A Romanized Transcription for Persian

Jalal Maleki
Natural Language Processing Laboratory
Department of Computer and Information Science
Linköping University
SE-581 83 Linköping
Sweden
Email: jma@ida.liu.se

Abstract

This paper presents, Dabire, a romanized transcription scheme which is based on the phonology and morphophonology of Persian. Dabire uses an extended Latin alphabet and a number of conventions with the aim of providing a simple, consistent and easy to learn writing system.

1 Introduction

This paper presents, Dabire, a romanization scheme for Persian. The motivation for working with this proposal is twofold. As well as facilitating linguistic analysis and processing, it will provide an alternative writing system for Persian speakers that are not familiar with the Arabic-based script.

Currently, Persian is predominantly written in variations of the Arabic writing system. The official writing system of Iran, for example, is the Perso-Arabic Script (PA-Script) [6]. In recent years, however, an increasing body of romanized Persian text has appeared on the Internet and in the context of mobile communication. For the majority of Persian speakers, who are well-acquainted with PA-Script, occasional use of the Latin alphabet (for example, when sending an SMS or email) is mainly due to the technological ease of use associated with the alphabet. For the second and third generation Persian speaking immigrants, who are less likely to have had the opportunity to learn PA-Script and more likely to have been educated in a language written in the Latin script, a romanized script for Persian is usually a natural choice.

Romanized Persian transcription has a relatively long history [15][10][17][7]. A Latin-script was introduced in Tadjikistan at the end of 1920's but was later abandoned in favor of the Cyrillic script. Since the beginning of the 20th century a number of proposals for romanization of Persian script have been presented [19][1][2] but the proposals do not include the details necessary for a writing system. Unfortunately, despite its long history, romanized Persian has not been the subject of standardization. We hope Dabire will serve as a starting point in this respect.

The rest of this paper outlines some phonological and morphophonological issues that are related to the orthography and describes some of the major Dabire-conventions, such as, writing compound words, foreign words and certain grammatical constructs such as the enclitic article *Ezâfe*. Unless it is stated otherwise, the main stream accent of Persian in Iran is used in the paper. [12] gives a more comprehensive description of Dabire.

2 Persian Phonology

This section gives a brief account of Persian phonology. A rather simple, "off-the-shelf", technology for designing a new alphabet for a language is to study its phonology and identify the minimal pairs which can subsequently form the basis for a linguistically sound (phonemic and consistent) writing system [4]. Persian is an Indo-European language and apart from some minor disagreements (such as existence of diphthongs or the phonemic status of the glottal plosive at onset) its

INFOS2008, March 27-29, 2008 Cairo-Egypt

phonology is quite well-understood [20][8][16][22].

Persian has 29 phonemes including 23 consonants and 6 vowels. Furthermore, Arabic loan words include the phonologically significant glottal stop ***Hamze*** (IPA: ʔ) and pharyngeal fricative ***Eyn*** (IPA: ʕ). However, in Persian, these two phonemes have exactly the same pronunciation and are therefore considered as one phoneme and denoted as /'/ - we also choose to call this phoneme *Ist* (stop). Persian consonants are shown in Table-2. The table also includes two columns ("P-Script" and "Dabire") that list the minimal sets of constant graphemes for Persian. P-Script is the subset of PA-Script which is usually used in transliteration of non-Arabic loan words and creation of new Persian words. The only difference between P-Script and PA-Script is that the latter contains some more graphemes that are, in principle, redundant for Persian since the adoption of the Arabic script did not affect Persian phonology. These redundant graphemes are listed in Table-1.

Phoneme	Graphemes	Used in Loan Words
/ʔ/	ع	Arabic
/h/	ح	Arabic
/q/	ق	Arabic, Turkish
/s/	ص ,ث	Arabic
/t/	ط	Arabic
/z/	ذ, ض, ظ	Arabic

Table 1. *Redundant graphemes of PA-Script*

Vowels are listed in Table-3. Three of the six vowels are the so called long vowels (*â*, *i*, *u*) and three are short vowels (*a*, *e*, *o*). The representation of vowels in P-Script depends on two factors: 1) the position of the vowel in the syllable, 2) whether it occurs isolated or at the beginning, the middle or the end of a cluster of graphemes written in the cursive format [16][12]. The details of vowel-representation in P-Script are outside the scope of this paper.

