# Syllable Based Transcription of English Words into Perso-Arabic Writing System

Jalal Maleki
Dept. of Computer and Information Science
Linkping University
SE-581 83 Linkping
Sweden
Email: jma@ida.liu.se

## Abstract

*This paper presents a rule-based method for transcription of English words into the Perso-Arabic orthography. The method relies on the phonetic representation of English words such as the CMU pronunciation dictionary. Some of the challenging problems are the context-based vowel representation in the Perso-Arabic writing system and the mismatch between the syllabic structures of English and Persian. With some minor extensions, the method can be applied to English to Arabic transliteration as well.*

## 1 Introduction

During the translation process from English to Persian certain words (usually names and trademarks) are transcribed rather than translated. This is a general issue in machine translation between language pairs. Unfortunately, there are no guidelines as to how these words should be written in the Perso-Arabic Script (PA-Script) and some words are written in more than 10 different ways ([9] ). This paper introduces a rule-base method for English to PA-Script transcription which is based on the syllable structure of words. Syllables are important since transcription of vowels is mainly determined by the structure of the syllable in which the vowel appears. Given an English word we use a syllabified version of the CMU pronunciation dictionary (CMUPD) to lookup its pronunciation and use it for generating a phonemic romanized Persian transcription of the word which is finally resyllabified and transcribed into the Perso-Arabic Script (PA-Script) according to the syllabification-based method described in [11]. The romanized scheme we use is the Dabire-romanization described in [10]. Since Arabic and Persian essentially use the same script and have the same syllabic structure, our method can easily be extended to the Arabic script.

## 2 Phonological Issues

The essence of our method is phonological mapping between English and Persian and is defined as phonemic mapping of consonents and vowels and resyllabification of the source word using Persian syllable constraints. Just like transliteration between Arabic and English ([2]), transcription between English and Persian is a dfficult task. However, although the mapping between the sounds of Persian and english consonants and vowels is non-trivial, the most complicated step is conversion of Persian vowels to PA-Script [11].

### 2.1 Consonants

Mapping English consonants into Persian phonology is imperfect but straightforward and it can be summarized as a lookup operation. The mapping is however not perfect and in many cases a consonant is mapped into a Persian consonant that only approximately reflects its original pronunciation. For example, /th/ in 'thanks' (/TH, AE1, NG, K, S/) is transcribed to /t/, whereas, the /th/ of 'that' (/DH, AE1, T/) is transcribed to Persian /d/.

### 2.2 Vowels

From a transcription point of view, vowel correspondence between Persian and English phonology is also imperfect and relatively simple. Some examples are shown in Table-1. Some English diphthongs are treated as two separate vowels whereas some others are interpreted as a single vowel.

Phonological mapping is followed by conversion of phonemic romanized Persian to PA-Script. Type of syllable containing a vowel and the characteristics of the neighboring graphemes determine the choice of grapheme (or allographs) for the vowel. As an example, Table-2 shows the various and digraphs used for writing the vowel /i/ in different contexts [11].

### 2.3 Syllable Constraints and Consonant Clusters

Syllable structure in Persian is restricted to (C)V(C)(C), whereas, English allows the more complex structure (C)(C)(C)V(C)(C)(C)(C).

One of the main problems in writing English words in PA-Script is the transformation of syllables. For example, the word 'question' represented as /K, W, EH1, S, CH, AH0, N/ in CMUPD with the syllables /K, W, EH1, S/ and /CH, AH0, N/ is transcribed to *kuesšen* one syllable at a time and finally resyllabified as *ku-es-šen* and transliterated to PA-Script کوئسشن. Resyllabification is necessary since consonant clusters are broken by vowel epenthesis.

In general, the Persian transcription of English words involves short vowel insertion into consonant clusters and resyllabification (See Table-3 for examples.)

## 3 The Implementation

Transcription of an English word *w* into P-Script involves a number of steps which are briefly discussed below.

1. *w* is looked up in the syllabified CMUPD dictionary [4] and its syllabified pronunciation $p(w)$ is retrieved. For example, given the word 'surgical', we get: ((S ER1) (JH IH0) (K AH0 L))
2. Syllables of $p(w)$ are transcribed to Dabire which is a phonemic orthorgraphy for Persian. For the 'surgical', we get *((s e r) (g i) (kâl))*.
3. The syllables are individually modified to fulfill the contraints of Persian syllable structures. For example, *spring* (CCCVCC) is transformed to *espering* (VCCVCVCC) using *e* epenthesis, `prompt` (CCVCCC) is transformed to *perompet* (CVCVCCVC). See Table-3 for more examples.
4. The resulting Dabire word is resyllabified. For example, *espering* is syllabified as *es.pe.ring*
5. Application of context-dependent replace rules [3] to enforce orthographical conventions of Persian [5, 13, 1]
6. Finally, the Dabire-word is transliterated to Perso-Arabic Unicode.

Step 1-3 are currently implemented in Lisp and steps 4-6 are implemented as transducers in XFST [3]

The syllabification step (4) which is one of the main modules of the system is explained further. 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. The correct syllabification would naturally lead to correct writing since as mentioned earlier, vowels are written differently depending on their position in the syllable.

