Converting Romanized Persian to the Arabic Writing System

Jalal Maleki, Lars Ahrenberg

Natural Language Processing Laboratory
Department of Computer and Information Science
Linköping University
SE-581 83 Linköping
Sweden
jma@ida.liu.se, lah@ida.liu.se

Abstract

This paper describes a syllabification based conversion method for converting romanized Persian text to the traditional Arabic-based writing system. The system is implemented in Xerox XFST and relies on rule based conversion of words rather than using morphological analysis. The paper presents a brief evaluation of the accuracy of the transcriptions generated by the method.

1. Introduction

Persian, is predominantly written in variations of the Arabic writing system. The writing system in Iran, for example, is the Perso-Arabic script (PA-Script) (Neysari, 1996), whereas Latin and Cyrillic scripts have been used elsewhere (Perry, 2005; Hashabeiky, 2005). However, due to the technological limitations in deploying PA-Script in many software systems, the Latin script is very common in blogs, email, SMS and other online settings. Furthermore, many Persian speakers (such as second generation immigrants) who are not familiar with PA-Script exclusively use the Latin alphabet in written Persian communication. It is, therefore, quite reasonable to conclude that, for all practical purposes and specifically in the online world, there are two parallel scripts for Persian.

Although, Latin-based scripts for Persian have existed for many decades, the relationship between these scripts and the traditional PA-Script is not well-understood. The main goal of our work is bridging this gap by developing rules and algorithms for converting back and forth between these writing systems.

Converting Persian text written in a latin-based script to the traditional Persian script is complicated. A considerable source of complication is the transcription of vowels. Factors such as the type of syllable containing the vowel and the characteristics of the neighboring graphemes determine the choice of grapheme (or allographs) for the vowel. As an example, Table-1 shows the various allographs and digraphs used for writing the vowel /i/ in different contexts. Each box in the table contains the graphic element used for /i/ and also an example word in which such a graphic element occurs.

The focus of this paper is on a rule-based method which uses syllabification for determining the selection of correct Arabic-based orthography for a romanized word. The method does not rely on morphological analysis or lexical information which means that it can, for example, be used for transcription of new loan-words which are written in a phonemic Latin-based writing system for Persian, called Dabire (Maleki, 2008). The implementation is based on the finite state transducer technology developed at Xerox (Beesley and Karttunen, 2003; Kaplan and Kay, 1994).

The rest of this paper briefly discusses some issues specific to the scripts and discusses the syllabification method for converting Dabire-text to the Arabic-based writing system. Although this method has many useful applications, it can not hanndle all Arabic loan-words or some of the compound words in Persian.

2. A Brief Account of the Writing Systems

Persian has 24 consonants and 6 vowels. Since there is a one-to-one correspondence between Persian phonemes and the graphemes of the Dabire-romanization (Maleki, 2008), we will use the same symbols to denote the phonemes of Persian and the graphemes of Dabire. It is usually clear from the context whether we are referring to a phoneme or a grapheme. Persian phonemes are:

â, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, š,
t, u, v, w, x, y, z, ž, ’

The vowels are usually divided into two groups: the ’short vowels’ *a*, *e*, *o*, and the ’long vowels’ *â*, *i*, *u*. The consonant ’ is the glottal stop and its occurrences in rime are phonologically significant in Arabic-loan words.

From the orthographic point of view, one of the important conventions in the romanized system, Dabire, is that compound words are written in three different formats: open, hyphenated and closed format. This is similar to conventions used in English (Ritter, 2002). This convention differs from the convention used in the traditional Arabic-based script where sub-words of a compound are expected to be written in an open format separated by a zero width space rather than a normal space.

The Perso-Arabic script used in Iran includes all Arabic graphemes and diacritics for representation of short vowels as well as four Persian-specific graphemes for representing /p/, /č/, /ž/ and /g/. However, in what follows, we will introduce a subset of PA-Script which we call P-Script. The graphemes of P-Script are the following:

/i/	Word Initial	Segment Initial	Segment Medial	Segment Final	Intra-Word Isolated
V, VC, VCC	ایـ این	ئیـ پائیز	ـئیـ لئیم	ـئی خالیئی	ای ,ئی رفته‌ای ,بانوئی
CVC, CVCC		یـ پردیس	ـیـ سیزده		
CV		یـ دیدار	ـیـ بیدار	ـی خاکی	ی کاری

Table 1: *Mapping the phoneme /i/ to P-Script graphemes depends on its position in the syllable and the word*

ا (Alef)	ب (Be)	پ (Pe)	ت (Te)	ج (Jim)
چ (Ce)	خ (Xe)	د (Dâl)	ر (Re)	ز (Ze)
ژ (Že)	س (Sin)	ش (Šin)	غ (Qeyn)	ف (Fe)
ك (Kâf)	گ (Gâf)	ل (Lâm)	م (Mim)	ن (Nun)
و (Vâv)	ه (He)	ی (Ye)		

Here, we are considering the graphemes آ (Â-ye madde) and ئ (Hamze) as allographs of ا (Alef). This subset of PA-Script graphemes is in principle equivalent to the list specified in (Kasravi, 1944).

