Transliteration editors for Arabic, Persian and Urdu

E.Veera Raghavendra, Prahallad Lavanya, Fahmy Mostafa
Carnegie Mellon University
IIIT Hyderabad, India.

Abstract: Transliteration editors are essential for keying-in language scripts into the computer using QWERTY keyboard. Applications of transliteration editors in the context of Universal Digital Library (UDL) include entry of meta-data and dictionaries for many languages both local and International. In this paper we propose a simple approach for building transliteration editors for International languages such as Arabic, Persian and Urdu using Unicode and by taking advantage of its rendering engine which is called Unicode rendering engine. We demonstrate the usefulness of the Unicode based approach to build transliteration editors for International Languages, and report its advantages needing little maintenance and few entries in the mapping table, and ease of adding new features such as adding letters, to the transliteration scheme. We also explain how easy it is to add any language and build a transliteration editor using Unicode and its mapping tables. We demonstrate the transliteration editor for 3 International languages and also explain how this approach can be adapted for any foreign language.

Keywords: Transliteration Editors, IT3, Arabic, Persian and Urdu, UDL.

1. Introduction

Several processes exist in realizing the concept of universal digital library (UDL). These processes include scanning of the books, improving the quality of scanned images, entry of meta-data of the scanned books, storage of the data, retrieval and access to the data as and when required.

1.1 Need for Transliteration Editors

Transliteration editors are essential for keying-in Indian language scripts into the computer using QWERTY keyboard. Applications of transliteration editors in the context of Universal Digital Library (UDL) include entry of meta-data and dictionaries for International languages says it all, In the Universal Digital Library, many international language books are scanned and the need to enter the metadata in their local languages becomes necessary. To enter the text in their local languages other than English is possible only through Transliteration editors. These editors let you enter the language you chose using the QWERTY language scripts like Arabic, Persian and Urdu.

1. 2 How are they useful in digital library?

Applications of transliteration editors in the context of Universal Digital Library (UDL) include entry of meta-data and dictionaries for Indian languages. The issues in building transliteration editors include design of a user-friendly and readable transliteration scheme, user interface to key-in the text and have the text rendered in native script, provide

transliteration code for the characters in International languages.

1.3 Are there any previous Transliteration Editors

There are many transliteration editors developed for many languages all over the world for example we have Indian language Transliteration editors such as Indian Unitrans and Om transliteration editor found at http://speech.iiit.ac.in/~speech/Transliteration/ http://www.cs.cmu.edu/~madhavi/OM The above editors mentioned also use IT3, but the only drawback for the above editors is they support only Indian languages and not extended to support any other foreign language. Since Universal digital library includes the scanning and Meta data entry of International languages like Arabic, Persian and Urdu, there is a need to develop a stable Transliteration editor where the Meta data entry can be easy and effective. All the local people who speak Arabic, Persian and Urdu can see their books in their own language.

2. Nature of Arabic, Persian and Urdu Scripts

Arabic, Persian and Urdu often called as Middle East Languages have many features in common. The nature of the script for all the three languages is mostly same and as described as follows:

- The Arabic alphabet contains 28 letters. Some additional letters are used in Persian and Urdu such as /p/ or /g/.
- Words are written in horizontal lines from **right to left**, numerals are written from left to right
- Most letters change form depending on whether they appear at the beginning, middle or end of a word, or on their own.
- The Arabic, Persian and Urdu script is cursive, and all primary letters have conditional forms for their glyphs, depending on whether they are at the beginning, middle or end of a word, so they may exhibit four distinct forms (initial, medial, final or isolated).

For Example:

Latin	Name	Final	Medial	Initial	Isolated
ṭ	ṭā'	ط	ـطـ	طـ	ط
ẓ	ẓā'	ظ	ـظـ	ظـ	ظ
ʿ	ʿayn	ع	ـعـ	عـ	ع
ġ	ġayn	غ	ـغـ	غـ	غ
f	fā'	ف	ـفـ	فـ	ف

- Letters that can be joined are always joined in both hand-written and printed Arabic, Persian and Urdu.
- The long vowels /a: /, /i: / and /u:/ are represented by the letters *'alif*, *yā'* and *wāw* respectively.
- Vowel diacritics, which are used to mark short vowels and other special symbols, appear only in the Qur'an. They are also used, though with less consistency, in other religious texts, in classical poetry, in textbooks children and foreign learners, and occasionally in complex texts to avoid ambiguity. Sometimes the diacritics are used for decorative purposes in book titles, letterheads, nameplates, etc.

