The Ambiguity of the Relations between Graphemes and Phonemes in the Persian Orthographic System

Tayebeh Mosavi Miangah[1]
Relja Vulanović[2]

Abstract

In this paper, the degree to which Persian orthography deviates from transparency is quantified and evaluated. We investigate the relations between graphemes and phonemes in Persian, in which the writing system is not fully representative of the spoken language, mostly due to the omission of the short-vowel graphemes. We measure the degree of the Persian orthographic system transparency using a heuristic mathematical model. We apply the same measures to orthographic systems of other languages and compare the results to those obtained for Persian. The results show a relatively high degree of transparency in Persian when it comes to writing, but a low degree of transparency when it comes to reading. We also consider models that avoid the problems related to the short vowels in Persian and these models demonstrate a considerable decrease of the uncertainty in the Persian orthographic system.

Keywords: *Persian language, orthographic system, orthographic uncertainty, phonemic uncertainty.*

1. Introduction

There is a principle in linguistics known as the one-meaning-one-form principle, based on which forms and meanings in any language system tend to correspond one to one (Anttila 1972, p. 181). The more a language obeys this principle, the more transparent, or the less opaque/uncertain/ambiguous it is. The concepts of transparency and opaqueness/uncertainty/ambiguity are not the same as the simplicity and complexity of languages. Hengeveld and Leufkens point out that "[l]anguages may be complex yet transparent, or simple, yet opaque." (Hengeveld & Leufkens 2018, 141). Yet, it is argued in Vulanović (2007) that language complexity should be measured relative to language transparency (if an increase in complexity is not accompanied by a decrease in transparency, the relative complexity of the language is even greater).

The notions of transparency and opaqueness can be extended to any pair of related linguistic units that do not necessarily involve meaning. In this paper, we consider the transparency of the relationship between phonemes and graphemes. More specifically, this paper aims to investigate the transparency of the Persian orthographic system using mathematical models. Our approach is similar to that found in Best & Altmann (2005) and the contributions to the volume *Analyses of script: Properties of characters and writing systems* (Altmann & Fan 2008). However, these works are mainly concerned with the *phoneme-to-grapheme map* and provide measures of the orthographic uncertainty of all phonemes in various languages. As opposed to this, we examine both directions of the relationship between graphemes and phonemes in the written and spoken forms of modern Persian and measure the degree of the overall Persian orthographic system transparency. In addition to the information-theory measure used in Altmann & Fan (2008), we also propose and use a new, slightly simpler one, which is akin to the measures of the degree of violation of the one-meaning-one-form principle (Vulanović & Ruff 2018, Vulanović & Mosavi Miangah 2020).

[1] Department of Linguistics, Payame Noor University, Tehran, Iran. E-mail: mosavit@pnu.ac.ir. Work done while visiting Kent State University at Stark. http://orcid.org/0000-0002-6528-2876

[2] Department of Mathematical Sciences, Kent State University at Stark, 6000 Frank Ave NW, North Canton, Ohio 44720, USA. E-mail: rvulanov@kent.edu. Corresponding author. http://orcid.org/0000-0003-2189-5133

If each grapheme corresponds to one and only one phoneme, and vice versa, we can say that the orthography of the given language is transparent. On the other hand, when such one-to-one correspondence does not exist, we deal with phonological and/or orthographic uncertainties. As languages differ from one another by the extent to which they are transparent or opaque, we compare the degree of transparency of the orthographic systems in different languages to locate the relative position of the Persian language on a scale.

In the next section, we provide an overview of some fundamental concepts and describe Persian orthography and phonology in general and to the extent that is needed for the present study. Then, in section 3, we survey the relevant literature. This is followed by the introduction of the mathematical models in section 4 and their application in section 5, where the results of the calculations are presented. Concluding remarks are given in section 6.

2. Preliminaries

2.1 Some fundamental concepts

The specific terms we use are as follows.

- Phoneme: a mental representation of a speech sound made by the mouth, which distinguishes one word from another in a particular language. We indicate phonemes by placing them between two forward slashes, / /.
- Letter: a visual building block of written words (the way a word looks, not the way it sounds).
- Grapheme: an individual letter or groups of letters that represent a single phoneme. A grapheme is a written symbol expressing a sound. When we need to emphasize that we are dealing with graphemes, we place them inside angle brackets, < >.

Consider, for example, the word "elephant" in which the grapheme <ph> consists of two letters, <p> and <h>, representing the phoneme /f/.

2.2. Persian orthography and phonology

Persian is the official language of Iran. The writing system of this language has been adopted from the Arabic script with some modifications, although the spoken form is very different from Arabic. Persian is written from right to left and most letters have to be joined to their adjacent letters according to their position in the word. Although there is no distinction between capital and small letters, most letters have more than one shape depending on their position in the word, known as initial, medial, and final shapes. There are 36 letters in Persian, out of which 10 letters can only be written joined to the preceding letter (آ - د - ر - ز - ذ - و - ژ - ؤ- أ - ا), not to the following one.

The Persian phonemic system has 24 consonants and 6 vowels, three long and three short ones.[3] The three long vowels (i, u, A) are realized in written form and the three short vowels (e, o, a) are usually not written, except in two special cases when the phonemes /o/ and /e/ are indicated using the letters و and ه, respectively. The letter و is mainly used to represent a consonant phoneme, as well as a long-vowel phoneme, and ه is mainly used to represent a consonant phoneme. Moreover, superscript and subscript diacritics exist that can be used along with the consonant letters to indicate the short vowels. When added to a letter, the diacritics make the pronunciation of that letter different from the original one. Table 1[4] illustrates the role

[3] The issue of whether Persian distinguish between the vowel lengths or the distinction between the vowels is solely based on their quality is of no relevance to this paper. Some general references to Persian phonology are Majidi (1986/1990) and Windfuhr (1997).

