International Journal of Information Science and Management

A Proposed UNICODE-Based Extended Romanization System for Persian Texts

M. A. Mahdavi, Ph.D.
Imam Khomeini International University, Iran
Email: mahdavi@researchattic.ca

Abstract

So far, various Romanization schemes have been proposed for capturing Persian text using Latin alphabet. However, each have served a very specific and yet limited function. This paper proposes an extended Romanization scheme that can facilitate a wide range of encoding needed in the field of Natural Language Processing. The proposed scheme endeavors to preserve both orthographic and phonological phenomena in the language. It also accounts for encoding hand-written manuscripts, in which glyph ambiguity is a salient feature. It is particularly relevant to Romanizing the Kufi script, in which diacritical marks are omitted. The current work also recommends orthographic rules in an effort to standardize future Romanization tasks.

Keywords: Romanization System, Persian Text, Natural Language Processing, Written Manuscript.

Introduction

Romanization

The process of capturing non-Roman script languages (such as Persian, Arabic, Hebrew, and Chinese) using Latin characters is called Romanization. Romanization, on one hand, pertains to the act of transliteration, which is the writing of a language using Latin characters. Transcription, on the other hand, is another form of Romanization that captures the speech utterances in form of written text using Latin characters. While transliteration remains faithful to capturing the orthography, transcription is mainly concerned with the phonographic features of a language.

Ambiguity in non-vocalized form

The non-vocalized Persian texts are highly ambiguous due to multiple pronunciations. Perhaps a less ambiguous writing scheme alleviates the burden of heavy and complex disambiguation algorithms. However, what is more appealing than having an unambiguous writing scheme would be an unambiguous transliteration scheme. Much research is done

using transliterated Persian texts, yet a comprehensive transliteration scheme is long overdue. There is a dire need for an extended transliteration system that covers the full range of letters, glyphs, and sounds. Furthermore, the conventions for rendering empirical transliterations are also missing.

Compiling a comprehensive corpus of transliterated Persian text requires a reliable system to preserve the orthographic as well as the phonological features of the language. This paper offers an alternative scheme to capture Persian text in Roman alphabet.

Key aim, stating the problem

There is an increasing need to write Persian text using Latin characters. Pieces of Persian text are frequently transliterated using Romanization. The aim of this paper is to introduce a transliteration system which provides an unambiguous one-to-one mapping between Latin characters in the UNICODE range and each Arabic script character. In other words, the Romanization would have to provide a reversible conversion. This paper proposes a Romanization system that is able to preserve both the pronunciation and the written forms of the text.

Literature on Existing Romanization Systems

Several Romanization systems for Persian have been introduced so far. Among many, one may mention those of *Encyclopedia Iranica*[1], *Encyclopedia of Islam(EI:1960)*[2], the American Library Association - Library of Congress (ALA-LC:1997)[3], the United Nations (UNGEGN:1972)[4], *Deutsche Morgenländeishe Gesellschaft* (DMG:1969), Deutches Institut für Normung standard (DIN 31 635:1982)[5], Board on Geographic Names (BGN/PCGN:1946,1958)[6], International Civil Aviation Organization (NTWG:2008)[7], the British Standard (BS 4280:1968), Buckwalter (Xerox)[8,9], FarsiTeX[10], UniPers[11], EuroFarsi[12], Dehdari[13], Maleki (Dabire)[14], The CJK Dictionary Institute (CJKI)[15], Standard Arabic Technical Transliteration System (SATTS), ASMO 449, ECMA[16], and International Standard Organization (ISO 233_3:1999)[17].

These Romanization systems are either based on phonology or orthography. Some Romanization schemes (such as UniPers[18] and EuroFarsi) focus primarily on the sound of the utterances rather than the orthographic variations of each sound. While this approach simplifies the transcription effort, it introduces further ambiguity in reproducing the original orthography.

In cases of some formal transliteration systems such as *Encyclopedia Iranica* and *Encyclopedia of Islam (EI^2)*, the notation preserves some aspects of the orthography while attempting to vocalize the words. The shortcoming of such transliteration systems is that, sometimes pronunciation distinction is asserted (as is the case with TEH MARBUTA ("ۀ")), while it is ignored elsewhere (such as the case for the definite article *al* "ال").