- Usually in normal texts the diacritics are not used.

3. Middle East Script Unicode Support

With the advent of Unicode support in many web browsers, Arabic, Persian and Urdu script display is no longer an issue. The Unicode rendering engine in XP will take care of the display of these language characters in any form whether it appears in initial, middle, end or isolated positions. XP comes with default fonts for the Arabic, Persian and Urdu as we use Unicode; however there are many freely downloadable fonts for the above languages if we wish to use them.

4. A transliteration Scheme for Arabic, Persian, Urdu

A transliteration scheme referred to as IT3 which is originally developed by IISc Bangalore and Carnegie Mellon University. This transliteration scheme is designed as an improvement over the ITRANS scheme for typing any language characters like Arabic etc… using the standard keyboard. This transliteration mapping is meant to add a few more features to enhance the usability and readability, and has been designed on the following principles:

- Easy readability
- Use of case-insensitive mapping: While preserving readability, this feature allows the use of standard natural language processing tools for parsing and information retrieval to be directly applied to the Indian language Texts.
- Phonetic mapping, as much as possible. This makes it easier for the user to remember the key combinations for different Indian characters.

Example of the IT3 is shown as: a aa k g h’ etc.

4.1 Mapping Table

The important part of the Arabic editor is to map the it3 symbol to the corresponding Arabic Unicode character Using the IT3 notation and the Unicode characters, one can build a simple transliteration editor in a short amount of time. Any new language can be added to it with minimal effort. To add a language, a mapping table has to build to map an IT3 character to the corresponding Unicode character. However attention should be paid while building the Arabic, Persian and Urdu mapping table, as all the letters in these languages are consonants and each letter can appear separately or in combination with one another, unlike Indian Languages we cannot have combination letters as the mapping of IT3 to Unicode for ex: sh is not equal to s and h.

A sample mapping table for Arabic is shown below:

Arabic Mapping Table – Help

a	ا	l	ل	x	ً
aa	آ	m	م	~	ٌ
b	ب	n	ن	!	ٍ
t	ت	w	و	ix	ِ
v	ث	h	ه	ax	ّ
j	ج	y	ى	ux	ُ
d	د	h'	ح	a1	َ
z-	ذ	k'	خ	aa1	ىَ
r	ر	c	ص	i1	ِ
z	ز	d'	ض	ii1	ى
s	س	t'	ط	u1	ُ
s'	ش	z'	ظ	uu1	وُ
f	ف	e	ع	au1	ُوَ
q	ق	g	غ	ai1	ُىَ
k	ك				

Fig2: Arabic Mapping Table

4. 2 User Interface Design

Now we have all the parameters to build the editor like the mapping table and we need to build the interface where we can see both the IT3 and the Arabic/Persian or Urdu characters depending on the language is selected.
To build the Arabic, Persian and Urdu language editor, the following is the pseudo code which is followed: Given an IT3 word, parse it into sequence of characters by phonifying them, for ex: if you give the sequence as nilu then the editor as the first step will phonify the sequence as n i l u as individual phones.
These phones are mapped to their corresponding Unicode numbers and the display of the Arabic/Persian or Urdu script is shown on the editor.
(See Fig 2)

The Middle East languages like Arabic, Persian and Indo-European language Urdu have some common characteristics. All these three languages are written from right to left. Persian and Urdu are derived from Arabic.
The characters of Arabic, Persian and Urdu characters are called alphabets. Each alphabet corresponds to a phone (Library of Congress, 1997).