[4] In Table 1 and throughout the paper, we use the SAMPA (Speech Assessment Methods Phonetic Alphabet) phonetics for Arabic (Wells 2002) with some modifications adjusted to Persian. The complete list of Persian characters along with their nearest English equivalents is presented in Appendix I.

of the diacritical marks in Persian and the way they make words different in terms of pronunciation and meaning.

Table 1.
Possible pronunciations of the written word کرم.

No.	Graphemes	Corresponding Phonemes	Pronunciation	Meaning	Graphemes with diacritics
1	کرم	/krm/	[karam]	generosity	کَرَم
2	کرم	/krm/	[kerm]	worm	کِرم
3	کرم	/krm/	[korom]	chrome	کُرُم
4	کرم	/krm/	[kerem]	cream	کِرِم

As Table 1 shows, a string of several consonant graphemes may have several different pronunciations and meanings depending on the type and the location of the short vowels, which are normally not written, in the graphemic string. As a result, phonological and semantic ambiguities arise, specifically for children and others learning Persian for the first time. Such ambiguities partly result from the existence of many Arabic and western loanwords in Persian (e.g., Nos. 1, 3, and 4 in Table 1).

Although short vowels are not normally realized in writing, they are written in two special circumstances. The first one is texts used for and by the beginner learners of Persian (including native Persian-speaking children in the early grades of school). The second case is the religious texts which are almost all borrowed from Arabic. In both cases, it is much easier to read the texts when the short-vowel diacritics are written. However, a major problem arises when the diacritics are omitted and inexperienced Persian readers are required to read the text correctly and understand its meaning. This problem also frequently happens to foreigners learning Persian as their second or foreign language when they try to pronounce written Persian words correctly and get the appropriate meaning. Still, the question is what happens to adult native or matured Persian speakers to be able to read the texts without diacritics and understand their meanings despite the phonological and semantic ambiguities created by the homographs. The answer to this question lies in the fact that after being exposed to texts with and without diacritics in the early years of Persian learning, Persian speakers reach a kind of cognitive maturity based on which reading texts without diacritics is largely possible. In other words, the early year textual materials act as a training corpus with the help of which the readers can develop their reading abilities through the strategies they have previously used. In most cases, their visual ability gained while learning Persian helps them to understand new occurrences of the words using patterns similar to those they have been exposed to earlier. It goes without saying that for the cases which have more than one correct pronunciation with different meanings, adult readers refer to the context to which the given word belongs to decide on the appropriate pronunciation and meaning.

There are also some additional discrepancies between phonemes and graphemes in Persian. An individual grapheme can represent several different phonemes and, similarly, an individual phoneme can be represented by several different graphemes. Tables 2 and 3 depict all possible relationships that exist between these two linguistic units in Persian. For complete lists of the graphemes and phonemes of Persian, see Appendices II and III.

Table 2.
The possible number of phonemes for individual graphemes in Persian (with diacritics)

Graphemes	**No. of phonemes**	**Examples**
1- آ (A) 2- او (‘u) 3- ؤ (?) 4- ئـ (?) 5- أ (?) 6- ایـ (i) 7- ای (ei)	1	1- آبادان (AbAdAn = a city name) 2- اوست (‘ust = s/he is) 3- مسؤول (mas?ul = responsible) 4- رئیس (re?is = boss) 5- هیأت (hei?at = committee) 6- ایران (‘irAn = Iran) 7- ای خدا (ei xodA = Oh God)
وا	2	واکسن (vAksan = vaccine) خواهر (xAhar = sister)
ا , ب , د , م , س , ت , ر , ن , ز , ش , ک , پ , گ , ف , خ , ق , ل , ج , ح , چ , ژ , ص , ع , ث , ض , ط , غ , ظ , ذ ,	4	سال (sAl = year) سَفیر (safir = ambassador) سُلطان (soltAn = king) سِفید (sefid = white) (Example only provided for the grapheme س.)
1- ی (i, y, ya, ye, yo) 2- ه (h, e, he, ha, ho)	5	1- سینی (sini = tray), یاقوت (yAqut = ruby), یَزد (yazd = a city name), یِک (yek = one), یُد(yod = iodine) 2- مهتاب (mahtAb = moonlight), خانه (xAne = home), هِل (hel = cardamom), هَمیشه (hamiSe = always), هُجوم (hojum = rush)

Apart from the short vowels, the long vowels also show some inconsistencies in the Persian orthography. In Table 4, which is a selected extraction from Appendix II, we list the graphemes for the three long vowels of Persian, as those occurring exclusively in initial positions and those appearing in both initial and non-initial positions (each long vowel has a different form when appearing in the initial position). As Table 4 shows, there are some inconsistencies between long-vowel graphemes and phonemes, specifically when the long vowels appear in middle or finial positions in the word. In the initial position, the different form of the long-vowel grapheme almost always helps to identify the corresponding phoneme unambiguously. To be concrete, the grapheme <آ>, the other variation of <ا> appearing only in the initial position in the word, has just one possible pronunciation, /A/. The grapheme <ایـ>, the other variation of <یـ> appearing only in the initial position in the word, has two possible pronunciations as /i/ and /ei/; the latter one can only be seen in the word ای (an interjection word which means "Oh!"). And lastly, the grapheme <او>, the other variation of <و > appearing only in the initial position in the word, has a very limited possible set of pronunciations which can be distinguished in a small number of words such as اَوِستا, اِوِرِست and a couple of other words that can be easily memorized. Therefore, it is not very realistic to mention these initial forms as ambiguous graphemes.