Certain combinations of the approximants [ʊ] and [ɟ], could, from phonetic point of view be considered as diphthongs. These are, [aɟ], [æɟ], [ɛɟ], [oɟ], [uɟ], [oʊ] [20]. For example, [ɛɟ] and [oʊ] in: *peyk* (courier), *Mowlavi* (Rumi). There is, however, no convincing evidence that these formations are diphthongs. For example, syllabic constructions of the form /ow/CC or C/ow/CC can not be found in Persian [22] which weakens the status of /ow/ as a diphthong. Furthermore, when these constructions appear in compound words, their second constituent, the semi-vowels, usually leave the partnership and initiate a new syllable. Consider, for example, *peyâpey* which consists of the three syllables *pe.yâ.pey* (CV.CV.CVC) rather than *pey.â.pey* (CVC.V.CVC). So, for all practical purposes, these instances of /w/ and /y/ could be considered as consonants rather than constituents of a diphthong [20][16].

The syllabic structure of Persian is shown by the following syntactic rules.

Syllable →[*Onset*]*Rime*
Onset → C
Rime → V[C[C]]

Unfolding these rules results in six possible syllable templates: V, VC, VCC, CV, CVC, CVCC. Optionality of the onset, however, is not a settled issue. Samare [20] considers onset to be compulsory whereas Neysari [16] and others [22][10] consider it optional.

3 Morphophonological Issues

Persian is a productive language and new words are constructed by combination of affixes (dominated by suffixes), nouns, stems and adjectives. When constructing a writing system for such a language, two main issues are of concern:

- what phonological or mophophonolical alternations or interactions occur between neighboring phonemes or morphemes
- how should the orthographic representation of a compound word relate to its constituents

In Persian, most of the inter-morphemic phonological alternations aim to improve the euphony in hiatus situations. These alternations mostly include epenthesis and elision.

3.1 Alternation Rules

This section discusses some phonological alternations that usually affect the orthographic represenation

of words. When automatically converting between writing systems or generating words according to the morphology of the language, realization of the alternations is important. A more complete account is presented in [12].

3.1.1 Epenthesis

Epenthesis is the insertion of a sound, a letter or a syllable into a word to facilitate its pronunciation. In Persian, intervocalic epenthesis is quite common.

Persian morphology is inflectional and derivational and includes some suffixes and enclitics that begin with vowels. When these suffixes are concatenated with words that end in a vowel, an interaction between the vowels results. This interaction is usually peaceful and in speech presents itself as a graceful transition from the sound of one vowel to the other. For example, in *pâiz* (autumn), *rauf* (kind), *Soed* (Sweden) and *zendei* (to be alive 2SG+PRES).

In other situations, a direct transition from one vowel to another is not smooth and certain consonants are inserted as mediators between the neighboring vowels. Here are some examples of epenthesis:

/g/ *zendei→zendegi* (life)
/y/ *pâe→pâye* (base)
[*j*] *siâh→siâh* [sɪjah] (black)

3.1.2 Epenthesis in Loan Words

As mentioned earlier, the syllable structure in Persian is [C]V[C[C]] which only accepts V, VC, VCC, CV, CVC and CVCC as syllables. Foreign words that have clusters of consonants at onset or more than two consonants in rime (for example, CCV (ska), CCC (krk), CCCVCC (Spring), CVCCC (Minsk)) are not accepted.

When foreign words enter Persian as new words or are pronounced according to the constraints of Persian rather than the source language, their syllabic structure is modified to fulfill the limitations imposed by [C]V[C[C]]. In particular clusters of consonants may be broken by inserting vowels between the constituents of the cluster in order to create syllables that are tolerated. For example, *b[o]luz* (blouse), *d[e].rink* (drink). A reasonable convention in Dabire is to write these words with the epenthesis.

3.1.3 Elision

Sometimes certain sounds are deleted to improve pronunciation. This is usually applied to unstressed sounds and sometimes it only affects the pronunciation and the word is written as usual.

/e/	*ke in*	*kin* (*ke* (that) , *in* (this))
/a/	*to ast*	*tost* (*to* (you 1SG), *ast* (is))
/o/	*begozašt*	*begzašt* (passed)

Another example of elision occurs in the context of the affix *-estân*. This suffix is added to nouns to create nouns representing a certain place, city or state, for example, *golestân* (garden) and *Tâjikestân* (Tadjikistan). When the suffixed word ends with a vowel, the *e* of *-estân* is eliminated, for example, in *Hendustân* (India).