Every Romanization scheme seems to have served a specific purpose. In some cases, they are devised to transliterate geographic names (e.g. BGN/PCGN), personal names (e.g. NTWG), morphological analyses (e.g. Buckwalter, Dehdari), or rendering pronunciations (e.g. UniPers). For NLP purposes, however, a much broader Romanization scheme is needed to meet various processing requirements. For instance, a generalized scheme is required to encode hand-written manuscripts, in which there is an abundance of ambiguous glyphs. In other words, a more extended set of Roman characters is needed to encode scripts that are missing diacritical marks and dots such as the Kufi script. Other areas of NLP such as encoding syllabification patterns and text-to-speech encoding may benefit from such broad and generalized Romanization system.

A generalized Romanization scheme is also needed to capture an entire text or build an extensive corpus. In the early works done on Persian and Arabic NLP, a customized set of ASCII characters has been adopted to develop morphological analyzers and parsers. Xerox, for instance, has adopted Buckwalter transliteration[19] scheme which is a set of 7-bit ASCII characters representing the full range of Arabic characters. Jon Dehdari has adopted a similar ASCII set to develop a Link Grammar parser for Persian.

While machine readable, these customized transliteration systems are not conducive to large scale corpus collection. The fundamental issue with such systems is that, while they provide a one-to-one mapping to the alphabet, they are not human-readable. For instance, Buckwalter system uses an asterisk (*) for the Arabic letter THAL ("ذ"). The overwhelming abundance of punctuation marks and non-alphabetic characters such as **"{, <, $, &, |, _, ~, >, }**" in the Buckwalter transliteration makes the text extremely unreadable by a human reader. Dehdari's transliteration offers an improved legibility compared to Buckwalter's. However, they are both case sensitive, which make the text less readable to a general reader when uppercase and lowercase of a letter mean different letters.

FarsiTex has also been used as a typographic convention to encode Persian text. Like ArabTex, FarsiTex is primarily concerned with encoding the written form of the language. Unlike Buckwalter transliteration system, FarsiTex convention is slightly more readable by humans. However, this readability is achieved through multi-character codes for encoding Persian letters. For instance, the Arabic letter SHEEN ("ش") is encoded by a circumflex accent followed by the Latin letter S (^s).

This method is very similar to the popular transliteration system used in the field of oriental studies, which utilizes "sh" to encode the Arabic letter SHEEN ("ش"), for example. However, this does not provide a reversible one-to-one mapping. For instance, the word TAMĀSHĀ would result in two distinct Persian words of TAMĀS-HĀ ("تماس‌ها" meaning contacts) and TAMĀSHĀ ("تماشا" meaning to watch). A multi-character transliteration system is also produced for encoding personal information on the machine readable travel documents. In this scheme, NTWG uses an escape character (letter X) for the

disambiguation purposes. Although the ambiguity issue is addressed, the readability by human seems to have been sacrificed.

What these systems have in common is their effort to capture Persian and Arabic writings using Latin characters. Despite the differences in their choice of characters for corresponding Arabic letters, almost all of them lose part of the textual information in the process, *albeit*, the phonetic realization or its glyph representation. This paper proposes a generalized Persian Romanization scheme that tries to remain faithful to both orthographic and phonological aspects of the Persian language.

Discussing the Principles

The underlying principle in this study is to achieve an unambiguous Romanization whereby each Persian letter is represented by a single and unique Latin letter. In addition to orthography, the Romanization scheme should provide a phonological clue to how a word is pronounced. Thus, the selection process follows a set of criteria, which would act as the guiding principles. The selection criteria are described as follows:

(P1) Every letter in the Romanization scheme should be captured by a single UNICODE representing a unique character. Combination of UNICODE characters should be avoided.

(P2) The non-language diacritical marks, which do not participate in the writing of the language, are not part of the Romanization. For instance, the notational guides for reciting the Quran are not considered as part of the alphabet. However, short vowels, MADDA ("ٓ"), SHADDA ("ّ"), SUKUN ("ْ"), and superscript ALEF ("ٰ") are counted as part of the alphabet.

(P3) Phonological choices would take precedence over glyph shapes. In other word, in cases where a Persian letter has a phonological correspondence in Latin, the character chosen should follow the phonological resemblance as its base character. For instance, although TEH MARBUTA ("ة") looks like a final HEH ("ه"), the corresponding Latin character should be a variant form of the letter "T" rather than "H" because it is pronounced the same as Latin letter T.

(P4) In cases where Persian and Arabic pronunciations differ, preference is given to the Persian pronunciation. Thus, the Latin character chosen should resemble the Persian phonological realization rather than the Arabic one. For instance, the pronunciation of the Arabic letter DAD ("ض") is uttered differently in Persian. While some Arabic transliteration systems use Latin letter "D" as the base character, for Persian pronunciation, the base character should be Latin letter "Z".