Table 3.
The possible number of graphemes for individual phonemes in Persian (without diacritics)

Phonemes	No. of graphemes	Examples
Phonemes other than the ones listed below	1	نمک (namak=salt)
1- /o/ 2- /i/ 3- /t/ 4- /h/ 5- /u/	2	1- استخوان (ostexAn=bone), خوردن (xordan=eating) 2- ایشان (iSAn = they), بینی (bini = nose) 3- طوفان (tufAn = storm), توپ (tup = ball) 4- هندوانه (hendevAne = watermelon), حیوان (heivAn = animal) 5- اورژانس (urZAns = emergent), سوخت (suxt = fuel)
1- /s/ 2- /A/	3	1- صورت (surat = face), سفید (sefid = white), ثمن (saman = price) 2- مادر (mAdar = mother), آفتاب (AftAb = sunlight), خواندن (xAndan = reading)
/z/	4	زیبا (zibA = nice), ذرت (zorat = corn), ظهر (zohr = noon), ضمن (zemn = while)
/?/	5	سؤال (so?Al = question), مسئله (mas?ale = problem), قلعه (qal?e = castle), امضاء (emzA? = signature), مأوا (ma?vA = residence)

Table 4.
Possible corresponding pronunciations of Persian long vowels

Long-vowel graphemes (initial position)	Possible phonemes	Examples
یـ (ایـ)	/i/, /y/, /ya/, /ye/, /yo/	مینا , دستیار , یواش , یگانه , یمن ,ایستاد
و (او)	/u/, /o/, /v/, /va/, /ve/, /vo/	جوراب , خورشید , دشوار دعوت , وصال , وضو , اوست
ا (آ)	/A/, /a/, /e/, /o/	سال , اسیر , امتحان , اسوه , آدم

So, in this paper, we want to calculate the degree of the Persian orthographic ambiguity from the standpoint of both matured and non-matured Persian users. The ambiguity is mainly due to the absence of the short-vowel diacritics from the Persian writing system, which makes the conversion of graphemes to phonemes, that is, reading, difficult. If we look at the Persian orthography as already mastered by adult native speakers, we find the grapheme-to-phoneme relation much more transparent. The same happens if we determine the degree of ambiguity of the Persian orthography by considering texts that use all the diacritics overtly. The other direction, the conversion of phonemes to graphemes, that is, writing, is not a big problem in Persian and we find the Persian orthography more transparent than the orthography of any other language in our sample. These results can be found in section 5 below.

3. Related works

There is a related line of work that has grown around the Persian orthographic issues. Of note is a thread of works in this vein by Baluch (Baluch & Besner 1991; Baluch & Shahidi 1991; Baluch 1993, 2005). We would like to mention two separate investigations on the reading of individual Persian words (naming) by children and adult learners. In the first research, Baluch and Shahidi studied the naming of Persian words with consonantal spelling, as opposed to those with vowel letters, by children with the mean age of 8.4 years. Their findings revealed that children made significantly more errors when dealing with opaque words (like /bCh/, meaning *child*) than with transparent words (like /bAzi/, meaning *play*). Consequently, the time taken to name a list of words with consonantal spelling was shown to be longer than the time taken to name a list of words with vowel-letter spelling (Baluch and Shahidi 1991). Baluch also reported similar findings while performing the same experiment on skilled adult Persian readers. When consonantal words had multiple meanings, their naming times were significantly longer than in the case of consonantal words with a unique meaning. He concluded from there that there was significant difficulty in naming words with consonantal spelling, caused by phonological processes (Baluch 1993).

In an investigation dealing with the Persian spoken in present-day Iran and the relationship between Persian orthography and literacy, Baluch has attempted to emphasize how literacy acquisition by Persian beginner or skilled readers may be affected by peculiarities of Persian orthography. He claimed it was the first time the issue of cognitive processes involved in literacy acquisition of Persian was reviewed. After an extensive elaboration on orthographic and phonological inconsistencies of Persian, he put forward that the main orthographic and phonological factors possibly affecting Persian literacy are the grapheme-phoneme regularity, the phoneme-grapheme ambiguity, and the absence of short vowels in written text. Finally, he points to the fact that perhaps some changes should be introduced into Persian orthography (Baluch 2005).

Another related work is the research reported by Kaveh Ashourinia in his Master's Dissertation. He attempted to quantify and visualize the ambiguities of the semi-consonantal Persian writing system with a glance at its consequences. He introduced an analytic approach as a tool to examine the difficulties resulting from the lack of short vowels in written Persian and the inconsistency of some long vowels with the written form. He concluded that these analytical data support the idea that the Persian writing system is not well-suited for the Persian language (Ashourinia 2019).

However, some other researchers completely ignored the problem of diacritics in Persian and categorized this language as a very transparent language in comparison to other languages. Gholamain & Geva, for instance, use the term "orthographic depth," previously used by Baluch (1993), Baluch & Besner (1991), and Frost, Katz & Bentin (1987), to categorize orthographic systems on a continuum ranging from shallow to deep. They claim that Persian (like some other languages such as Turkish, Hebrew, and Arabic) can be labeled as "shallow" because it has a simple grapheme-phoneme relationship in comparison to other scripts, e.g., the English script, which they labeled as "deep" for its more complicated grapheme-phoneme relations. They argue that learning Persian and mastering to decode Persian words and reading Persian texts accurately for children in the early grades is much easier than it might be the case for reading English. They implicitly conclude that such regularity may have an impact on learning to read, referring to other investigations (Gholamain & Geva 1999). As we have mentioned in subsection 2.2, when we consider Persian texts with all short vowels overt in the script (vowelized script), the grapheme-to-phoneme relationship becomes sufficiently regular.

The only point that is absent in the above literature on Persian orthography is how to measure the degree of orthographic uncertainty in Persian so that it can be compared to other languages. Our work is set to measure for the first time the orthographic ambiguity of Persian by heuristic and robust mathematical formulas. In this way, we can concretely calculate the degree of deviation of the Persian orthographic system from the ideal transparency and compare it to orthographic systems in other languages through an objective criterion. Although this has not been done for Persian before, there are many investigations on quantifying the relationship between phonemes and graphemes in other languages. In addition to Altmann & Fan (2008), which we have already discussed in the introduction, we can mention Best & Altmann (2005). A review of ways to measure orthographic transparency/uncertainty, and of other related issues, is given in Borleffs, Maassen, Lyytinen & Zwarts (2017).