3.1.4 Vowel Transitions

Some vowel transitions are quite common in Persian,

e→i	*beâ*	*biâ* (come!)
u→o	*bihude*	*bihode* (useless)
â→a	*âgâh*	*âgah* (conscious)
i→e	*niku*	*neku* (good)
e→o	*bero*	*boro* (go!)

Sometimes transitions are combinations of simple transitions, for example, in the following cases, first the *a* in *ast* (is) is dropped and then the *e* of the previous word is changed to *i*.

ke ast → kist (who is she/he?)
ce ast → cist (what is it?)

4 Dabire, A Romanized Transcription for Persian

Representation of Persian phonemes in a romanized writing system is straightforward, there is a one-to-one correspondence between phonemes and graphemes. Dabire uses an extended Latin alphabet consisting of 30 graphemes. These graphemes and their correspondence to phonemes are shown in Table-2 and Table-3.

Defining a writing system for a language involves much more that specifying a minimal set of graphemes for the language. Specifying conventions for writing compound words, foreign names and special constructions such as *Ezâfe* in Persian are examples of issues that need to be addressed. Conventions should be easy to learn and should also be formulated to minimize exceptions. There should also be a certain level of toler-

ance for exceptions which are dictated by factors such as historical use of certain words.

The overall design principles for Dabire are based on creating a morphophonemic writing system which respects the phonological alternations that have evolved over time. In contrast to English writing, Dabire is strongly phonemic. However, just like in English, morphophonemic principles have higher priority than phonemic principles. For example, the advantage of writing the plural of the English word *thing* as *things* and not *thingz* is that the identity of the possessive *s* is maintained. Similarly, in Dabire we write *šanbe* (Saturday) rather than *šambe* which better reflects the real pronunciation.

In the rest of this section some of the choices that we have made in Dabire will be presented.

4.1 Writing *Ezâfe*

The enclitic article *Ezâfe* (written as *e* or *ye*) is used to connect the head of a noun phrase to its modifier (*mozâf* and *mozâfon elayh*). It is usually not written in the PA-Script. In Dabire, we propose that it be written as a separate word since it links two words and it will be inappropriate to connect it to either of the two, for example, *dar e kiosk* (kiosk door), *sib e sorx* (red apple), *jâ ye man* (my place). Various formats have been adopted by various authors, for example, Neysari [15] and [2] write *Ezâfe* as a separate word, whereas [10] and [22] use a hyphon to separate the head from *Ezâfe*. Unipers [1] attaches *Ezâfe* to the word which is being modified (*sibe sorx*).

4.2 Gemination

In PA-Script, gemination is indicated by ***Tašdid***, a special diacritic which is placed on a consonant indicating that it is duplicated. For example, see *xatt* (line) and *arre* (saw) in the glossary.

Gemination, in general, means that the consonant is pronounced twice, the first time ending a syllable and the second time initiating a new one. In some loan Arabic words, such as *hadd* (limi), where gemination concerns the final letter of the word, the second instance of the geminated consonant is only pronounced if the following word starts with a vowel. For example, in *hadd dârad* (has a limit), the second *d* is not pronounced, in *hadd o marz* it is pronounced and if the word precedes a paus then the pronunciation of the geminated consonant is prolonged rather than doubled.

A reasonable romanized representation of gemination is to repeat the consonant irrespective of how it is pronounced. There is, however, an exception: when the geminated letter is the PA-letter ***Ye***, it could either mean that the consonant /y/ is doubled (*Sorayya*) or that we have a sequence of /i/ and /y/ (*tahiye*). In Dabire, these are distinguished in transcription as *yy* and *iy* respectively. In PA-Script, *yy* and *iy* have the same orthographic representation.

4.3 Writing *Ist*

As we mentioned earlier, there is no consensus on the phonemic status of the glottal *Ist* at syllable onset. The question is whether the "hard attack" glottal stop [21] is a consonant phoneme or not. Glottal stops exist in many languages including a number of Indo-European languages. Longman's Pronunciation Dictionary [21] defines *hard attack* as,

> When a word or a syllable begins with a vowel sound, it is possible to start the vowel from a position where the vocal folds are first held closed, then burst open for the vowel: that is, to precede the vowel by a GLOTTAL STOP. This way of starting a vowel is called a **hard attack**.