(P5) If several Persian letters have the same phonological realization, the alternative glyph forms should be captured using Latin diacritical mark. For instance, letters THAL ("ذ"), ZAIN ("ظ"), DAD, and ZAH ("ز"), in Persian, are all realized phonologically the

same as the letter "Z". In this case, one of them is encoded as "Z", while others would take alternative diacritical forms of the letter "Z".

(P6) Characters that have no visual realization, but their presence is implied, should also be given an equivalent Latin character. For instance, the non-spacing UNICODE characters ZERO WIDTH NON-JOINER (ZWNJ), and ZERO WIDTH JOINER (ZWJ) which affect the glyph shape of the previous character should be given an equivalent Latin character.

(P7) Grammatical markers that are not written but their presence is conventionally implied such as the third person marker for the perfect form of the verb should be assigned a Latin character. This is particularly applicable to morphological and grammar rules.

(P8) Although Persian writing system lacks a character for marking the syllabic stress on words, the envisaged Romanization system should reserve a character to mark the stress. The stress mark can be a combining diacritic such as the acute accent.

(P9) The radical forms of letters used in archaic writing and classical texts (i.e. "□", "ف", "□", "ك", and "ں") should each get a unique Latin character. These characters facilitate the encoding of the Kufi script. An extension of this principle is also applied to the ambiguous base forms such as "ر", "س", "ص", "ط", and "ح" that may have multiple interpretations in some Kufi writings.

(P10) The previous principle is also extended to the ambiguous forms such as "ز", "ب", "ج", "ک", which in some hand-written manuscripts are used to represent "ژ", "پ", "چ", and "گ" respectively.

(P11) If a Persian letter is written but not pronounced, a unique Latin character must be assigned to mark the non-vocal property of the glyph. In other words, silent letters should be treated as a phonological realization. For instance, one phonological realization of letter WAW in Persian is a silent character that is written but not pronounced, as in the word خواهر (meaning sister). In this case, a unique Latin character should be assigned to mark its non-vocal.

Selection Process

In addition to meeting the selection criteria, efforts have been made to follow a logic that achieves a consistent semantic for the use of Latin diacritic. In other words, the selection process faces a challenge to find a set of rules that can govern the way in which diacritics are applied to the base characters. This is because the UNICODE system does not captured the full range of Latin diacritical forms using single codes. For most cases in the UNICODE range, the diacritical forms are achieved by combining the base letters with the diacritics. Unless private ranges of the UNICODE are used to define custom characters, it is very difficult to achieve consistent diacritics for all of the base consonants. For instance, UNICODE does not contain a single code for the Latin letter Q/q with a dot above.

Nonetheless, efforts have been made to standardize the use of diacritics in such a way that a consistent semantic is implied.

The followings are the rules for capturing some of the orthographic and phonological phenomena in Persian alphabet.

(SP1) In cases where a character is written but not pronounced, the Latin combining CARON is used to denote that the character has the same glyph shape as the base character but it is not pronounced. For instance, Ǎ, Ȟ, Ǐ, Ǔ, T̰, and L̰ are all written but not pronounced. The Latin letter "T" and "L" with a CARON above are not in the UNICODE. Instead, a combining CIRCUMFLEX below is used to represent the voiceless TEH MARBUTA ("ة") and the unvoiced letter "L" ("ل") of the definite article "AL" in words beginning with sun letters.

(SP2) Whenever an Arabic letter HAMZA ("ء") is placed above other base letters, the Latin combining DIAERESIS should be used to mark the existence of a HAMZA ("ء") above the base character as in the cases of Ӛ, Ḧ, Ä, Ë, Ö, Ẅ, Ÿ and Ï.

(SP3) Any vowel that has an elongated duration should be represented by the Latin combining MACRON above the base vowel. These vowels are Ā, Ō, Ē, Ī, and Ū.

(SP4) There are a few characters in Arabic script that are the base forms and act as the seat for diacritical dots and strokes. These occurrences should be represented using a combining DOT above or below the base character in their corresponding Latin representation. These characters are Ḃ, Ḣ, Ḟ, Ṅ, and Ḳ. In the case of the base form of the letter "Q", since the Latin character "Q" with a dot above is not part of the UNICODE, the letter K with a dot below is selected instead to respect the rule for using a single code for each character.

(SP5) In classical Persian texts, letters P ("پ"), G ("گ"), ZH ("ژ"), and CH ("چ") are sometimes written as B ("ب"), K ("ک"), Z ("ز"), and J ("ج"). While these characters are not in the base form, the rule applied to these would be the same as the base forms. These characters are Ṗ, Ġ, Ż, and Ċ.