The only difference between P-Script and PA-Script is that the latter includes some Arabic-specific letters as well. These letters are, however, redundant in Persian (Perry, 2005). The significance of P-Script is that it is sufficient for representing Persian phonemes and transcription of words from non-Arabic languages. Arabic words and terms usually keep their original spelling.

P-Script is a semi-cursive writing system where words are written from right to left by joining the graphemes according to certain constraints. Even the typed variations of the writing system simulate the hand-written cursive style. The mapping between the consonant phonemes and graphemes is straightforward. Vowel representation is, however, more complicated, see, for example, the mapping of /i/ to P-Script in Table-1.

A piece of text in Persian consists of a sequence of words written from right to left and separated by usual delimiters. Here is an expression with five words: من امروز يك سيب خوردم (I ate an apple today)

An alphabetic word is written as a sequence of one or more segments written from right to left. Segments are separated by a zero-width space. A *segment* is a sequence of conjoined graphemes and is merely a graphical concept and does not necessarily correspond to any linguistic unit.

Because of the cursive nature of P-Script, both in handwriting and mechanized writing, each grapheme can appear as a number of allographs. An allograph is essentially adaptation of a grapheme so that it can properly join its neighboring allographs. There are four different positions in which a cursive allograph can appear: *Segment-Initial*, *Segment-Medial*, *Segment-Final* and *Isolated*.

An alphabetic word-segment is written as a sequence of one or more allographs written from right to left. If a segment consists of a single grapheme then its isolated form is used, otherwise, the word-segment begins with a segment-initial allograph and is followed by zero or more segment-medial allographs and ends with a segment-final allograph. The segment گفتند,[1] for example, starts with the segment-initial allograph گـ which is followed by segment-medial ـفـ, ـتـ and ـنـ, and ends with the segment-final ـد.

The graphemes ا, د, ذ, ر, ز, ژ and و, lack segment-medial allographs and, thereby, never join the subsequent grapheme. We refer to these graphemes as R-joiners since they can only join the grapheme on their right. All other graphemes are LR-joiners since they can join on both sides. R-joiners always terminate the segment.

Factors that complicate the traditional script can be summarized as follows:

- Choice of graphemes for representing vowels depends on its position in the word and in the syllable
- The script has a relatively ad hoc set of conventions for writing compound words. These are presented in a recent publication by the Persian Academy (Farhangestan, 2003). Many consider these conventions as inconsistent and unacceptable (Malek, 2001).
- Some graphemes have multiple roles, for example, ه (He) is used for denoting /h/ as well as word final /a/ and /e/. Here are some words where it represents a vowel:

  [*parvâne*, پروانه, **prwˀnh**, *pæɹvanɛ*, butterfly]

  [*korre*, کرّه, **kr̈̄h**, *koɹ ɹɛ*, foal]

  [*na*, نه, **nh**, *næ*, no]

  When such a word forms the non-final sub-word of a compound word, the ه being a LR-joiner, will join the initial grapheme of the following word and its shape will change from the segment-final form to the segment-medial. This leads to misinterpretation since ه represents a vowel only if it occurs in the segment-final or isolated form. For example, forming a com-

[1] [*goftand*, گفتند, **gftnd**, *goftænd*, they said]

[2] [*parvâne*, پروانه, **prwˀnh**, *pæɹvanɛ*, butterfly]

pound using پروانه[2] (butterfly) and وار (like) can be either written as پروانهوار or پروانه‌وار, but the latter is less prone to misinterpretation. But the situation is more complicated, for example, it is fine to write the plural of ماه[3] as ماهها but it is inappropriate to write the plural of آیت اللّه[4] as آیت اللّهها.

- Certain short vowels are sometimes written and sometimes not. For example, /o/ is sometimes written as و, sometimes as a diacritic اُ (here placed on Alef), and in general not represented writing.

When converting Dabire to PA-Script we also face the problem of 1-to-many mapping between phonemes and graphemes.

3. The XFST Implementation

The conversion system is implemented using Xerox XFST (Beesley and Karttunen, 2003) and is defined as a multi-level composition of transducers `Sfy .o. CR .o. D2P` that contains a syllabification transducer `Sfy`, a transducer for realizing context sensitive orthographic representation `CR` and a simple transducer `D2P` for mapping to P-Script. A more detailed description is given below.