LPD continues by saying that, in English, hard attack is not customary and is only used for emphasis. For example, "to eat" is usually expressed as [tu i:t] but sometimes as [tu ʔi:t]. Despite the fact that glottal stop occurs in English, there is no orthographical representation of it since there is no need for it.

In fact, the same kind of reasoning holds in Persian phonology. When a syllable starts with a vowel (V, VC, VCC), our speech generation mechanism would place a glottal plosive at the start of the process in order to initiate the production of the vowel. Irrespective of the phonemic status of *Ist* in syllable onset, it is doubtful whether it should have an orthographical representation, for example, it is unnecessary to write *'âb* instead of *âb* (water). The latter is simpler and there is little doubt as to how it should be pronounced.

Another reason for not assigning graphic representation to glottal stops at onset is that they happily give up thier position and disappear, for example, when *gol* (flower) joins *'âb* (water) to form *golâb* (rose-water), the resulting word has the syllabification *go.lâb*.

In Arabic loan words, however, *Ist* also occurs in coda and is phonemically significant. For example, consider the Persian word *bad* (bad) and the Arabic word *ba'd* (after). Dropping *'* from the second word leads to ambiguity. In summary, the adopted convention in Dabire is to only write *Ist* when it occurs in coda, never as a consonant initiating a syllable.

4.4 Writing Compound Words

The conventions of Dabire and PA-Script in writing compound words are different. In PA-Script, due to the semi-cursive nature of the script, context sensitive letter shapes, similarity between letter shapes and multiple roles for certain letters, the general orthographic rule is to keep the original shape of the sub-words of a compound in order to minimize the effort put into word identification. Unfortunately, the existing conventions for PA-Script contain a large number of exceptions and some ad hoc conventions [16][6][3][11].

The conventions for writing compounds in Dabire are similar to some European writing systems such as Swedish and German. The default format is to write compound words in the closed format [18]. For example, *ravânpezešk* (psychologist) with subwords *ravân* (soul) and *pezešk* (physician) or *badtarin* (worst) with subwords *bad* (bad) and *-tarin* (suffix for indicating superlative form of adjectives). Depending on the structure and the nature of a compound word, it may be written in one of the formats: *open*, *hyphenated* or *closed* as described in the following subsections. Compound words that are used extensively during a long period of time usually adopt the closed format.

4.4.1 Open Format

Open format refers to the case where subwords of a compound are written separately using a space as delimiter. This format is mainly applied to the following four cases: 1) nominal compounds containing *Ezâfe*, for example, *Emârât e Arabi* (United Arab Emirates), 2) nominal compounds in which *Ezâfe* has been removed and the order of the modified noun and its modifier is switched (called *qalb e ezâfe*), for example, *gardande falak*[1] (rotating heavenly wheel) which is equivalent to *falak e gardande*, 3) nominals formed using the connective *o* (and), for example, *xert o pert* (frippery), *kâr o bâr* (job, work), 4) compound and complex verbs and their derivatives such as *barkenâr šod* (was dismissed) and *vazir e barkenâr šode* (the dismissed minister).

4.4.2 Hyphenated Format

The hyphenated format is used for situations similar to the open format but a hyphen rather than a space is used as the delimiter. For example, *pic-gušti* (screwdriver). This format is generally applicable to newly formed compounds. Extensive and prolonged use of a hyphenated compound word justifies removal of the hyphens and the adoption of the closed format for the word - as is indeed the case for *picgušti*.

4.4.3 Closed Format

The closed format serves as the default format in Dabire, the words of a compound term are concatenated to form a new word. For example, *golâb* (rosewater) formed by using *gol* (flower) and *âb* (water). This format also applies to the indefinite article *i* and the Arabic definite article *al*, for example, *mardi* (a man), *alqamar* (the moon), *alšams*. Notice that we are not making any distinction between "solar" and "lunar" Arabic letters.

The closed format is also used for all kinds of affixes.

4.5 Names and Trademarks

The transcription conventions of Dabire do not apply to trademarks and names. In general, names of people, places and registered products are exempted

[1] From the following poem by the great poet and mathematician Khayyam:

Piš az man o to, leyl o nahâri budast
Gardande falak niz, be kâri budast

Har jâ ke qadam nahi to bar ru ye zamin
Ân mardomak e cešm e negâri budast

from these conventions. For example, Smirnoff, Linux, k.d. lang and VOLVO must be written as they are and not according to how these words are pronounced in Persian. The same is true for the name of individuals, for example, although a reasonable way of writing 'Geoffrey' in Dabire would be as *Jefri*, it would be inappropriate to do so. However, if a name's original script is not Latin, then it is transcribed into Dabire following the phonological constraints of Persian.