(SP6) In some sources, the combination of letters KHEH ("خ") followed by a silent WAW ("و") is considered as a single letter KHO ("خو"). This letter is believed to have existed in the old Persian. However, in modern Persian, they are treated as two distinct letters. As a means to accommodate both cases, the Latin character "X" with DIAERESIS "Ẍ" is chosen to provide a single character and the combination of letter "X" followed by Ŵ is used to represent the two character alternation.

Character Chart

The following is the full range of the Romanized character set representing the Persian characters. Lower case letters have been removed from this table.

Table 1

Character chart representing the proposed Romanization

LAT	LATIN HEX Code	UNICODE_NAME	Example
ː	2D0	MODIFIER LETTER TRIANGULAR COLON	ZWNJ
ˑ	2D1	MODIFIER LETTER HALF TRIANGULAR COLON	ZWJ
Ø	D8	LATIN CAPITAL LETTER O WITH STROKE	تهی
'	27	APOSTROPHE	Accent
°	B0	DEGREE SIGN	سکون
²	B2	SUPERSCRIPT TWO	تشدید
³	B3	SUPERSCRIPT THREE	مدّه
A	41	LATIN CAPITAL LETTER A	فتحه
E	45	LATIN CAPITAL LETTER E	کسره
O	4F	LATIN CAPITAL LETTER O	ضمه
Ȁ	200	LATIN CAPITAL LETTER A WITH DOUBLE GRAVE	فتحتین
Ȅ	204	LATIN CAPITAL LETTER E WITH DOUBLE GRAVE	کسرتین
Ȍ	20C	LATIN CAPITAL LETTER O WITH DOUBLE GRAVE	ضمتین
Ӭ	4EC	CYRILLIC CAPITAL LETTER E WITH DIAERESIS	حمزه
Ä	C4	LATIN CAPITAL LETTER A WITH DIAERESIS	أ
Ë	CB	LATIN CAPITAL LETTER E WITH DIAERESIS	إ
Ö	D6	LATIN CAPITAL LETTER O WITH DIAERESIS	أ
Ï	CF	LATIN CAPITAL LETTER I WITH DIAERESIS	إی
Ü	DC	LATIN CAPITAL LETTER U WITH DIAERESIS	أو
Å	C5	LATIN CAPITAL LETTER A WITH RING ABOVE	Superscript Alef ٰ
Ǎ	1CD	LATIN CAPITAL LETTER A WITH CARON	□
Ā	100	LATIN CAPITAL LETTER A WITH MACRON	ا
Ã	C3	LATIN CAPITAL LETTER A WITH TILDE	آ
Ḃ	1E02	LATIN CAPITAL LETTER B WITH DOT ABOVE	Base form □
B	42	LATIN CAPITAL LETTER B	ب
□	1E56	LATIN CAPITAL LETTER P WITH DOT ABOVE	ب
P	50	LATIN CAPITAL LETTER P	پ
T	54	LATIN CAPITAL LETTER T	ت
□	1A7	LATIN CAPITAL LETTER TONE TWO	ث
J	4A	LATIN CAPITAL LETTER J	ج
Ç	C7	LATIN CAPITAL LETTER C WITH CEDILLA	چ
Ċ	10A	LATIN CAPITAL LETTER C WITH DOT ABOVE	Ambiguous (چ) ج