4. Mathematical description

Let us first introduce some mathematical notation. For any set A, let $|A|$ denote the number of elements in the set and let A^* be the set of all strings of elements in A including the empty string denoted by λ. For two nonempty finite sets X and Y, we define $X \times Y$ as the set of ordered pairs, $X \times Y = \{(x, y) : x \in X,\ y \in Y\}$. A relation Φ between X and Y is a subset of $X \times Y$. The corresponding inverse relation is denoted by Φ^{-1}, $\Phi^{-1} = \{(y, x) : (x, y) \in \Phi\}$. Let n_x indicate how many elements of Y are paired up in the relation Φ with a particular $x \in X$,

$$n_x = n_x(\Phi) = |\{y \in Y : (x, y) \in \Phi\}|.$$

We assume of any relation Φ that it involves every element of both X and Y, so that for each $x \in X$ there exists an element $y \in Y$ such that $(x, y) \in \Phi$, and vice versa. This implies that $n_x \geq 1$ for each $x \in X$.

We next define the frequency f_k of elements in X that have $n_x = k$,

$$f_k = f_k(\Phi) = |\{x \in X : n_x(\Phi) = k\}|,\ \ k = 1,2,3, \dots$$

Note that $0 \leq f_1 \leq |X|$. If $f_1 = |X|$, that is, if $f_k = 0$ for all $k \geq 2$, then the relation Φ is a function from X onto Y. To measure how far a relation Φ is from a function, we can use the formula

$$m(\Phi) = \frac{S(\Phi)}{|X|},\ \ S(\Phi) = \sum_{k \geq 2} f_k(\Phi)(k-1). \qquad (1)$$

In a function, each $x \in X$ is paired with exactly one $y \in Y$, so the above sum $S(\Phi)$ indicates the number of pairs in Φ that goes over the count which would be present in a function. If this count is 0, that is, if $m(\Phi) = 0$, then (and only then) Φ is a function. The sum $S(\Phi)$ is divided by the total number of elements in X, which makes $m(\phi)$ a relative measure. The reason for this is illustrated by the following simple abstract example.

Example. Consider $X = \{1\}$, $Y = \{a_1, a_2\}$, and $\Phi = \{(1, a_1), (1, a_2)\}$. We have $S(\phi) = 1$ and $m(\Phi) = 1$. Now, add 98 more pairs $(2, a_3), (3, a_4), \dots, (99, a_{100})$ into Φ to create a new relation Φ'. This new relation is not a function because of only one of the 100 pairs in it, whereas Φ is not a function because of one of the two pairs. Nevertheless, the sum $S(\Phi)$ shows no difference between Φ and Φ' since $S(\Phi')$ is still equal to 1. On the other hand, $m(\Phi') = \frac{1}{99} = 0.0101$, thus the relative measure $m(\Phi)$ places Φ and Φ' in more appropriate positions on a scale indicating for any relation how far it is from a function.

If Φ^{-1} is also a function, then Φ is a bijection (one-to-one correspondence) between X and Y. A measure of how far a relation is from a bijection is introduced in (Vulanović & Ruff 2018) and applied to linguistics. Further modifications and applications of this measure can be found in Vulanović & Mosavi Miangah (2020). As we are about to see, bijections are not suited for the analysis of phoneme and grapheme systems, so we cannot apply the formulas from Vulanović & Ruff (2018) or Vulanović & Mosavi Miangah (2020) in this paper. Nevertheless, there is some similarity between the formulas used in these two works and the formula in (1).

Let P be the set of all phonemes and G the set of all graphemes of a language. As Appendix III indicates in the case of Persian, the relation between the phonemes and the graphemes cannot be represented as a subset of $P \times G$, but rather as a subset of $P \times \tilde{G}$, where $\tilde{G} \subset G^*$. This is because $\tilde{G}$ includes the empty grapheme λ, which otherwise would not be considered an element of G. Kelih (2008) describes the situation in the Slovene language in a similar way. He uses the grapheme λ to present the phoneme-grapheme relation, and, although he does not consider the inverse relation, in some other analyses of the grapheme system, he does not include λ as a grapheme. Similarly, Appendix II shows that it is easier to describe the grapheme-phoneme relation in Persian not by referring to a subset of $G \times P$, but to a subset of $G \times \tilde{P}$, where $\tilde{P} \subset P^*$. This is because strings of phonemes need to be used.

Let the phoneme-grapheme relation (also called the *phoneme-to-grapheme map*), as a subset of $P \times \tilde{G}$, be denoted by Φ_1 and let Φ_2 stand for the grapheme-phoneme relation (or the *grapheme-to-phoneme map*), which is a subset of $G \times \tilde{P}$. Then, the measure $m_1 \coloneqq m(\Phi_1)$ evaluates the orthographic uncertainty of all phonemes, while $m_2 \coloneqq m(\Phi_2)$ measures the phonemic uncertainty of all graphemes. Let also $p_k = f_k(\Phi_1)$ and $g_k = f_k(\Phi_2)$. Then, according to the formulas in (1),

$$m_1 = \frac{1}{|P|}\sum_{k\geq 2} p_k(k-1), \quad m_2 = \frac{1}{|G|}\sum_{k\geq 2} g_k(k-1). \tag{2}$$

The greater the value of m_1, the harder it is to write. If $m_1 = 0$, this indicates the easiest writing system in terms of knowing what grapheme to use for any given phoneme. This is so because each phoneme has exactly one graphemic representation (although one grapheme may be used to represent more than one phoneme). Similarly, the greater the value of m_2, the harder it is to read. An orthography that enables the easiest reading has $m_2 = 0$ because each grapheme represents exactly one phoneme (although several different graphemes may be used for the same phoneme). When the values of m_1 and m_2 are close, this indicates an orthography in which it is approximately equally easy to write and to read. If m_1 is considerably greater than m_2, writing is harder than reading, and the other way around.