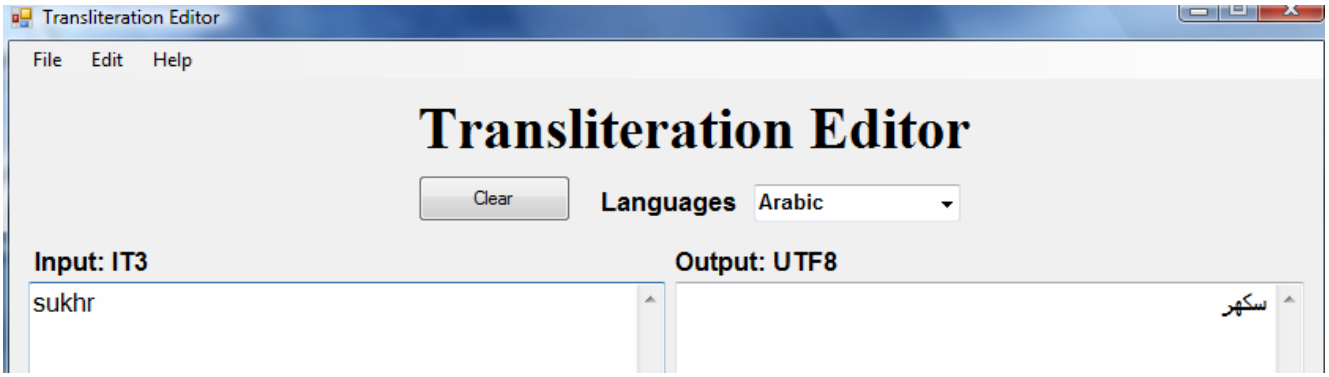
Arabic:
Each alphabet corresponds to a phone. Arabic has 9 vowels and 28 consonants (Library of Congress,
1997; Qur'an Transliteration, 2005).
In Arabic it should be noticed that the vowels do not appear independently as seen in many Indian languages, they occur only with a consonant. All the Arabic characters are written in different ways by the occurrence of the same, i.e. the alphabet is written differently if it occurs in the initial position and it is written differently if it occurs in the middle or end of a word. This rule is same for Persian and Urdu.

Persian:
Persian also has 9 vowels and 32 consonants. Modern Persian uses a modified version of the Arabic alphabet. Persian adds four extra alphabets due to the fact that four sounds that exist in Persian do not exist in Arabic. The additional four alphabets are shown in Table 1 (Omniglot, 2005).