Another issue related to Iranian names is that, in spoken Persian, an *Ezâfe* is placed between the first name and the surname of a person - unless the first name ends with the vowel *â*. In Dabire these instances of *Ezâfe* are not represented in writing, for example, we will write *Shirin Ebâdi* although her name is pronounced as *Shirin e Ebâdi*.

4.6 Foreign Loan Words

In a proposal such as Dabire, dealing with foreign loan-words is not trivial. At the same time that we strive for general writing rules, we need to respect well-established usage of words. There are essentially two possibilities, either the word already has a well-established pronunciation in Persian or it doesn't. In the former case, we base the Dabire-representation on the established pronunciation, otherwise, its pronunciation in the original language will be used for determining a suitable orthography for it. Obviously, as mentioned in section 3.1.2, the phonological constraints of Persian must be imposed on the new word. For example, if the English word "star" were to enter Persian, an initial *e* would be inserted and it would be written as *estâr* ensuring a VC.CVC syllabification and rejecting the original syllabification CCVC.

4.7 Capitalization

Since capitalization of certain letters improves the readability of text, in line with many other Latin-based scripts, we propose similar rules for it. The disadvantage of capitalization is that the number of graphemes are doubled. Anyhow, just in case future developments speak for capitalization, we list a number of cases where capitalization is useful.

- The first word of a sentence is capitalized.
- The first word of a syntactically complete quoted sentence is capitalized. For example, *Mahnâz goft, "Xoš âmadid."* (Mahnâz said, "You are welcome.".)
- Proper names and geographic names are always capitalized.
- Even abbreviations that have through time obtained the status of a word could be capitalised. For example, we not only can write *UNESCO*, but also *Unesco*.
- Contrary to the practice in English, the names of weekdays, months, and years are not capitalized. For example, *došanbe* (Monday), *farvardin* (First month of the Iranian calender), *ut* (August)
- Abbreviated titles are always capitalized. For example, *Du. Lâle Kermâni* (Miss. Lâle Kermâni)
- In articles and books, all main words appearing in a title or chapter names are capitalized, for example, *Fasl e Yek: Zabân e Fârsi* (Chapter One: Persian Language)

4.8 Abbreviation

We consider the following types of abbreviations.

- Abbreviation of single words should follow an standardization. These abbreviations should end with a period (.). For example, *Teh.* as a possible abbreviations for *Tehran*, *Dank. Barq* (Dept. of Electrical Engineering) as a possible abbreviation for *Dâneškade ye Barq*, *q.* as a possible abbreviation of *qeyd* (adverb). A simple rule for creating such abbreviations would be to start with the first letter of the word and continue including letters until a vowel is reached or a unique abbreviation is obtained. If this method leads to abbreviations that are somehow unreasonable then we can include the first syllable of the original word and proceed with the rest of the word according to the rule described earlier, and start with the second syllable and do as before. Another alternative would be to start with the first syllable and then select a letter from the rest of the word so that a unique and suitable abbreviation is formed. Some examples follow.

Iran	*Ir.* (Iran)
Tehrân	*Teh.* (Tehran)
dânešgâh	*dan.* (university)
dâneškade	*dank.* (department)

Naturally, some words can be exempted from these rules. For example, when we create abrviations for days of the week or months of the year. In such cases other requirements, such as constant length of abbreviations, may determine the format of the abbreviation.

- When abbreviating a compound name, the first letter of each major word in the compound name should be included in the abbreviation, for example, *ŠNI* as the abbreviation of *Šerkat e Naft e Irân* (Iranian Petroleum Company), *RI* as the abbreviation of *Râdio Irân* (Radio Iran).

  When the resulting abbreviations are used as words in Persian, then we may write them as normal words. If we imagine there were a ministry called *Vezârat e Âb Va Enerži e Kešvar*, then it could be abbreviated as *VÂVEK* which may gradually turn into a normal word that may be written (and pronounced) as vâvek or Vâvek.