Typically, $\Phi_1^{-1} \neq \Phi_2$. Therefore, it is not appropriate in the present context to ask how far a relation is from a bijection and the measures from Vulanović & Ruff (2018) and Vulanović & Mosavi Miangah (2020) cannot be used. Nevertheless, we can still measure the uncertainty of the whole system of phonemes and graphemes by averaging the two measures given in (2),

$$m = \frac{1}{2}(m_1 + m_2). \tag{3}$$

Most of the works on the systems of phonemes and graphemes in various languages are focused on the relation Φ_1 and do not consider Φ_2 (Best & Altmann 2005, Altmann & Fan 2008). The orthographic uncertainty of phonemes is measured in these works not by m_1 but by the quantity U_1,

$$U_1 = \frac{1}{|P|}\sum_{k\geq 1} f_k(\Phi_1)\log_2 k = \frac{1}{|P|}\sum_{k\geq 2} p_k \log_2 k. \tag{4}$$

By the way, this is the correct mathematical rendering of the formula in Best & Altmann (2005) and Altmann & Fan (2008), which instead of the above sum uses $\sum_{x\in P} f_x \log_2 n_x$ although the frequency f_x does not depend on a single phoneme x.

The measure U_1 is based on information theory. The corresponding weighted measure of the orthographic uncertainty of a single phoneme reduces to entropy (Best & Altmann 2005, Borleffs et al. 2017). We shall apply both measures m_1 in (2) and U_1 in (4) to all the languages in our sample and show that there is a strong correlation between them. Therefore, either measure can be used to arrive at the same general conclusions and m_1 is a little simpler because it does not require the calculations of binary logarithms.

At the same time, it is possible to measure the phonemic uncertainty of graphemes by a quantity U_2 which is defined analogously to U_1 in (4),

$$U_2 = \frac{1}{|G|} \sum_{k \geq 2} g_k \log_2 k. \tag{5}$$

Then, the uncertainty of the whole system of phonemes and graphemes can also be measured by

$$U = \frac{1}{2}(U_1 + U_2). \tag{6}$$

We shall only be able to compare the corresponding measure m_2 in (2) and U_2 in (5), as well as m in (1) and U in (6), in the case of Persian and Greek.

5. Results

5.1. Persian

The table in Appendix III shows the 30 phonemes in Persian. We present in Table 5 the number f_k of phonemes that have k graphemic representations.

Table 5.
The number p_k of Persian phonemes
that have k graphemic representations

$\boldsymbol{k}$	1	2	3	4	5
$\boldsymbol{p_k}$	19	5	4	1	1

Based on this table, we calculate the orthographic uncertainty of phonemes in Persian using the measure m_1 given in (2),

$$m_1 = \frac{5 \cdot 1 + 4 \cdot 2 + 1 \cdot 3 + 1 \cdot 4}{30} = \frac{20}{30} = 0.6667.$$

The other formula, (4), yields

$$U_1 = \frac{5 \cdot 1 + 4 \log_2 3 + 1 \cdot 2 + 1 \cdot \log_2 5}{30} = 0.5221.$$

Because U_1 uses logarithms, this measure produces values that are less than those of m_1.

We now consider the phonemic uncertainty of Persian graphemes. Table 6 is derived from Appendix II.

Table 6.
The number g_k of Persian graphemes
that have k phonemic representations

$\boldsymbol{k}$	1	2	3	4	5	6
$\boldsymbol{g_k}$	7	1	0	29	2	1

We calculate the value of m_2 using the formula in (2),

$$m_2 = \frac{1 \cdot 1 + 29 \cdot 3 + 2 \cdot 4 + 1 \cdot 5}{40} = \frac{101}{40} = 2.525.$$

Also, from formula (5), we get

$$U_2 = \frac{1 \cdot 1 + 29 \cdot 2 + 2 \log_2 5 + 1 \cdot \log_2 6}{40} = 1.6557.$$

Comparing these values to the above-calculated m_1 and U_1, we can see that it is much more difficult to read than to write Persian. This is so because the three short-vowel phonemes, /e/, /a/, and /o/, are not written after the consonants. It should be mentioned that we only consider the basic phoneme-to-grapheme and grapheme-to-phoneme maps in Persian. Otherwise, going into details, like in the discussion related to Table 4, can make the maps much more complicated, see also Mohseni Behbahani, Babaali & Turdalyuly (2016).

We can also calculate the overall measures m and U using the formulas (3) and (6), respectively,

$$m = \frac{0.6667 + 2.525}{2} = 1.5959, \quad U = \frac{0.5221 + 1.6557}{2} = 1.0889.$$

Let us now consider the situation when three different diacritics are used to indicate the three short vowels after the consonants. This increases the readability of Persian drastically. Table 5 does not change because the empty grapheme λ in Appendix III is replaced with the corresponding short-vowel diacritic. Therefore, the values of m_1 and U_1 remain the same. However, m_2 and U_2 become much smaller. This is because Table 6 changes to Table 7.