LAT	LATIN HEX Code	UNICODE_NAME	Example
□	1E28	LATIN CAPITAL LETTER H WITH CEDILLA	ح
□	1E22	LATIN CAPITAL LETTER H WITH DOT ABOVE	Ambiguous (خ) خ̣
X	58	LATIN CAPITAL LETTER X	خ
□	1E8C	LATIN CAPITAL LETTER X WITH DIAERESIS	خو
D	44	LATIN CAPITAL LETTER D	د
□	1E0A	LATIN CAPITAL LETTER D WITH DOT ABOVE	Ambiguous (د) ذ
□	1B8	LATIN CAPITAL LETTER EZH REVERSED	ذ
R	52	LATIN CAPITAL LETTER R	ر
□	1E58	LATIN CAPITAL LETTER R WITH DOT ABOVE	Ambiguous (ر) ṙ
Z	5A	LATIN CAPITAL LETTER Z	ز
Ĵ	134	LATIN CAPITAL LETTER J WITH CIRCUMFLEX	ژ
Ż	17B	LATIN CAPITAL LETTER Z WITH DOT ABOVE	Ambiguous (ژ) ż
S	53	LATIN CAPITAL LETTER S	س
Ş	15E	LATIN CAPITAL LETTER S WITH CEDILLA	ش
Ś	15A	LATIN CAPITAL LETTER S WITH ACUTE	ص
Ź	179	LATIN CAPITAL LETTER Z WITH ACUTE	ض
□	1E6E	LATIN CAPITAL LETTER T WITH LINE BELOW	ط
□	1E94	LATIN CAPITAL LETTER Z WITH LINE BELOW	ظ
□	186	LATIN CAPITAL LETTER OPEN O	ع
Ğ	11E	LATIN CAPITAL LETTER G WITH BREVE	غ
F	46	LATIN CAPITAL LETTER F	ف
□	1E1E	LATIN CAPITAL LETTER F WITH DOT ABOVE	Base form ف
Q	51	LATIN CAPITAL LETTER Q	ق
□	1E32	LATIN CAPITAL LETTER K WITH DOT BELOW	Base form □
K	4B	LATIN CAPITAL LETTER K	ك
G	47	LATIN CAPITAL LETTER G	گ
Ġ	120	LATIN CAPITAL LETTER G WITH DOT ABOVE	Ambiguous گ
□	1E3C	LATIN CAPITAL LETTER L WITH CIRCUMFLEX BELOW	Voiceless ل
L	4C	LATIN CAPITAL LETTER L	ل
M	4D	LATIN CAPITAL LETTER M	م
□	1E44	LATIN CAPITAL LETTER N WITH DOT ABOVE	Base form ں
N	4E	LATIN CAPITAL LETTER N	ن
□	21E	LATIN CAPITAL LETTER H WITH CARON	ه

LAT	LATIN HEX Code	UNICODE_NAME	Example
□	1E26	LATIN CAPITAL LETTER H WITH DIAERESIS	ۀ
□	1E70	LATIN CAPITAL LETTER T WITH CIRCUMFLEX BELOW	ة
□	1E6A	LATIN CAPITAL LETTER T WITH DOT ABOVE	ة
H	48	LATIN CAPITAL LETTER H	ه
Ŵ	174	LATIN CAPITAL LETTER W WITH CIRCUMFLEX	Voiceless و
Ō	14C	LATIN CAPITAL LETTER O WITH MACRON	Short Vowel و
Ẅ	1E84	LATIN CAPITAL LETTER W WITH DIAERESIS	ؤ
Ū	16A	LATIN CAPITAL LETTER U WITH MACRON	Long Vowel و
W	57	LATIN CAPITAL LETTER W	Arabic (W) و
V	56	LATIN CAPITAL LETTER V	Persian (V) و
□	1E00	LATIN CAPITAL LETTER A WITH RING BELOW	Alef Maqsura ى
Ē	112	LATIN CAPITAL LETTER E WITH MACRON	Short Vowel ى
□	4F0	CYRILLIC CAPITAL LETTER U WITH DIAERESIS	ئ
Ī	12A	LATIN CAPITAL LETTER I WITH MACRON	Long Vowel ي
Y	59	LATIN CAPITAL LETTER Y	Consonant ى
·	B7	MIDDLE DOT	Syll. Mark

Standardization

The Romanization table, outlined in previous sections, facilitates the capturing of a range of texts, from modern to classical. However, for practical reasons, a standard system of Romanization is recommended. In what follows, a series of conventions are introduced to provide a guideline for standard Romanization. The rules are written in such a way that the content of hand-written manuscripts may also be captured using the standard Romanization.

Morpheme Boundaries

(R1) ZERO WIDTH NON-JOINER (ZWNJ): In Persian orthography, sometimes, it is necessary to force the final context of a letter. In Persian texts, ZWNJ character is used to mark free morpheme boundaries. For instance, when the noun "ایران" is prefixed to another noun "زمین", the new word "ایرانزمین" is formed, in which the morpheme boundary between the original nouns is marked using the ZWNJ character in UNICODE. For Romanization purposes, it is recommended to use ZWNJ consistently to mark the free morpheme boundaries. However, the junction between bound and free morphemes should not be

marked by ZWNJ.

(R2) ZERO WIDTH JOINER (ZWJ): The use of word delimiters such as space and other non-alphabetic characters forces the final context of the letters. However, if a non-final context is desired despite the present of word delimiter, ZWJ is used to force the non-final presentation form. The use of ZWJ marker in the standard Romanization system is recommended to mark acronyms and the morpheme boundaries where a bound morpheme attaches to its following morphemes.

