



A Comparative Analysis of Phonetic Representation: Pashto Script in Contrast with Farsi, Urdu, and Arabic

Mohammad Qaseem Kashaf^{1*} & Abdul Jabbar Sapand²

¹Department of Pashto, Faculty of Languages and Literature, Nangarhar University, Jalalabad, Afghanistan.

²Department of Pashto, Faculty of Languages and Literature, Kabul Education University, Kabul, Afghanistan.

ARTICLE	DATES	CITE THIS ARTICLE: A Comparative Analysis of Phonetic Representation: Pashto Script in
INFO	27-09-2026	Contrast with Farsi, Urdu, and Arabic. (n.d.). Nangarhar University Journal of Arts and
	06-10-2026	Humanities, 2(01), 1-5. https://doi.org/10.70436/nujah.v2i01.33

ABSTRACT

The adaptation of the Perso-Arabic script across multiple linguistic traditions has resulted in varying degrees of phonetic adequacy and orthographic stability. This study presents a comparative analysis of grapheme-phoneme correspondence in Pashto, Farsi, Urdu, and Arabic, with particular emphasis on the structural limitations of the script when applied to Pashto. While Farsi and Urdu demonstrate relatively successful orthographic innovations through the addition of new graphemes and partial standardization, Pashto exhibits a significantly higher degree of phonetic ambiguity and under-specification. Using a comparative-analytical methodology, this paper evaluates consonantal and vocalic representation across the four languages. The findings reveal that Pashto's complex phonemic inventory, including retroflex fricatives, a dense sibilant system, and a central vowel, is inadequately represented within the inherited script framework. This results in increased reliance on contextual interpretation and optional diacritics. The study concludes that Pashto represents a critical case for understanding the functional limits of the Perso-Arabic script, with important implications for literacy, language standardization, and computational processing.

KEYWORDS: Comparative, Grapheme-phoneme, Linguistics, Orthography, Phonetic-representation, Script.

Introduction

Writing systems are not neutral tools but historically shaped frameworks that reflect the phonological needs of the languages for which they were originally designed. The Arabic script, developed as a consonantal abjad, was highly efficient for representing the phonology of Arabic. However, its adaptation to non-Semitic languages such as Farsi, Urdu, and Pashto introduced varying degrees of mismatch between script and sound system.

Farsi and Urdu adapted the script through relatively systematic innovations, including the addition of new letters to represent non-Arabic sounds. These modifications allowed both languages to achieve a workable balance between phonetic accuracy and orthographic stability. In contrast, Pashto presents a more complex case. Its phonological system includes features that exceed the representational capacity of the Perso-Arabic script, resulting in inconsistencies and ambiguities that affect both reading and writing.

This study examines the extent to which the Perso-Arabic script accommodates the phonetic structure of Pashto in comparison with Farsi, Urdu, and Arabic. It focuses on identifying areas of orthographic strain and evaluating the effectiveness of adaptation strategies across these languages. It further discussed the effective orthographic strategies used in Pashto compared to those employed in Farsi and Urdu.

This research contributes to both theoretical and applied linguistics. From a theoretical perspective, it advances the understanding of orthographic adaptation and the concept of orthographic depth. From a practical perspective, it highlights challenges in literacy acquisition, language teaching, and computational processing of Pashto. The findings are particularly relevant for educators, linguists, and developers working on natural language processing tools for under-resourced languages.

*Corresponding Author: (Kashaf M. Q)



Email: qaseem_kashaf@nu.edu.af



<https://doi.org/10.70436/nujah.v2i01.33>



© 2024 The Authors. Published by Nangarhar University, this is an open access article published under the CC-BY license <https://creativecommons.org/licenses/by/4.0>

[Daniels, P. T., & Bright, W. \(1996\)](#) Theories about the origins and development of writing systems have emerged alongside the systems themselves. According to [Coulmas \(2003\)](#) Several languages, including Korean, Mongolian, Persian, Romanian, Turkish, and Vietnamese, have been represented using different writing systems over time, often reflecting shifts in political affiliations and alignments. The study of writing systems has long emphasized the relationship between orthographic structure and phonological representation. One of the central concepts in this field is grapheme-phoneme correspondence, which refers to the systematic relationship between written symbols and spoken sounds. According to [Aronoff \(1994\)](#) writing systems are not purely phonetic, but are shaped by historical, morphological, and functional constraints. This perspective is essential for understanding the adaptation of the Perso-Arabic script across multiple languages.

Another influential concept is the Orthographic Depth Hypothesis, developed by [Frith \(1985\)](#) and later expanded by [Katz and Frost \(1992\)](#). This hypothesis distinguishes between shallow orthographies, where there is a direct correspondence between letters and sounds, and deep orthographies, where this relationship is more complex and context-dependent. Languages such as Spanish are considered shallow, while English is categorized as deep. Within this framework, the Perso-Arabic script tends to produce deeper orthographies when applied to non-Semitic languages.