- When abbreviating an expression or a construction consisting of two words or more, the first letter of each major word is included in the abbreviation and succeeded by a period. For example, *b.m.* as an abbreviation for *barâye mesâl* (for example), *v.e.a.* for *va elâ âxar* (and so on), *b.b.i.* as an abbreviation of *banâ bar in* (therefore, hence). No spaces should be included in the abbreviation. Some abbreviations such as units of measurements should be exempeted from this kind of punctuation. For example, it is better to *5cm* (5 centimeters) rather than *5c.m.* or *80 GB* (80 gigabytes) rather than *80 G.B.*.

- Titles should normally be abbreviated, for example, *Porofesor Pari Mehrabân* (Professor Pari Mehrabân) could be abbreviated as *Porof. Pari Mehrabân*.

- It is quite practical to have abbreviations for ordinal numbers. In Persian, ordinals are constructed by adding the suffix *-om* to a number. For example, *yekom* (first), *dovom* (second), *sevom* (third), and so on. We propose abbreviations as thus: 1om (first), 2om (second), 3om (third), . . ., or as 1^{om}, 2^{om}, 3^{om}, 4^{om}, . . ., in mathematical texts.

- Dash – (in Persian, *xatt-e-fâsele*) can be used as an abbreviation of the word *tâ* (to) which is used to specify intervals, for example, *s. 11-23* as an abbreviation of *az safhe ye 11 tâ 23* (from page 11 to 23).

- Dates may be written in any of the following ways. We exemplify these formats using the Friday, 31st day of the first month *farvardin*, year 1363.

  âdine, 31 farvardin 1363
  31 farvardin 1363
  28 far. 1363
  28-9-63
  28-9-1363 hš.

  In the final two cases, the date may be succeeded by one of the abbreviations *hš.*, *hq.* or *mi.* to indicate whether one is using the Iranian solar calender (*hejri e šamsi*), Islamic lunar calender (*hejri e qamari*) or the the Christian solar calender (*milâdi*).

- Hours of the day may be written in any of the following formats:

  1:15 pi (1:15 am)
  12:00 pa (12:00 pm)
  5pa (5pm)
  13:15 (13:15)

  As these examples illustrate, we propose using *pa* and *pi* as abbreviations of *pas* (after) and *piš* (before).

4.9 Punctuation

Dabire and PA-Script, in principle, follow the punctuation rules that are practiced in most other writing systems. We list some of the common conventions.

- In order to specify a list of items in the text flow, end the word preceeding the list with a colon (:) and separate the items with a comma. For example, *Yek hafte haft ruz dârad: šanbe, yekšanbe, došanbe, sešanbe, cahâršanbe, panjšanbe, âdine* (A week has seven days: Saturday, Sunday, Monday, Tuesday, Wednesday, Thursday, Friday).

- Use commas to improve readablity of the text. For example, *Barâye mesâl, farvardin 31 ruz ast va mehr 30 ruz* (For example, month of farvardin is 31 days and mehr 30 days)

- Quoting rules are as usual, for example, *Âraš goft: Midâni "omniscience" be Fârsi ci miše?* (Araš said: Do you know what "omniscience" would be in Persian?)

- A punctuation character should be followed by a space unless it is internal to a word. For example, in *b.b.i.* which is the abbreviation of *banâ bar in* (therefore) no spaces follow the internal punctuations.

- A period should be used to mark the decimal part of a number, for example, 3.14. Commas should be used to improve readability of large numbers, for example, $5,678$, $10,000,000$, $12,345.99$.

5 Concluding Remarks

In his talk at COLING-2004, Martin Kay [9] underlined the importance of explicit representation of vowels as a prerequisite for a fair and correct study of languages with Arabic-based scripts. Our proposal is inline with his recommendations. During the past year, we have used Dabire in a number of NLP-projects [13]. Although, most of this work could have been done without specifically using Dabire, we have realized that using a romanized system has facilitated our work and enabled us to communicate our work with colleagues that are not familiar with the Persian language or the traditional orthography.

Another issue of interest is related to the so called orthographic depth principle. Psychologists classify orthographies as shallow or deep depending on whether it is easy or difficult to correctly predict the pronunciation of a word based on its orthographical representation [5]. English writing system is, therefore, a deep orthography whereas Serbo-Croatian enjoys a shallow orthography.