Syllabification: The first step in the conversion system is to syllabify an input string (a word in Dabire). The syllabification transducer works from left to right on the input string and ensures that the number of consonants in the onset is maximized. Given the syllabic structure of Persian, this essentially means that if a vowel, V, is preceded by a consonant, C, then CV initiates a syllable. For example, for a word such as *jârue*, the syllabification *jâ.ru.e* (CV.CV.V) is selected and *jâr.u.e* (CVC.V.V) is rejected. Here are some more examples:

hadd	*hadd*
haddi	*had.di*
Jalâl	*ja.lâl*
âbâd	*â.bâd*
Jalâlâbâd	*Ja.lâ.lâ.bâd*

The following XFST-definitions form the core of the syllabification:

```
define Sy V|VC|VCC|CV|CVC|CVCC;

define Sfy C* V C* @> ... "." ||  _ Sy;
```

The first statement defines a language (`Sy`) containing all syllables of Dabire. `V`, `VC` etc. are defined as regular languages that represent well-formed syllables in Dabire. For example, CVCC is defined as,

```
define CVCC [C V C C] .o. ~$NotAllowed;
```

which defines the language containing all possible CVCC syllables and excluding the consonant clusters in NotAllowed such as *bp*, *kq*, and *cc* which are not tolerated.

The second statement defines, `Sfy`, a replacement rule (Beesley and Karttunen, 2003) that represents the syllabification process. The replacement operator `@>` ensures that shortest possible strings (of the form `C* V C*`) are selected in a left to right direction and identified as syllables which are separated by dots.

Context Rules: The second component of the FST-system is a transducer that implements the rules and conventions of the writing system. The `CR` transducer is the composition of a large number of replace rules that determine the appropriate graphic representation in the Arabic-based script. For example, the rule,

```
  i -> [A y] || .#. _
```

would replace word-initial occurrences of *i* with the intermediate sequence **ˀy** (denoted as `A y` in the code) which subsequently is transliterated by a rather simple transducer `D2P` to P-Script-graphemes in Unicode (Esfahbod, 2004). Table-2 includes examples that illustrate how this method works. In the third example, occurrences of /i/ are transliterated in three different ways: **y**, **'y**, and **ˀy**. The input to `D2P` contains the short vowels enabling us to generate the appropriate diacritics diacritics if we wish to do so, but following general practice, we currently ignore all short vowels.

Transcription to Unicode. The final step (`D2P`) is the transliteration to the Unicode characters.

4. Conclusions

This paper presents a method for converting Persian text written in a romanized writing system (Dabire) to the traditional Arabic-based orthography. Related work includes (Johanson, 2007) which introduces a method for converting names written in the PA-Script to Latin. (Kashani et al., 2007) uses letter-based alignment in automatic transliteration of proper nouns in Arabic to English, and (Beesley, 2007) applies morphological analysis to transcription to Arabic. However, we are not aware of any earlier work on automatic transcription from romanized to arabized Persian.

The transcription method we use is implemented in XFST, the finite state transducer technology of Xerox. The method is completely rule-based and does not rely on the existence of a dictionary or any other lexical information enabling us to transcribe new words.

The system performs well in the context of transcription of new terms and loan-words, but if these terms and words have Arabic origin, the generated Arabic-based transcription may be different from the original Arabic spelling for the word. This is obviously due to the fact that during the original Arabic–Dabire transcription process there is loss of information. As we mentioned earlier, many Arabic letters are redundant in Persian and are mapped to the same phoneme (and thereby, the same Dabire-grapheme). Therefore, 'back'-transcription (from Dabire to Arabic) will not

[3] [*mâh*, ماه, **mˀh**, *mah*, the moon, month]

[4] [*Âyatollâh*, آیت اللّه, **ãyt ąllh**, *ajætollah*, Ayatollah]

	Dabire Word	Syllabification Sfy	Context Rules CR	Transliteration D2P	Focus of Example
1	Irân	i.rân	**ʾy.rʾn**	ایران	Initial /i/
2	zibâi	zi.bâ.i	**zy.bʾ.’y**	زیبائی	Initial, Final /i/
3	zibâii	zi.bâ.i.i	**zy.bʾ.’y.ʾy**	زیبائی ای	All /i/
4	be	be	**beh**	به	Final /e/
5	na	na	**nah**	نه	Final /a/
6	Tâjikestân	tâ.ji.kes.tân	**tʾ.jy.kes.tʾn**	تاجیکستان	/â/, /i/
7	Jorj Buš	jorj buš	**jorj bwš**	جرج بوش	/o/, /u/

Table 2: *Examples of stepwise conversion from Dabire to P-Script.*
The transcription letters ʾ *and* ʿ *(*Alef *and* Eyn*) are shown as* `A` *an* `E` *in the code examples in the text*

be able to recover the lost information and may generate apparently inaccurate transcriptions.