The Arabic script itself has been extensively studied as a consonantal writing system, or Abjad, which primarily represents consonants and long vowels while leaving short vowels unmarked in most contexts. Ryding (2005) notes that this system functions efficiently for Arabic due to its root-based morphology, where consonantal patterns carry the core semantic meaning. However, when this script is transferred to languages with different phonological structures, its limitations become more apparent.

The adaptation of the Arabic script to Persian represents one of the earliest and most successful cases of orthographic modification. [Windfuhr \(2009\)](#) and Perry (2005) highlight how Persian introduced new graphemes, including چ, گ, and پ, to represent sounds absent in Arabic. At the same time, Persian simplified the system by merging several Arabic phonemes that were not contrastive in Persian. This combination of innovation and simplification resulted in a relatively stable orthographic system ([Hock, 2016](#)).

Urdu represents a further stage in the adaptation of the Perso-Arabic script, particularly within the Indo-Aryan linguistic context. According to Shackle (2003) Urdu developed additional letters such as ڈ, ٲ, and ڙ to represent retroflex consonants, a prominent feature of South Asian phonology. These innovations demonstrate a systematic approach to orthographic expansion, allowing Urdu to maintain a relatively consistent grapheme-phoneme correspondence despite its complex sound system ([Hock, 2016](#)).

In contrast, research on Pashto orthography reveals a more problematic adaptation. Early descriptive work by [Penzl \(1955\)](#) and later by [MacKenzie \(1987\)](#) documented the phonological richness of Pashto, including its retroflex consonants, sibilant distinctions, and complex vowel system. However, these studies also noted inconsistencies in how these sounds are represented in writing.

More recently [Bashir \(2013\)](#) highlights the role of dialectal variation in exacerbating orthographic ambiguity, particularly in the pronunciation of letters such as ښ and ږ. These variations make it difficult to establish a single standardized mapping between graphemes and phonemes.

Despite these contributions, the existing literature reveals a significant gap. Most studies focus on describing Pashto phonology or orthography independently, without conducting a systematic comparison with related languages such as Farsi and Urdu. Furthermore, limited attention has been given to the practical implications of orthographic ambiguity for literacy and computational processing.

This study addresses these gaps by providing a comparative analysis grounded in the theoretical frameworks of grapheme-phoneme correspondence and orthographic depth. It extends previous research by examining not only the structural features of Pashto orthography but also its functional consequences.

Methods and Materials

This study adopts a comparative-analytical approach. Data were collected from standard orthographies, linguistic descriptions, and representative texts in Pashto, Farsi, Urdu, and Arabic. The analysis focuses on two main dimensions: consonantal and vocalic representation.

A comparative matrix was developed to map phonemes to their corresponding graphemes in each language. The analysis evaluates transparency, consistency, and completeness of grapheme-phoneme correspondence. Case studies were also used to illustrate specific instances of ambiguity in Pashto.

Results and Discussion

The analysis reveals clear differences in the degree of orthographic adaptation among the four languages. Farsi demonstrates a high level of stability, with a limited number of added letters and manageable vowel ambiguity. Urdu shows greater complexity but compensates through a well-developed system of diacritics and additional graphemes.

Table 1. Phonetic Representative of all four languages.

Alphabet	Pashto IPA	Farsi IPA	Urdu IPA	Arabic IPA
ا (Hamzah)	/a/	/a/	/ə/, /ɪ/,	/a/
ا (Alif):	/ā/	/ā/	/ɑ/, /ʊ/	/a/
ب (Bay):	/b/	/b/	/b/	/b/
پ (Pay):	/p/	/p/	/p/	
ت (Tay):	/t/	/T/	/t̪/	/th/
ٹ (Tay):	/t̪/		/t̪/	/t̪/
ث (Say / Thay):	/s/	/s/	/s/	/s/
ج (Jeem):	/j/	/J/	/dʒ/	/j/
چ (Chay):	/č/	/ʻc/	/tʃ/	/č/
غ (tsay):	/c/	/c/	/c/	/c/
ح (Hay):	/h/	/ʻh/	/h/	/h/
خ (Khay):	/x/	/x/	/x/	/Kh/
د (Daal):	/d/	/d/	/d̪/	/d/
ډ (Daal):	/d̪/		/d̪/	/d̪/
ذ (Zaal):	/z/	/z/	/z/	/Dh/
ر (Ray):	/r/	/r/	/r/	/r/
ز (Zay):	/z/	/z/	/z/	/z/
ځ (Zhay):	/ž/	/ʻz/	/ʒ/	/z./
ږ (Ray):	/r̪/		/r̪/	/r̪/
ځ (Gae):	/ğ/		/ğ/	/ğ/
س (Seen):	/s/	/s/	/s/	/s/
ش (Sheen):	/š/	/ʻs/	/ʃ/	/sh/
ښ (khean):	/s/			/ ʃ /
ص (Swaad):	/s/	/s/	/s/	/ʃ/