In the PA-Script, the diacritics are seldom written and in practice the script becomes rather opaque specially for the novices [14]. Furthermore, there is a many-to-many correspondence between phonemes and graphemes which challenges the spelling capabilities of most people. From orthographic-depth point of view, Dabire is phonologically more transparent and therefore easier to learn and process. We have performed a number of small pedagogical experiments where we have tried to teach Dabire to Persian speakers and non-speakers who are familiar with the Latin alphabet, in all cases, learning Dabire has been a matter of a few hours.

Although Dabire-like writing systems have a number of advantages, PA-Script will naturally continue to be the main orthography used for writing Persian. For this reason, automatic transcription systems for converting back and forth between Dabire and PA-Script will have an important role in future applications on the Internet.

Finally, we hope that Dabire will serve as an initial step towards creating a standard for work in Persian linguistics. The conventions we have listed in this paper are proposals that need to be discussed and developed further.

References

[1] Unipers. http://www.unipers.com, 2007.

[2] Xatt e Now. http://www.eurofarsi.com, 2007.

[3] M. S. Adib-Soltâni. *An Introduction to Persian Orthography - (in Persian)*. Amir Kabir Publishing House, Tehrân, 2000.

[4] S. Bird. Orthography and Identity in Cameroon. 2001.

[5] D. Bresner and M. C. Smith. *Basic Processes in Reading: Is the Orthographic Depth Hypothesis Sinking?*, chapter 3. Advances in Psychology. North Holland, 1992.

[6] Farhangestan. *Dastur e Khatt e Farsi (Persian Orthography)*, volume Supplement No. 7 , February 2000. Persian Academy, Tehran, February 2003.

[7] F. Hashabeiky. *Persian Orthography - Modification or Changeover?* Acta Universitatis Upsalienis, 2005.

[8] C. Jahani. The Glottal Plosive: A Phoneme in Spoken Modern Persian or Not? In Éva Ágnes Csató,

B. Isaksson, and C. Jahani, editors, *Linguistic Convergence and Areal Diffusion: Case studies fron Iranian, Semitic and Turkic*, pages 79–96. Routledge Curzon, 2005.

[9] M. Kay. Arabic Script-based languages deserve to be studied linguistically. In A. Farghaly and K. Megerdoomian, editors, *Computational Approaches to Arabic Script-Based Languages*, 2004.

[10] G. Lazard. *Grammaire du Persan Contemporain*. Paris: Klincksieck, 1957.

[11] R. R. Z. Malek. *Qavâed e Emlâ ye Fârsi*. Golâb, 2001 (1380 hš).

[12] J. Maleki. A Romanized Transcription for Persian. Working Paper, NLPLAB, IDA, Linköping University, 2007.

[13] J. Maleki and L. Ahrenberg. Converting Romanized Persian to Perso-Arabic Writing System Using Syllabification. In *Proceedings of the LREC2008, Marakech*, 2008.

[14] P. Natel-Khanlari. *The History of Persian Language*, volume I. New Delhi Press, 1979.

[15] S. Neysari. *Ketâb-e Avval Fârsi be xatt-e jahâni*. Publisher Unknown, 1956.

[16] S. Neysari. *A Study on Persian Orthography - (in Persian)*. Sâzmân e Câp o Entešârât, 1996.

[17] J. P. Perry. *A Tajik Persian Reference Grammar*. Brill, 2005.

[18] R. M. Ritter. *The Oxford Guide to Style*. Oxford University Press, 2002.

[19] S. H. Taqizâde. Jonbeše Melli e Adabi. *Armaqân*, (8/9), 1941.

[20] Y. Samare. *Âvâšenâsi e zabân e fârsi - (in Persian)*. Markaz-e našr-e dânešgâhi, 2nd edition, 1997.

[21] J. C. Wells. *Longman Pronunciation Dictionary*, pages 307, 327. Longman, 1990.

[22] G. Windfuhr. Persian. In B. Comrie, editor, *The World's Major Languages*, pages 523–546. Routledge, 1989.

Glossary

Because of some spacing conflicts in LaTeX, apparently due to the size of the ArabTeX fonts, the following words are moved from the body of the text to this section.