Combining Diacritics

(R3) SUKUN ("ْ"): If the original text contains the SUKUN diacritic over a character, the Romanized text should use the DEGREE SIGN to mark the presence of the SUKUN in the original text. However, in a standard Romanization system, the use of SUKUN should be avoided.

(R4) SHADDA ("ّ"): In cases where the original text has used a SHADDA to mark the doubling of a consonant, the Romanized text should also use the superscript number 2 to mark the existence of the SHADDA diacritic in the original text. However, for standardization purposes, the doubling of a consonant should be captured by writing it twice.

(R5) MADDA ("~"): With the exception of the overlong ALEF ("آ"), where it is part of the glyph, MADDA is used as a combining diacritic in Persian texts. For the purposes of preserving the original text, the combining diacritic MADDA can be represented by the superscript digit ("^{3}"). In the standard form, MADDA is almost never present.

(R6) The ALEF WITH MADDA ABOVE ("آ") is a contracted form of HAMZA ("ء") followed by a long ALEF ("ا"), for standard Romanization, the contracted form Ã should be avoided. Instead the non-contracted combination □Ā should be used. For example, the word ÃMADAM for the word "آمدم" is the non-standard form, whereas □ĀMADAM is the standard Romanization.

(R7) THIRD PERSON MARKER: Third person singular marker for the perfect tense of verbs is a convention that does not have visual manifestation in Persian texts. For the purposes of standard Romanization, the Latin letter "O" with a stroke "Ø" should only be used in grammatical or lexicographical contexts. Standard Romanized text should not contain this marker.

Consonants

HAMZA ("ء"): In Persian orthography, the letter HAMZA is borrowed from Arabic. While the current Romanization set accommodates any combination of orthographic forms of HAMZA, for standardization purposes, the use of HAMZA may be simplified.

(R8) HAMZA in Arabic script can occur either in free-standing form or it may be carried by other letters of the alphabet, which act as a seat for the HAMZA. There are

several letters that can be used as a seat for HAMZA; they are ALEF ("أ" and "إ"), YEH ("ئ"), WAW ("ؤ"), and sometimes HEH ("ۀ"). The rules governing the seat of HAMZA differ depending on the grammar source used. However, for the purpose of a standard Romanization, the free standing HAMZA should be used. For example, to transliterate the word "تأمين", TAƏMĪN should be used instead of TÄMĪN, for the word "مسؤول", MASƏŪL should be used instead of MAS□ŪL, and for the word "مؤثر", MOƏA□□ER should be used instead of MOẄA□□ER.

(R9) Since Persian words may not begin with vowels, the initial HAMZA in words is almost always carried by the letter ALEF. The combination of initial HAMZA with its following vowel may culminate in a range of variants. As mentioned above, for the purposes of preserving the original orthography, alternative forms of HAMZA may be used. However, for standardization purposes, the isolated form of HAMZA should be used in conjunction with vowels. For example, ЄŪ should be used instead of ÄŪ/Ü for "أو", □OMĪD instead of ÖMĪD/ÄOMĪD for "أميد", □ENSĀN instead of ËNSĀN/ËENSĀN for "إنسان".

(R10) Unvoiced HEH ("ه" in "خانه"): In Persian some words end in a non-vocal "H" as in XĀNA□ or XĀNE□ ("خانه" meaning "home"). Since the letter HEH ("ه") has a visual realization, the Romanization should also give that indication. Thus, the unvoiced letter "□" is always written in the standard form.

(R11) Unvoiced WAW ("و" in "خواندن"): In cases where WAW is written but not pronounced, the standard Romanization should capture this silent WAW using unvoiced letter "Ŵ".

(R12) Unvoiced YEH ("ی" in "موسی"): In one case, the current Romanization system differs from the UNICODE interpretation. In the UNICODE characters, the Arabic letter YEH (U+0649) is introduced to mark ALEF MAKSURA ("ٰ"), as in the case of the proper noun □ĪSĀ. However, the current Romanization system interprets the letter "ی" as the silent seat for carrying ALEF MAKSURA, as in the ending of □ĪS□ "عیسیٰ". Here, rather than using two glyphs to represent a single phenomenon, "□" represents the silent YEH "ی". In other words, in these cases the superscript ALEF ending of the words is implicit. Other examples are: BAN□ "بنی", □EL□ ("إلی"), and □AL□ ("علی").