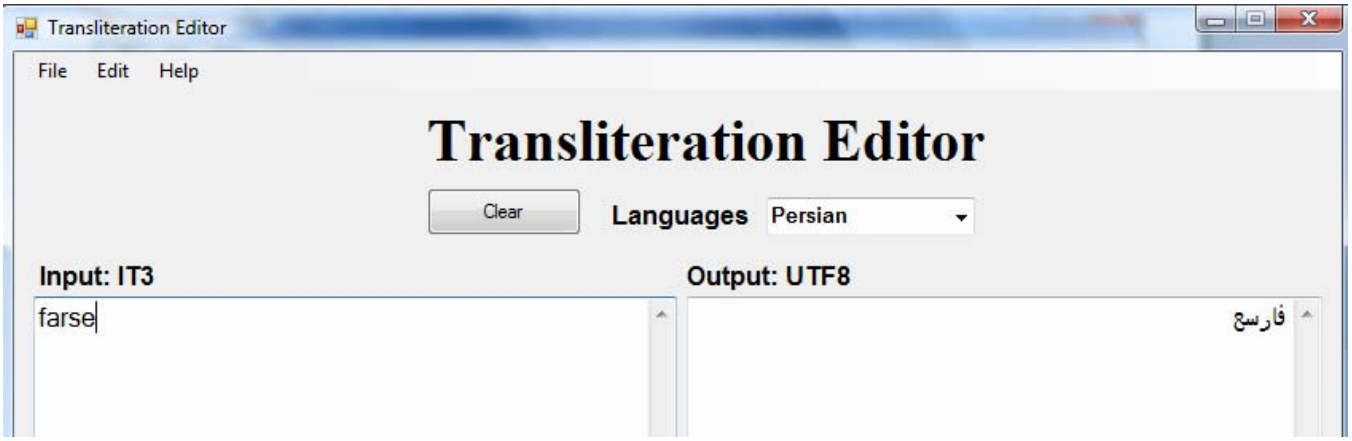
Table 1 Extended Persian set

Sound	Shape	Unicode name
[p]	پ	Peh
[tʃ] (ch)	چ	Tcheh
[ʒ] (zh)	ژ	Jeh
[g]	گ	Gaf

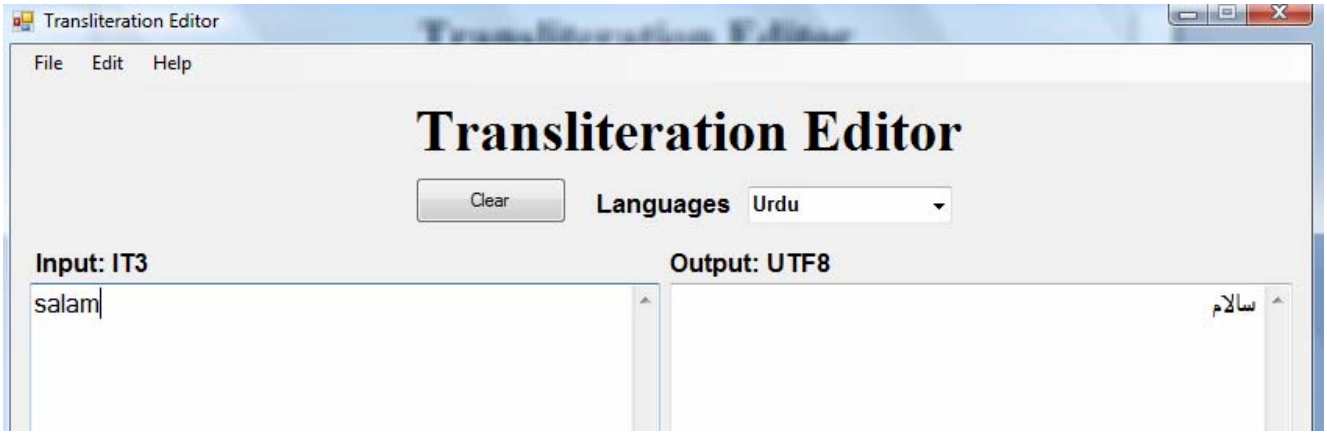
Urdu:
Urdu is derived from Persian which in turn is derived from Arabic. Urdu uses more complex and sinuous Nastaliq script. It is said that Arabic is a subset of Urdu. Urdu has 11 vowels in addition to the 9 vowels of the Arabic alphabet and 35 consonants (alphabets) (Hugo's Website, 2005; U-TRANS, 2002).
Urdu language has two noon (n), one is noon and the other is noon gunna, where as Persian and Arabic has one only noon.



(a)



(b)



(c)

Fig 2: Screen shots of the Arabic, Persian and Urdu Unicode Editor (a) (b) (c) respectively