[*âb*, آب, **ãb**, *ʔab*, water]

[*alqamar*, القمر, ›**lq̣mr**, *ælqæmæɪ*, the moon]

[*alšams*, الشمس, ›**lšms**, *æʃːæmsˌ* the sun]

[*arre*, ارّه, ›**r̈h**, *æɪ ɪɛ*, saw]

[*bad*, بد, **bd**, *bæd*, bad]

[*ba'd*, بعد, **b‹d**, *bæʔd*, after, later]

[*gol*, گل, **gl**, *gol*, flower]

[*golâb*, گلاب, **gl›b**, *golab*, rose water]

[*hadd*, حد, **ḥd**, *hædˑ*, limit]

[*Mowlavi*, مولوی, **mwlwy**, *moʊlævɪ*, Rumi]

[*pâiz*, پائیز, **p›i̯yz**, *paʔɪz*, autumn]

[*peyâpey*, پیا پی, **py›py**, *pəjɑpəj*, successively]

[*peyk*, پیك, **pyk**, *pəjk*, courier]

[*siâh*, سیاه, **sy›h**, *sɪɑh*, black]

[*Sorayyâ*, ثریا, **ṯrÿ›**, *soɪæj jɑ*, Arabic name for girls]

[*tahiye*, تهیّه, **thÿh**, *tæhɪ jɛ*, prepare]

[*xatt*, خطّ, **xẗ**, *xætˑ*, line, orthogrpahy]

[*zendegi*, زندگی, **zndgy**, *zɛndɛgɪ*, life]

[*zendei*, زنده ای, **zndh›y**, *zɛndɛʔɪ*, you are alive]

	Phoneme	Characteristics	IPA	P-Script	Dabire	Example
Stop	/b/	Voiced, Bilabial	*b*	ب	b	bi, بو (smell)
	/p/	Voiceless, Bilabial	*p*	پ	p	pâ, پا (foot)
	/t/	Voiceless, Denti-Alveolar	*t*	ت	t	tu, تو (inside)
	/d/	Voiced, Denti-Alveolar	*d*	د	d	dar, در (door)
	/k/	Voiceless, Velar	*k*	ك	k	ki, کی (who)
	/g/	Voiced, Velar	*k*	گ	g	gâv, گاو (cow)
	/’/	Voiceless, Glottal	*ʔ*	ء	’	joz’, جزء (part)
Affricate	/č/	Voiceless, Alveo-Palatal	*tʃ*	چ	c	ce, چه (what)
	/ǰ/	Voiced, Alveo-Palatal	*dʒ*	ج	j	jâ, جا (place)
Fricative	/f/	Voiceless, Labio-Dental	*f*	ف	f	fut, فوت (blow)
	/v/	Voiced, Labio-Dental	*v*	و	v	vâke, واکه (vowel)
	/s/	Voiceless, Alveolar	*s*	س	s	sar, سر (head)
	/z/	Voiced, Alveolar	*z*	ز	z	zar, زر (gold)
	/š/	Voiceless, Alveo-Palatal	*ʃ*	ش	š	šab, شب (night)
	/ž/	Voiced, Alveo-Palatal	*ʒ*	ژ	ž	dež, دژ (castle)
	/x/	Voiceless, Uvular	*x*	خ	x	xâk, خاك (soil)
	/q/	Voiced/Voiceless, Uvular	*q*	غ	q	qam, غم (sorrow)
	/h/	Voiceless, Glottal	*h*	ه	h	har, هر (every)
Liquid	/r/	Voiced, Alveolar, Trill	*ɹ*	ر	r	ru, رو (face)
	/l/	Voiced, Alveolar, Lateral	*l*	ل	l	liz, لیز (slippery)
Nasal	/m/	Voiced, Bilabial	*m*	م	m	mu, مو (hair)
	/n/	Voiced, Alveolar	*n*	ن	n	ney, نی (cane)
Semi-vowel	/y/	Voiced, Palatal, Approximant	*j*	ی	y	yâ, یا (or)
	/w/	Bilabial (Rounded), Approximant	*ʊ*	و	w	now, نو (new)

Table 2. *Persian Consonants and their correspondence to the graphemes of PA-Script and Dabire*

Phoneme	Characteristics	IPA	Dabire	Example
/a/	Front, Short, Open, Low, Spread	*æ*	a	bad (bad)
/e/	Front, Short, Half-Close, Mid, Spread	*ɛ*	e	be (to)
/i/	Front, Long, Close, High, Spread	*ɪ*	i	si (thirty)
/â/	Back, Long, Open, Low, Rounded	*ɑ*	â	bâ (with)
/o/	Back, Short, Half-Close, Mid, Rounded	*o*	o	do (two)
/u/	Back, Long, Close, High, Rounded	*u*	u	bu (smell)

Table 3. *Persian Phonemes – Vowels*