ض (Zwad):	/z/	/z/	/z/	/ḏ/
ط (Toay):	/t/	/t/	/t/	/ṭ/
ظ (Zoay):	/z/	/z/	/z/	/ẓ/
ع (Ain):	/a-ʕ/	/ʕ/	/a/, /ɾ/	‘
غ (Ghain):	/g-ʕ/	/g/	/ʕ/	/gh/
ف (Fay):	/f/	/f/	/f/	/f/
ق (Qaaf):	/q/	/q/	/q/	/q/
ک (Kaaf):	/k/	/k/	/k/	/k/
گ (Gaaf):	/g/	/g/	/g/	/g/
ل (Laam):	/l/	/L/	/l/	/l/
م (Meem):	/m/	/m/	/m/	/m/
ن (Noon):	/n/	/n/	/n/	/n/
ښ (Noondh):	/ṇ/		/ṇ/	
و (Wao):	/o/	/o/	/o/	
و (خرگند و او):	/u/	/ú/	/u/	
و (نیمواک و او):	/w/	/w/	/w/	/w/
ه / ه (Narma/ Do-Chashmi Hay):	/h/	/h/	/h/	/h/
ی (نیمواک ی)	/y/	/y/		
ی (narma Ye):	/ay/		/ay/	
ې (ugda Ye):	/e/		/e/	
ي (landa Ye):	/i/	/ i /	/i/	/y/
ئ/ی (zwrakiwala/amria Ye):	/əy/		/əy/	

Note: [Ziar \(2005\)](#), [Yamin \(2018\)](#), [Bruce \(2021\)](#), and [Awde \(2000\)](#).

According to Table 1. The phonetic representation of all four language almost same in many alphabets but in some alphabets the phonetic representation is different. It is notable, that there are some alphabets that exist in some languages and some languages do not have it as seen in the above table.

Pashto, however, exhibits significant challenges. Consonantal representation is marked by ambiguity, particularly in the case of retroflex fricatives such as ښ and ږ, which correspond to multiple phonetic values depending on dialect. The sibilant system is also under-differentiated, leading to overlap between sounds.

Vowel representation presents an even greater problem. The central vowel is represented only by an optional diacritic, which is typically omitted in writing. This results in widespread ambiguity, as multiple pronunciations can correspond to the same written form.

The findings confirm that Pashto orthography is significantly deeper than that of Farsi and Urdu. This depth is characterized by a lack of consistent grapheme-phoneme mapping and a heavy reliance on contextual interpretation.

Unlike Farsi and Urdu, which introduced systematic solutions to address phonological gaps, Pashto relies on less consistent strategies. The use of modified letters and diacritics does not fully resolve ambiguity, particularly in the case of vowels.

This has important implications for literacy. Readers must rely on prior knowledge and context to interpret written text, increasing cognitive load. In education, this complicates the teaching of reading and writing, as learners cannot depend solely on phonetic decoding.

In computational contexts, these issues create significant challenges for text processing, speech recognition, and machine translation. The lack of explicit phonetic information in the script limits the effectiveness of automated systems.

Conclusion

This study demonstrates that while the Perso-Arabic script has been successfully adapted for Farsi and Urdu, it faces significant limitations when applied to Pashto. The language's complex phonemic inventory exceeds the representational capacity of the script, resulting in a high degree of orthographic ambiguity.

The analysis highlights the need for more systematic approaches to Pashto orthography, particularly in the representation of vowels and retroflex consonants. Addressing these challenges is essential for improving literacy, standardization, and technological development.

References

- Aronoff, M. (1994). *Morphology by itself: Stems and inflectional classes*. MIT Press.
- Awde, N. (2000). *The Arabic alphabet: How to read and write it*. Lyle Stuart.
- Bashir, E. (2013). Wakhi. In *The Iranian languages* (pp. 825-862). Routledge.
- Bruce, G. M. (2021). *Urdu vocabulary: a workbook for intermediate and advanced students*. Edinburgh University Press.
- Coulmas, F. (2003). *Writing systems: An introduction to their linguistic analysis*. Cambridge University Press.
- Daniels, P. T., & Bright, W. (1996). *The world's writing systems*. Oxford University Press.
- Frith, U. (1985). Beneath the surface of developmental dyslexia. In *Surface dyslexia*. Routledge.
- Hock, H. H., & Bashir, E. (2016). The languages and linguistics of South Asia. *A Comprehensive Guide*. Berlin & Boston: De Gruyter Mouton.
- Katz, L., & Frost, R. (1992). The reading process is different for different orthographies. In *Orthography, phonology, morphology, and meaning*. Elsevier.
- MacKenzie, D. N. (1987). *A standard Pashto*. Oxford University Press.
- Penzl, H. (1955). *A grammar of Pashto*. American Council of Learned Societies.
- Seymour, P. H. K. (2005). Early reading development in European orthographies. *British Journal of Psychology*.
- Windfuhr, G. (2009). *The Iranian languages*. Routledge.
- Yamin, M. H. (2018). *Contemporary Grammar of Fari-Dari Language*. Banga Publication. Kabul.
- Ziar, M.A. (2005). *Grammar*. Danish Publishing Association, Jalalabad.