The following XFST-definitions form the core of the syllabification [11]:

```
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 untolerated consonant clusters in NotAllowed such as *bp*, *kq*, and *cc*.

| Vowel | Example Word | Phonemes | Persian Phoneme | Romanized Persian | Perso-Arabic |
|---|---|---|---|---|---|
| AA | odd | AA D | â | âd | آد |
| AE | at | AE T | a | at | ات |
| AH | hut | HH AH T | â | hât | هات |
| AO | ought | AO T | o | ot | اوت |
| AW | cow | K AW | â | kâv | کاو |
| AY | hide | HH AY D | ây | hâyd | هاید |

**Table 1.** *Some Vowels from CMU Pronunciation Dictionary with Examples*

The second statement defines a replacement rule [3] that represents the syllabification process. The operator `@>` ensures that the shortest possible strings (of the form `C* V C*`) are selected in left to right direction and identified as syllables which are separated by a dot.

Table-4 includes examples that illustrate examples of input/output for this.

## 4 Discussion and Evaluation

We have introduced a rule based transcription of English to PA-Script. Earlier work [2, 8, 6, 7] mainly relies on statistical methods.

Our method produces correct transcriptions for most of the data-set randomly selected from CMUPD. Quantitative evaluation of the method is in progress. The performance of the system is dependent on the availability of syllabified English words and future improvements would require use of statistical methods for automatically handling words that do not exist in the dictionary. Some early experiments [14] based on CMUPD show a success rate of 71.6% in automatic grapheme to phoneme conversion of English words not present in CMUPD. Further development would also require integration of automatic syllabification of English [12] into the system.

## References

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

[2] Y. Al-Onaizan and K. Knight. Machine transliteration of names in arabic text. In *ACL Workshop on Computational Approaches to Semitic Languages*, 2002.

[3] K. R. Beesley and L. Karttunen. *Finite State Morphology.* CSLI Publications, 2003.

[4] Carnegie Mellon University. CMU pronunciation dictionary. *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*, 2008.

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

[6] J. Johanson. Transcription of names written in farsi into english. In A. Farghaly and K. Megerdoomian, editors, *Proceedings of the 2nd workshop on computational approaches to Arabic Script-based languages*, pages 74–80, 2007.

[7] S. Karimi, A. Turpin, and F. Scholer. English to persian transliteration. In *Lecture Notes in Computer Science*, volume 4209, pages 255–266. Springer, 2006.

[8] M. M. Kashani, F. Popowich, and A. Sarkar. Automatic transliteration of proper nouns from arabic to english. In A. Farghaly and K. Megerdoomian, editors, *Proceedings of the 2nd workshop on computational approaches to Arabic Script-based languages*, pages 81–87, 2007.

[9] R. R. Z. Malek. *Qavâed e Emlâ ye Fârsi.* Golâb, 2001.

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

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

[12] Y. Marchand, C. R. Adsett, and R. I. Damper. Automatic Syllabification in English: A Comparison of Different Algorithms. *Language and Speech*, 52(1):1–27, 2009.

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

[14] S. Stymne. Private communication. *Linköping*, 2010.

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

**Table 2.** *Mapping /i/ to P-Script Graphemes*

| English | Onset/Coda | Transcription | Example | Clusters |
|---|---|---|---|---|
| /šr/ | Onset | /šer/ | shrink→šerink | /šr/ |
| /s$C_1$/ | Onset | /es$C_1$/ | school→eskul | /sp, st, sk, sm, sn, sl/ |
| /š$C_2$/ | Onset | /eš$C_2$/ | schmock→ešmâk | /šp, št, šk, šm, šn, šl/ |
| /$C_3C_1$/ | Onset | /$C_3$e$C_1$/ | trunk→terânk | /pr, pl, bl, br, .../ |
| /sCw/ | Onset | /esCu/ | squash→eskuâš | /skw/ |
| /sCy/ | Onset | /esCiy/ | student→estiyudent | /spy, sty/ |
| /sC$C_1$/ | Onset | /esCe$C_1$/ | spring→espering | /spl, spr, str, skr/ |
| /$C_1$Cs/ | Coda | /$C_1$Ces/ | corps→korpes | /lps, rps, rts, rks/ |
| /CCCC/ | Coda | /CCeCeC/ | prompts→perâmpetes | |

**Table 3.** *Epenthesis in consonant cluster transcription. $C_1$ stands for all consonants except /w/ and /y/. $C_2$ stands for all consonants except /w/, /y/ and /r/. $C_3$ Stands for all consonants except /s/ and /š/.*

| English<br>Word | CMU<br>Pronunciation | Dabire<br>Romanization | Syllabification | PA-Script |
|---|---|---|---|---|
| GEORGE | JH AO1 R JH | jorj | jorj | جورج |
| BUSH | B UH1 SH | buš | buš | بوش |
| BIOGEN | B AY1 OW0 JH EH2 N | bâyojen | bâ.yo.jen | بایوجن |
| LOUISE | L UW0 IY1 Z | luiz | lu.iz | لوئیز |
| LOUISIANA | L UW0 IY2 Z IY0 AE1 N AH0 | luizianâ | lu.i.zi.a.nâ | لوئیزیئنا |
| INDOSUEZ | IH1 N D OW0 S UW0 EY1 Z | indosuez | in.do.su.ez | ایندو سوئز |
| SPRITE | S P R AY1 T | esperâyt | es.pe.râyt | اسپرایت |

**Table 4.** *Examples showing some of the steps in the transliteration*