(R12) According to the morphotactic rules, a suffix may cause the termination of ALEF MAKSURA to either change to a long ALEF or a diphthong YEH as in words. For instance, RAM□ ("رمی") becomes RAMĀHĀ ("رماها") after adding the suffix HĀ at the end. Also, in the case of □EL□ ("إلی"), the ending changes to □ELAYHĀ ("إليها"). Although Persian language does not have extensive use of such examples – since the words are themselves borrowed from Arabic – the same rule should apply in Persian Romanization.

(R13) Definite article ("ال"): In Persian, all nouns are definite by default. Otherwise,

an indefinite suffix is used to mark the explicit indefiniteness. In Arabic, however, besides the definite case ending, an ALEF followed by a LAM is used to mark the definiteness. In possessive (EZAFEH) and adjectival constructions, the definite ALEF may be written as the ALEF WASLA. In these cases the letter Ă is used to mark the unvoiced status of the ALEF. In other cases, it is just a common usage of the initial HAMZA which should be written as an isolated HAMZA followed by a vowel. For instance, □ALKETĀB (“ألكتاب”) is a definite noun that does not participate in a construction, while FĪ ĂLKETĀB (“فی الكتاب”) is a construction in which ALEF is not voiced (Ă).

(R14) The letter L in the definite marker is pronounced differently depending on the context of the next letter. When definite article is followed by the sun letters, the letter “L” loses its voice and doubles the following letter. In these cases, the Latin character “□” is used to indicate that the “□” should be treated as the same sound as its following letter. For instance, □A□ŞAMS (“ألشمس”) and FĪ Ă□ŞAMS (“فی الشمس”) are two of such cases, where “□“ is pronounced as “Ş” causing a letter doubling of Ş.

(R15) Damped YEH (“ی”): There is another variant in Persian for the long vowel “Ī”. In some dialects, this long vowel is subsumed by the previous short vowel “E”, hence producing a slightly longer and more emphasized E. Although it is written as YEH “ی”, it is pronounced “E”. To represent this context the Latin letter “Ē” is used. For instance, “میرود” is transliterated as MĒRAVAD rather than MIRAVAD.

(R16) Damped WAW (“و”): In some contexts, the letter WAW gets consumed by the previous letter “O”. In this case, the Latin letter “Ō” is used to denote the damped WAW. For instance, “گور” is transliterated as GŌR rather than GŪR.

(R17) Unvoiced TEH MARBUTA (“ة”): In cases where TEH MARBUTA is not pronounced, the Latin letter □ should be used instead of the normal TEH MARBUTA “□”. For instance, “حجة” should be transliterated as □OJJA□ rather than □OJJA□.

Capitalization

For standardization purposes, the English language capitalization rules should be adopted here. Library of Congress, as well as other major schemes adopted by the scholarly community, treats the definite article ALEF LAM as an exception to this rule. The proposed Romanization system, however, does not make this distinction for the definite article.

Conclusion

Several Romanization schemes are currently used by the scholarly community. Almost all of the current schemes fall under two major categories. The first category is the transliteration scheme, which emphasizes the written form of the language. These schemes simply capture the way a text is written. The second category, on the other hand, is the

transcription scheme, which focuses on how a text is read. Very few have attempted to achieve a combination of the two methods. None of them, however, lend themselves to a large scale Romanized corpus collection of Persian texts.

The Romanization system proposed here will provide a comprehensive set of Latin characters that will provide orthographic as well as phonological representation for Persian writing system. The UNICODE system may not yet be supported by many fonts and applications. However, for the language research community, it will provide a sound system to capture any form of text, be it classical or modern. In other words, in addition to providing an unambiguous method to produce an extensive tagged corpus for Persian NLP, the current transliteration scheme also provides a range of encodings for capturing hand-written manuscripts.

The proposed method here uses diacritics consistently. It also provides a set of rules to standardize the transliteration process. While it may not be an immediate alternative to replace the existing transliterated texts, it is certainly a suitable alternative to capture extensive Persian texts. It owes its strength to its readability and its unambiguous use of the glyphs.

Future Works

The success of this project is only secured if extensive resources are provided for employing the proposed scheme. The resources needed for the use of this transliteration system include a variety of UNICODE fonts and an input method such as a keyboard method. To achieve an efficient keyboard layout, a data-intensive method may be used to suggest the best key layout based on the frequency of glyph usage in Persian texts.

It is also envisaged that a series of converters are needed to automatically convert Persian texts to their standard transliterations. Conversion tools may also be required to assist with convert existing transliterated texts to the current scheme.