Table 7.
The number g_k of Persian graphemes,
including the diacritics to indicate the three short vowels,
that have k phonemic representations

k	1	2	3	4
g_k	38	3	1	1

This gives

$$m_2 = \frac{3 \cdot 1 + 1 \cdot 2 + 1 \cdot 3}{43} = 0.186, \quad U_2 = \frac{3 \cdot 1 + 1 \cdot \log_2 3 + 1 \cdot 2}{43} = 0.1531.$$

and

$$m = \frac{0.6333 + 0.186}{2} = 0.4097, \quad U = \frac{0.3732 + 0.1531}{2} = 0.2632.$$

Obviously, the inclusion of the short-vowel diacritics turns the situation in Persian completely around, making its orthography easier for reading than for writing. The same would happen if we treated each consonant grapheme in Appendix II as only representing the corresponding consonant phoneme, thus ignoring the short-vowel phonemes after the consonants. We mention this approach to the analysis of Persian orthography because it is followed in Frost et al. (1987), but short vowels after the consonants should not be ignored because they are phonemes. In this case, m_1 and U_1 would still be the same as above, but the values of m_2 and U_2 would become

$$m_2 = \frac{3 \cdot 1 + 1 \cdot 2 + 1 \cdot 3}{40} = 0.2, \quad U_2 = \frac{3 \cdot 1 + 1 \cdot \log_2 3 + 1 \cdot 2}{40} = 0.1646,$$

which would give

$$m = \frac{0.6333 + 0.2}{2} = 0.4167, \quad U = \frac{0.3732 + 0.1646}{2} = 0.2689.$$

Therefore, when the short vowels are ignored in this way, the ambiguity of the Persian orthography is reduced.

Each of the two cases considered above, viz., the inclusion of the short-vowel diacritics and the ignoring of the short-vowel phonemes after consonants, simulates a matured Persian user who has no or very little difficulty reading Persian.

5.2. A comparison of Persian and Greek

Another paper that analyzes both phoneme-to-grapheme and grapheme-to-phoneme maps is Protopapas & Vlahou (2009). The analysis is applied to the Greek language. In this subsection, we do all the calculations like in 4.1 using the Greek data from (ibid.). Table 8 summarizes the counts.

Table 8.
The numbers p_k and g_k in the Greek orthography

k	1	2	3	4	5	6	7	8	9	10	11
p_k	7	14	4	4	3	0	2	2	0	0	1
g_k	63	14	4	1	1	1	0	0	0	0	0

When counting the Greek phonemes to find the values of p_k, we made one slight modification of the original data. Protopapas & Vlahou (2009) consider the phoneme strings /k//s/ and /p//s/ as single phonemes /ks/ and /ps/ because they have unique graphemic representations <ξ> and <ψ>, respectively. We did not follow this approach since orthography is not a factor in determining phonemes in a language. Therefore, we did not treat /ks/ and /ps/ as single phonemes. Instead, we added <ξ> to the list of graphemes representing /k/ and also to those representing /s/. Similarly, <ψ> was added to both list of graphemes representing /p/ and /s/. Nemcová & Altmann (2008) did the same with the Slovak grapheme <x>, which they included in each list of graphemes representing the phonemes /k/, /g/, /s/, and /z/.

Based on Table 8, we have for Greek that

$$m_1 = 2.2162, \quad U_1 = 1.3616, \quad m_2 = 0.4048, \quad U_2 = 0.3244.$$

Since m_1 is considerably greater than m_2 (the same relation holds between U_1 and U_2), we see that Greek is much easier to read than to write, as opposed to the situation we find in Persian. Using the above values, we can also calculate the overall measures m and U. They are given in Table 9 together with the results for the three cases of Persian analysis done in the previous subsection. Although Table 9 only contains four pairs of values of m and U, it can be used to verify the correlation between the two measures. The coefficient of correlation is $R = 0.9988$, so the correlation between m and U is nearly perfect.

Table 9.
The overall uncertainty of the Persian and Greek Orthographies

Language	m	U
Persian (without the short-vowel diacritics)	1.5833	1.0764
Persian with the short-vowel diacritics	0.4097	0.2632
Persian with short vowels after consonants ignored	0.4167	0.2689
Greek	1.3105	0.8430

Table 10.
The orthographic uncertainty of phonemes across languages

Language	m_1	U_1
Greek	2.2162	1.3616
German	1.1795	0.9661
Swedish	1.0556	0.7983
Slovak	0.9318	0.7599
Slovene	0.9310	0.7847
Oriya	0.9167	0.8475
Italian	0.6949	0.5648
Persian	0.6667	0.5221

5.3. Other languages

Six other languages, in addition to Persian and Greek, are considered here. They are German and Swedish (Best & Altmann 2005), Slovak (Nemcová & Altmann 2008), Slovene (Kelih 2008), Oriya (Mohanty & Altmann 2008), and Italian (Bernhard & Altmann 2008). However, for these six languages, we only have phoneme-grapheme relations, and because of this, we only calculate m_1 and U_1. Without changing any of the original data, we got the values presented in Table 10. The results for U_1 were also calculated in the original works. The values we show are essentially the same, but we carried out the calculations with more decimal places. The languages are listed in Table 10 in the decreasing order of the m_1 values. We see that, of the eight languages, Greek is the hardest one to write, whereas Italian and Persian are the easiest ones. Comparisons like this should be taken with a degree of reservation because the eight phoneme-to-grapheme maps are not necessarily given with the same amount of detail. As mentioned before, we here only consider the basic map for Persian. The correlation between the m_1 and U_1 values is very strong, $R = 0.9620$.

6. Conclusion and implications

In this paper, we have quantitatively described the extent to which the Persian orthography deviates from an ideally transparent orthography. Many languages across the world have different types of inconsistencies in their phonological and orthographic systems due to various factors, but the Persian language has its peculiar problems that make the written form of the language rather hard to read especially for beginner learners. The Persian graphemes are basically taken from Arabic resulting in inconsistencies with the phonological system of Persian. As we have demonstrated, the main problem of Persian orthography is that the short vowels are absent in writing, which leads to ambiguities when reading Persian texts. We have also shown that the ambiguity of the Persian orthographic system becomes much lower when the short vowels are ignored in the pronunciation of Persian words, or when the diacritical marks indicating the short vowels are fully implemented.

