer-mediated digraphia". In: *Standard Languages and Language Standards: Greek, Past and Present*. Ed. by A. Georgakopoulou and M. Silk. Farnam: Ashgate.

Atta, A. (2021). "Scripts on linguistic landscapes: A marker of hybrid identity in urban areas of Pakistan". af. In: *Journal of Nusantara Studies (JONUS)* 6.2, pp. 58–96.

Bali, K. et al. (2014). ""I am borrowing *ya* mixing?": an analysis of English-Hſindi code mixing in Facebook". In: *Proceedings of the first workshop on computational approaches to code switching*, pp. 116–126.

Center for Language Engineering (n.d.). *Urdu 5000 Most Frequently Used Words List*. `https : / / cle . org . pk / software / ling _ resources / UrduHighFreqWords.htm`.

Cole, J. et al. (2010). "Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach". In: *Journal of phonetics* 38.2, pp. 167–184.

Faruqi, S. R. (2003). "A Long History of Urdu Literary Culture, Part I". In: *Literary cultures in history: Reconstructions from South Asia*. Ed. by S. Pollock. University of California Press, pp. 805–63.

Fatima, S. (2011). "Phonological Variation in Perso-Arabic Words in Urdu". af. In: *Language in India* 11.7.

Gilchrist, J. B. (1803). *Polyglot fables*. printed at the Hurkaru Office.

Grover, V. (2016). "Perception and Production of /v/ and /w/ in Hindi speakers". PhD thesis. City University of New York.

Harbour, Daniel (2021). "Grammar Drives Writing System Evolution". In: *Grapholinguistics in the 21st Century 2020. Proceedings: Grapholinguistics and Its Applications*. Ed. by Y. Haralambous. Fluxus Editions, pp. 201–221.

Kachru, Y. (2006). *Hindi*. John Benjamins Publishing.

Kanoongo, U. (2016). "Sociolinguistic Functions of Roman-Romanagari Code-switching in WhatsApp Instant Messaging". In: *The Journal of Rase* 14.28, p. 14.

Khan, M. Ahmad (2000). *Urdu phonology*. Dept. of Linguistics, Aligarh Muslim University.

Khurshid, K., S. A. Usman, and N. Javaid (2003). "Possibility of Existence and Identification of Diphthongs and Triphthongs in Urdu Language". In: *Crulp Annual Student Report*. Center for Language Engineering.

Mangrio, R. A. and G. Ali (2014). "Rise and Fall of Arabic Loan Phonemes in Urdu". In: *Literary Globalization and the Politics and Poetics of Translations* 15.

Masica, C. P. (1991). *The Indo-Aryan languages*. Cambridge University Press.

Meletis, D. (2018). "What is natural in writing?: Prolegomena to a Natural Grapholinguistics". In: *Written Language & Literacy* 21.1, pp. 52–88.

Meletis, D. and C. Dürscheid (2022). *Writing Systems and Their Use*. De Gruyter Mouton.

Meyer, C. M. and I. Gurevych (2012). "Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography". In: *Electronic Lexicography*. Ed. by Granger and Paquot. Oxford University Press, pp. 259–291.

Mishra, D. and K. Bali (2011). "A Comparative Phonological Study of the Dialects of Hindi". In: *ICPhS* 17, pp. 17–21.

Mouresioti, Evgenia and Marina Terkourafi (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". In: *Journal of Greek Linguistics* 21.2, pp. 224–262.

Mukherjee, J. (2010). "The development of the English language in India." In: *The Routledge Handbook of World Englishes*. Ed. by A. Kirkpatrick. Routledge, pp. 167–180.

National Assembly of Pakistan (1973). *The Constitution of the Islamic Republic of Pakistan*. https://na.gov.pk/uploads/documents/1333523681_951.pdf.

Nema, N. and J. K. Chawla (2018). "The Dialectics of Hinglish: A Perspective". In: *Applied Linguistics Papers* 25.2), pp. 37–51.

Noruzi, A. (2023). "Wiktionary Popularity from a Citation Analysis Point of View". In: *Informology* 2.1, pp. 1–6.

Oberlies, T. (2005). *A historical grammar of Hindi*. Leykam.

Ohala, M. (1983). *Aspects of Hindi phonology*. Motilal Banarsidass Publishers Pvt. Limited.

———— (1994). "Hindi". In: *Journal of the International Phonetic Association* 24.1, pp. 35–38.

Oxford Dictionary Press (1893). "Crore". In: *Oxford English Dictionary*. https://doi.org/10.1093/OED/8116865719. DOI: 10.1093/OED/8116865719.

Pandey, Pramod Kumar (1991). "Schwa fronting in Hindi". In: *Studies in the Linguistic Sciences* 21.1.

Parliament of India (1963). *The Official Languages Act*. en. https://www.indiacode.nic.in/bitstream/123456789/1526/1/A1963__19.pdf.

Pierrehumbert, J. and R. Nair (1996). "Implications of Hindi prosodic structure. Current trends in phonology". In: *Models and methods* 2, pp. 549–584.

Putra, R. A. and S. Triyono (2018). "Outlandish spelling system invented by Indonesian internet society: The case of language usage in social networking site". In: *International Journal of Applied Linguistics and English Literature* 7.7, pp. 66–73.

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.

———— (2020). "Pakistani English". In: *The handbook of Asian englishes*. Ed. by K. Bolton, W. Botha, and A. Kirkpatrick. Wiley-Blackwell, pp. 279–296.

Ramanathan (2016). "English Education Policy in India". In: *English Language Education Policy in Asia*. Ed. by R. Kirkpatrick. Springer.

Rao, C. et al. (2011). "Orthographic characteristics speed Hindi word naming but slow Urdu naming: Evidence from Hindi/Urdu biliterates". In: *Reading and Writing* 24.6, pp. 679–695.

Rosowsky, A. (2010). "'Writing it in English': script choices among young multilingual muslims in the UK". In: *Journal of Multilingual and Multicultural Development* 31.2, pp. 163–179.
Salomon, R. (2003). "Writing systems of the Indo-Aryan languages". In: *The Indo-Aryan languages*. Ed. by G. Cardona and D. Jain. Routledge, pp. 75–114.
Schmidt, R. L. (2003). "Urdu". In: *The Indo-Aryan languages*. Ed. by G. Cardona and D. Jain. Routledge, pp. 286–350.
Sebba, M. (2007). "Orthography as social practice". In: *Spelling and Society*. Ed. by M. Sebba. Cambridge University Press, pp. 26–57.
Shapiro, M. C. (2003). "Hindi". In: *The Indo-Aryan languages*. Ed. by G. Cardona and D. Jain. Routledge, pp. 250–285.
Sharma, B. D. (2018). "Vowel Phonemes in Hindi". In: *East European Journal of Psycholinguistics* 5.2, pp. 71–91.
Sharma, R. N. (1937). *Roman Urdu: A comprehensive study in Hindustani*. R.N. Sharma.
Speer, R. (2022). *rspeer/wordfreq: v3.0 (v3.0.2)*. Version v3.0.2. DOI: 10.5281/zenodo.7199437.
Unseth, P. (2005). "Sociolinguistic parallels between choosing scripts and languages". In: *Written language & literacy* 8.1, pp. 19–42.
Wiktionary (2023). *Wiktionary, the free dictionary*. URL: https://en.wiktionary.org/wiki/Wiktionary:Main_Page.
Zaidi, Syeda Bushra and Sajida Zaki (2017). "English language in Pakistan: Expansion and resulting implications". In: *Journal of Education & Social Sciences* 5.1, pp. 52–67.