Endnotes

1. http://www.iranica.com/
2. http://www.brill.nl/default.aspx?partid=227&pid=7560
3. The document was last revised in 1997. http://www.loc.gov/catdir/cpso/romanization/persian.pdf
4. A special publication dated 1946. http://www.un.org/depts/dhl/maplib/ungegn/session-1/misc/joint-rules.pdf
5. http://en.wikipedia.org/wiki/DIN-31635
6. http://unstats.un.org/unsd/geoinfo/gegn22wp54.pdf
7. (International Civil Aviation Organization, 2007)
8. http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/translit-chart.html

9. http://www.xrce.xerox.com/competencies/content-analysis/arabic-inxight/arabic-surf-lang-unicode.pdf
10. http://www.tug.org/TUGboat/Articles/tb23-1/farsitex.pdf
11. http://www.unipers.com/up.htm#alphabet
12. http://www.eurofarsi.com/n_alpha.html
13. http://www.ling.ohio-state.edu/~jonsafari/persian_charmaps.pdf
14. (Maleki, A Romanization Transcription for Persian, 2008)
15. (Halpern, 2009)
16. http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-114.pdf
17. The last revision was approved in 1999, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=2398.
18. http://www.unipers.com/up.htm#alphabet
19. http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/translit-chart.html. A revised version of his transliteration is published at http://www.qamus.org/transliteration.htm.

Bibliography

Brill. (1960). *Encyclopaedia of Islam* (New Edition ed.). (T. B. P.J. Bearman, Ed.) Leiden: Brill.

Congress, L. O. (1997). *ALA-LC Romanization Table for Persian*. Retrieved on October 31, 2010 from http://lcweb.loc.gov/catdir/cpso/persian.pdf

Encyclopaedia Iranica. (1996). *Encyclopaedia Iranica, Online Edition*. Retrieved on October 31, 2010 from Encyclopaedia Iranica: http://www.iranica.com/pages/citing-iranica

EuroFarsi. (1996). *EuroFarsi convention.* Retrieved on October 31, 2010, from EuroFarsi: http://www.eurofarsi.com/

Farhangestan Zaban va Adabiyat-e Farsi. (1384). *Dastur-e Khatt-e Farsi (Orthographic converntions for Persian)*. Retrieved on October 31, 2010 from the Persian Academy: http://www.persianacademy.ir/fa/dastoorpdf.aspx

Halpern, J. (n.d.). *CJKI Arabic Romanization System*. Retrieved on October 31, 2010 from the CJK Dictionary Institute, Inc.: www.kanji.org/cjk/arabic/cars/cars_paper.pdf

International Civil Aviation Organization. (2007). *Technical Advisory Group on Machine Readable Travel Documents*. (N. T. (NTWG), Producer, & Technical Advisory Group on Machine Readable Travel Documents) Retrieved on October 31, 2010 from Internationa Civil Aviation Organization, Air Transport Bureau: http://www.icao.int/icao/en/atb/sgm/mrtd/TAG_MRTD17/TagMrtd17_WP019.pdf

International Standard Organization. (1984). *ISO 233:1984. Documentation* --

Transliteration of Arabic characters into Latin characters. (International Standard Organization) Retrieved on October 31, 2010 from International Standard Organization: http://www.iso.org/iso/catalogue_detail.htm?csnumber=4117

Maleki, J. (2008). A Romanization Transcription for Persian. *Proceedings of International Conference on Intelligent Information and Engineering Systems* , 166-175.

Maleki, J. (2003). *EFarsi - A Latin-Based Writing Scheme for Persian*. Retrieved on October 31, 2010 from Linkoping University: www.ida.liu.se/~jalma/efarsi.pdf

PCGN. (1958). *Romanization System for Persian (Dari and Farsi)*. Retrieved on October 31, 2010 from The Permanent Committee on Geographical Names: http://earth-info.nga.mil/gns/html/Romanization/Romanization_Persian.pdf

UN. (2003). *Eighth United Nations Conference on the Standardization of Geographical Names.* New York: The United Nations.

UniPers. (n.d.). *UniPers: A 21st Century Alphabet for the Persian Language*. Retrieved on October 31, 2010 from UniPers: http://unipers.com/

United Nations Group of Experts on Geographical Names (UNGEGN). (2003). *Report on the Current Status of United Nations Romanization Systems for Geographical Names*. Retrieved on October 31, 2010 from United Nation's Statistics Division - Geographical Names and Information Systems: http://www.eki.ee/wgrs/rom1_fa.htm

Wikipedia, the free encyclopedia. (n.d.). *Romanization of Arabic*. Retrieved on October 31, 2010 from Wikipedia, the free encyclopedia: http://en.wikipedia.org/wiki/Romanization_of_Arabic