We have not investigated the distribution of phonemes or graphemes here, but this also plays a role in the transparency of Persian orthography. The more restricted the distribution, the easier it is to disambiguate phonemes or graphemes in context. This can be analyzed in a continuation of the present study.

Keeping the Persian orthographic system with such ambiguities may have not only national, but also global consequences. As Hengeveld and Leufkens suggest, transparency is an important factor in the learnability of languages, and transparent features of a language are the first to be mastered by young children acquiring their mother tongue (Hengeveld & Leufkens 2018). Thus, regarding literacy acquisition, the inconsistencies of Persian orthography give rise to many problems for both Iranian children learning to read and write Persian in early grades of school and for foreigners trying to learn Persian as a foreign or second language. From the computational linguistics point of view, such ambiguities have serious effects on the quality of Text-To-Speech (TTS) systems, spell checking systems (especially in the second and third phases—generating and ranking candidates—of the whole process, Mosavi Miangah 2014), and the like. These are some arguments indicating that a major reform of the Persian script may be needed, a reform that would render the Persian script much easier to read, not only by the inclusion of the short-vowel graphemes, but also by making the graphemes more uniform (Appendix III shows different written forms for single-consonant phonemes). Nickjoo also points out that the peculiarities of written Persian have implications for literacy. He argues for the abolition of the Persian alphabet and the creation of a Latinized version of Persian (Nickjoo 1979).

The existing semi-Arabic script of Persian cannot answer all the needs of modern life, especially in the digital environment. It is hoped that this paper will stimulate further investigations in the field and motivate appropriate measures towards forward-looking decisions regarding the Persian language and its future challenges.

References

Anttila, R. (1972). *An introduction to historical and comparative linguistics*. New York: Macmillan.

Altmann, G., Fan F. (eds.) (2008). *Analyses of script: Properties of characters and writing systems*, Berlin/New York: Mouton de Gruyter.

Ashourinia, K. (2019). *Quantifying and visualizing the ambiguities of the semi-consonantal Persian writing system, and its consequences*, Master's Dissertation, OCAD University, Toronto, Ontario, Canada.

Baluch, B., Besner, D. (1991). Visual word recognition: Evidence for strategic control of lexical and non-lexical routines in oral reading. *Journal of Experimental Psychology, Learning, Memory and Cognition* 17, 644–651.

Baluch, B., Shahidi, S. (1991). Visual word recognition in beginning readers of Persian. *Perceptual and Motor Skills* 72, 1327–1331.

Baluch, B. (1993). Lexical decisions in Persian: A test of the orthographic depth hypothesis. *International Journal of Psychology* 28, 19–27.

Baluch, B. (2005). *Handbook of orthography and literacy*. Routledge Handbooks Online. Cutting edge scholarship from Routledge and CRC Press.

Bernhard, G., Altmann, G. (2008). The phoneme-grapheme relationship in Italian. In Altmann, G., Fan F. (eds.), 13–24.

Best, K.-H., Altmann, G. (2005). Some properties of graphemic systems. *Glottometrics* 9, 29–39.

Borleffs, E., Ben A. M. Maassen, Ben A. M., Lyytinen, H., Frans Z. (2017). Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and Writing* 30, 1617–1638.

Frost, R., Katz, L., Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance* 13, 104–115.

Gholamain, M., Geva, E. (1999). Orthographic and cognitive factors in the concurrent development of basic reading skills in English and Persian. *Language Learning* 49, 183–217.

Kelih, E. (2008). The phoneme-grapheme relationship in Slovene. In Altmann, G. & Fan F. (eds.), 61–74.

Hengeveld, K., Leufkens, St. (2018). Transparent and non-transparent languages. *Folia Linguistica* 52, 139–175.

Majidi, M.-R. (1986/1990). *Strukturelle Grammatik des Neupersischen (Fārsi). Bd. I: Phonologie; Bd. II: Morphologie. Forum Phoneticum* 34, 1–2. Hamburg: Buske.

Mohanty, P., Altmann, G (2008). On graphemic representation of the Oriya phonemes. In Altmann, G., Fan F. (eds.), 121–140.

Mohseni Behbahani, Y., Babaali, B., Turdalyuly, M. (2016). Persian sentences to phoneme sequences conversion based on recurrent neural networks. *Open Comput. Sci.* 6, 219–225.

Mosavi Miangah, T. (2014). FarsiSpell: A spell-checking system for Persian using a large monolingual corpus. *Literary and Linguistic Computing* 29, 56–73.

Nemcová, E., Altmann, G. (2008). The phoneme-grapheme relation in Slovak. In Altmann, G., Fan F. (eds.), 79–90.

Nickjoo, M. (1979). A century of struggle for the reform of the Persian script. *The Reading Teacher*, 926–929.

Protopapas, A., Vlahou, E. L. (2009). A comparative quantitative analysis of Greek orthographic transparency. *Behavior Research Methods* 41, 991–1008.

Vulanović, R. (2007). On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics* 20, 399–427.

Vulanović, R., Ruff, O. (2018). Measuring the degree of violation of the One-Meaning–One-Form Principle. In Wang, L., Köhler, R., Tuzzi, A. (Eds.), *Structure, function and process in texts*, 67–77. Lüdenscheid: RAM.

Vulanović, R., Mosavi Miangah, T. (2020). The flexibility of parts-of-speech systems and their grammar efficiency. In Kelih, E., Köhler. R. (eds.), *Words and Numbers* (*In Memory of Peter Grzybek*), 129–147. Lüdenscheid: RAM.

Wells, J. C. (2002). SAMPA for Arabic. http://www.phon.ucl.ac.uk/home/sampa/arabic.htm, accessed 01.02.2021.

Windfuhr, G. (1997). Persian phonology. In Kaye, A. S. (ed.) *Phonologies of Asia and Africa (Including the Caucasus)*. Vol 2, 675–689. Winona Lake, Indiana: Eisenbrauns.