The degree of the precision is relative to the conventions outlined by the Persian Academy (Farhangestan, 2003). Our method generates close to 100% correct transcriptions for non-Arabic foreign names written in Dabire. Loss of precision in general cases and specially in case of existing words is due to the following factors,

1. Lack of Arabic-specific graphemes in P-Script.

2. The conventions for writing compound words in the traditional orthography is relatively ad hoc. The main principle is to write separate the sub-word components of a compound with a zero-width space in order to avoid situations where the cursive nature of the writing system lead to word shapes that are not easy to identify. But unfortunately, there are many exceptions to this principle. The mismatch between the orthographic conventions for writing compounds in Dabire-romanization and P-Script create some problems for the conversion system. Including lexical information and morphological analysis would remedy some of the problems.

3. Multiple roles for certain graphemes in P-Script complicate the conversion from Dabire to P-Script, For example, پیاده رو (footpath) which is a compound word with two sub-words is written as *piâderow* in Dabire. In the P-Script version the first word ends with a letter ه that usually represents /h/ but it is also used to represent word final /e/ and during the transcription to Dabire this information is lost and it is no longer clear whether the *e* in *piâderow* should be ignored when converting to P-Script or it should be converted to ه. Even in this case, use of lexical information would be useful in determining the correct orthography for the *e* in *piâderow* since the word is broken down to *piâde* and *row* and the *e* is clearly word final short vowel which is transcribed to ه by the replace rules in the `CR` transducers described earlier.

4. In P-Script, certain words have irregular spelling. For example, the letter و in خویش (self) and خواهر (sister) does not contribute to the pronunciation of these words, and when these words are transcribed into Dabire as *xiš* and *xâhar*, all evidence for the existence of و is lost, and therefore, transcribing these words back to P-Script produces innaccurate answers.

A brief evaluation of the system is summarized as Table-3. The numbers in the third column refer to the above factors. As it is apparent from the table, simpler Persian text (for example, Ferdowsi's epic Shahname) and foreign words written in Dabire are transcribed more accurately to P-Script but texts containing Arabic words and compound words which are common in contemporary texts are more difficult to convert more accurately. It is important to note that, except for words containing Arabic-specific graphemes and compound words that contain sub-words ending with /a/ or /e/, the shortcomings of the conversion system are related to style issues rather than spelling. Most of the these issues can be remedied by using morphological analysis and lexical information and we are currently extending the implementation in this direction. However, the syllabification-based implementation provides a good solution for transcribing new words which are absent in the lexicon.

5. References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

Kenneth R. Beesley. 2007. Arabic morphological analysis and generation. *http://www.arabic-morphology.com/.*

Behdad Esfahbod. 2004. Persian computing with unicode. In *25th Internationalization and Unicode Conference, Washington, DC.*

Sample Words	Percentage of Correctness	Reasons for Loss of Precision
100 English and Swedish names	100%	-
100 words selected from a random page of the Persian epic Shâhnâme	97%	2
100 Arabic words selected randomly from a Persian dictionary	58%	1, 3
100 words selected from a random report on www.bbcpersian.com	83%	1, 2, 3
100 words selected from a book on winter celebrations in Iran	63%	1, 2, 3

Table 3: Precision of the generated transcriptions

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

Forogh Hashabeiky. 2005. *Persian Orthography - Modification or Changeover?* Acta Universitatis Upsalienis.

Joshua Johanson. 2007. Transcription of names written in farsi into english. In Ali Farghaly and Karine Megerdoomian, editors, *Prooceedings of the 2nd workshop on computational approaches to Arabic Script-based languages*, pages 74–80.

Ronald M. Kaplan and Martin Kay. 1994. Regular models of phonological rules systems. *Computational Linguistics*, 20(3):331:378.

Mehdi M. Kashani, Fred Popowich, and Anoop Sarkar. 2007. Automatic transliteration of proper nouns rom arabic to english. In Ali Farghaly and Karine Megerdoomian, editors, *Prooceedings of the 2nd workshop on computational approaches to Arabic Script-based languages*, pages 81–87.

Ahmad Kasravi. 1944. Gâmhâyi dar râh-e alefbâ bar xâhim dâšt. *Zabân-e pâk*, pages 57–64.

Rahim Rezâ Zâde Malek. 2001. *Qavâed e Emlâ ye Fârsi*. Golâb.

Jalal Maleki. 2008. A Romanized Transcription for Persian. In *Proceedings of Natural Language Processing Track (INFOS2008), Cairo*.

Salim Neysari. 1996. *A Study on Persian Orthography - (in Persian)*. Sâzmân e Câp o Entešârât.

John P. Perry. 2005. *A Tajik Persian Reference Grammar*. Brill.

Robert M. Ritter. 2002. *The Oxford Guide to Style*. Oxford University Press.