Fig3: Screen shot of the Meta data of the Arabic Books

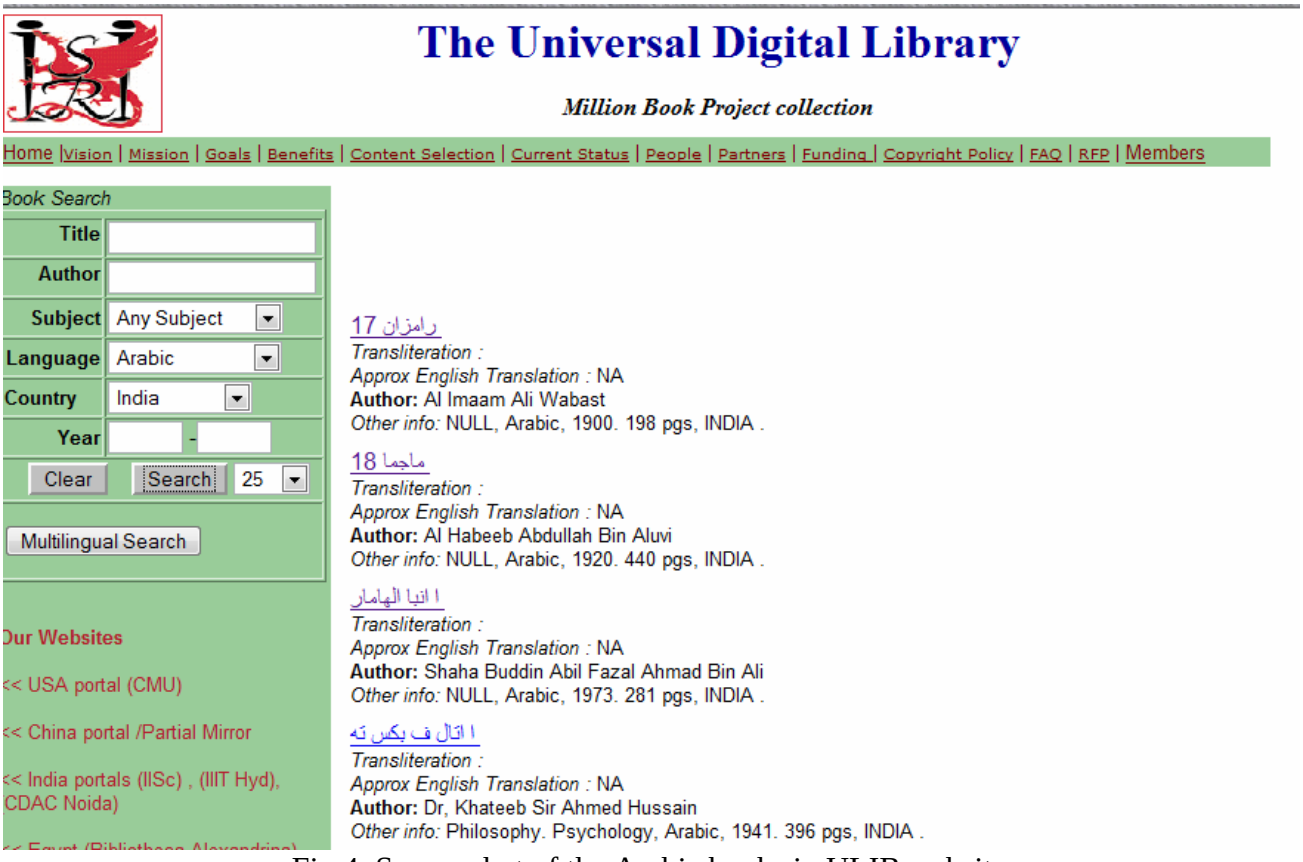


Fig 4: Screen shot of the Arabic books in ULIB website

IT3 codes such as ain (e or o) and zheh (z') and many such more are created and added to accommodate middle-east languages.
The following steps are followed to develop the Transliteration:
(1) Once all the alphabets are assigned IT3 codes, then each IT3 code should be mapped to the corresponding Unicode number.
(2) The above languages have explicit Unicode number assigned to each of the alphabet.
(3) Unlike Indian languages, a consonant alphabet represents a consonant alone and a vowel alphabet represents a vowel. For example in Indian languages "k" is mapped to the Unicode representing /ka/. But in Middle East languages "k" is explicitly mapped to /k/ (kaf). There is no syllabification required in these languages.
(4) Easy to adapt for new languages. Unicode based approaches require minimal knowledge to work in new languages, whereas ASCII font based approach requires a better understanding of the language to handle exceptions related to rendering of consonant clusters.

CONCLUSION

In this work, we have described the process of building Arabic, Persian and Urdu language editors using a simple scheme based on Unicode. This simple approach has the following advantages:
(1) Lesser number of entries in the mapping table. There are only 37 entries for Arabic language.
(2) Automatic rendering of the Unicode characters by the Unicode rendering engine in Windows XP/Linux.
(3) Using Unicode based approach, a single module can render all the languages. The mapping table changes, but the parsing of IT3 sequence and syllabifications are the same across all of the Arabic, Persian and Urdu Languages.
(4) Easy to adapt for new languages. Unicode based approaches require minimal knowledge to work in new languages, whereas ASCII font based approach requires a better understanding of the language to handle exceptions related to rendering of consonant clusters.
(5) Our editor is used in the Universal Digital Libraries to enter the metadata for the Arabic, Persian and Urdu books.
The screen shot of the Meta data is shown in the Fig 3
(6) Using our editor code, it possible to display the Arabic, Persian and Urdu characters even on the webpage, this is also incorporated in the webpage of the Universal Digital Libararies (www.ulib.org) and search for the Arabic books, we can see all the Arabic books shown in Arabic script.

ACKNOWLEDGEMENT

We would like to deeply thank Prof. Raj Reddy for guiding us through the project and has given us the initiative. We would also like to thank Dr. Nayel Shafei for giving us the feedback through out this project. We would like to thank Mr. S P Kishore for helping us in completing this project.

References

(1) Alan, W., 2005. Unicode Resources. http://www.alanwood.net/unicode/index.html.
(2) A Simple Approach for building Transliteration editors

http://www.zju.edu.cn/jzus/2005/A0511/A051125.pdf
Journal of Zhejiang University, 2005
(3) Hugo's Website. 2005. Urdu & Arabic Pages. http://users.skynet.be/hugocoolens/
(4) Library of Congress, 1997. ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts. http://www.loc.gov/catdir/cpso/roman.html.
(5)Markus, K., 2005. UTF-8 and Unicode FAQ for Unix/Linux. http://www.cl.cam.ac.uk/~mgk25/unicode.html
(6) Omniglot, 2005. Persian. http://www.omniglot.com/writing/persian.htm.
(7)Qur'an Transliteration, 2005. Qur'an Transliteration.http://transliteration.org/
(8) The Unicode Consortium, 2003. The Unicode Standard, Version 4.0. Addison-Wesley, Boston, MA.
U-TRANS,2002.Urdu. http://tabish.freeshell.org/u-trans/