Appendix I

Transcription standards of Persian used in this paper.

Persian character	SAMPA character	Persian example	SAMPA transliteration	English translation
آ	A	آسیب	Asib	damage
ا ـا	A	گرما	garmA	warmth
اَ	‘a	اَدیان	‘adyAn	religions
اِ	‘e	اِرتباطاتي	‘ertebAtAti	communicative
اُ	‘o	اُردك	‘ordak	duck
ب بـ	B	باادب	bAadab	polite
پ پـ	P	پتو	patu	blanket
ت تـ	T	تابستان	tAbestAn	summer
ث ثـ	s	ثابت	sAbet	fixed
ج جـ	j	جمله	jomle	sentence
چ چـ	C	چرا	CerA	why
ح حـ	h	حیاط	hayAt	yard
خ خـ	X	خانواده	xAnevAde	family
د	D	دوست	dust	friend
ذ	Z	ذرت	zorrat	corn
ر	R	روز	ruz	day
ز	Z	زانو	zAnu	knee
ژ	Z	ژرف	Zarf	profound
س سـ	S	سفید	sefid	white
ش شـ	S	شناگر	SenAgar	swimmer
ص صـ	s	صیاد	sayyAd	hunter
ض ضـ	z	ضروري	zaruri	necessary
ط	t	طاووس	tAvus	peacock
ظ	z	ظهر	zohr	noon
ع عـ ـعـ ـع	‘	عجیب	‘ajib	strange
غ غـ ـغـ ـغ	Q	غایب	QAyeb	absent
ف فـ	f	فردي	fardi	personal
ق قـ	q	قدیمي	qadimi	ancient
ک کـ	k	کثیف	kasif	dirty
گ گـ	g	گزارشگر	gozAreSgar	reporter
ل لـ	l	لیوان	livAn	glass
م مـ	m	مرطوب	martub	wet
ن نـ	n	نکته	nokte	point
ه ـه ـهـ هـ	h	همپایه	hampAye	coordinate
و	v	وظیفه	vazife	task
یـ	y	یخچال	yaxCAl	refrigerator
و	u	مربوط	marbut	related
و	o	خود	xod	Self

ـُ	o	مُدام	modAm	forever
ـَ	a	مَدرك	madrak	document
ـِ	e	مِهربان	mehrabAn	Kind
ـی ـیـ ی	i	شهری	Sahri	urban
ؤ	‘	مؤسس	mo‘asses	founder
ئـ أ ـأ ئ	‘	ارائه	‘erA‘e	presentation

Appendix II
Persian graphemes with their possible corresponding phonemes.

	Grapheme	**Phonemes** **(how the grapheme can be read)**
1	آ	/A/
2	ا	/A/, /e/, /a/, /o/
3	ب	/b/, /be/, /ba/, /bo/
4	د	/d/, /de/, /da/, /do/
5	م	/m/, /me/, /ma/, /mo/
6	س	/s/, /se/, /sa/, /so/
7	او	/u/
8	ت	/t/, /te/, /ta/, /to/
9	ر	/r/, /re/, /ra/, /ro/
10	ن	/n/, /ne/, /na/, /no/
11	ای	/i/
12	ی	/i/ , /y/, /ye/, /ya/, /yo/
13	ز	/z/, /ze/, /za/, /zo/
14	ه	/e/ , /h/, /he/, /ha/, /ho/
15	ش	/S/, /Se/, /Sa/, /So/
16	ک	/k/, /ke/, /ka/, /ko/
17	و	/o/ , /u/ , /v/, /ve/, /va/, /vo/
18	پ	/p/, /pe/, /pa/, /po/
19	گ	/g/, /ge/, /ga/, /go/
20	ف	/f/, /fe/, /fa/, /fo/
21	خ	/x/, /xe/, /xa/, /xo/
22	ق	/q/, /qe/, /qa/, /qo/
23	ل	/l/, /le/, /la/, /lo/
24	ج	/j/, /je/, /ja/, /jo/
25	ح	/h/, /he/, /ha/, /ho/
26	چ	/C/, /Ce/, /Ca/, /Co/
27	ژ	/Z/, /Ze/, /Za/, /Zo/
28	ص	/s/, /se/, /sa/, /so/
29	ع	/?/, /?e/, /?a/, /?o/
30	ث	/s/, /se/, /sa/, /so/
31	ض	/z/, /ze/, /za/, /zo/

32	ط	/t/, /te/, /ta/, /to/
33	غ	/Q/, /Qe/, /Qa/, /Qo/
34	ظ	/z/, /ze/, /za/, /zo/
35	ذ	/z/, /ze/, /za/, /zo/
36	وا	/A/, /vA/
37	ؤ	/?/
38	ئ	/?/
39	أ	/?/
40	ء	/?/

Appendix III

Persian phonemes with their possible correspondent graphemes
(λ denotes an empty grapheme).

	Phoneme	Graphemes
1	/A/	آ , وا , ا
2	/b/	ب
3	/d/	د
4	/m/	م
5	/s/	ص,ث , س
6	/u/	او , و
7	/t/	ت , ط
8	/r/	ر
9	/n/	ن
10	/i/	ای , ی
11	/y/	ی
12	/z/	ز , ض , ذ , ظ
13	/h/	ه , ح
14	/S/	ش
15	/k/	ک
16	/v/	و
17	/p/	پ
18	/g/	گ
19	/f/	ف
20	/x/	خ
21	/q/	ق
22	/l/	ل
23	/j/	ج
24	/C/	چ
25	/Z/	ژ
26	/?/	ع , ء , ؤ , ئ ,أ

27	/Q/	غ
28	/a/	ا, λ
29	/e/	ه, ا , λ
30	/o/	و,ا, λ