Transcription of the Persian language in the Electronic Format

Behrang QasemiZadeh
Presented at the CLUKI 2009, Dublin, Ireland

Abstract

Persian language, also known as Farsi, is the official language of Iran and Tajikistan, and one of the two main languages spoken in Afghanistan. Persian language is one of the popular languages on the web, with an approximation of 100,000 bloggers. This paper addresses problems with manipulation of Persian electronic texts due to its transcription, and encoding in the electronic format; the paper suggests e-orthography as a solution for a number of these problems. Persian is an agglutinative language, and a member of Indo-European languages. However, it is transliterated by Arabic cursive scripts, which cannot serve representation of important features of the language such as morphosyntactic one.

1 Introduction

Persian language is a member of Indo-European family of languages, and within that family it belongs to Indo-Iranian branch. Persian language generally has the properties of agglutinative languages; words are inflected by adding affixes to a root. Each affix typically represents one meaningful unit. Besides, and most importantly, affixes do not become fused with others. The majority of affixes in Persian are suffix with limited prefixes as well. There is no infix detected in Persian. (Iran Kalbasi, 2001)

Three major phases are distinguished in development of the Persian language, namely, Old, Middle and New Persian language. New Persian language now is the official language of Iran, and the Republic of Tajikistan, and one of the two main languages spoken in Afghanistan (where it is referred to as Dari in Afghanistan, and Tajiki in Tajikistan). Local environments such as Arabic language and Russian language have influenced the Persian language in different geographical regions. It is estimated around a hundred million people speak Persian, 70 millions in Iran, 20 millions in Afghanistan, and 10 millions in Tajikistan[*].

Old Persian was based on the cuneiform writing system (pictogram style) as early as the 6th century B.C. Later, the Persians invented a new alphabet called Pahlavi to replace the uniform alphabet. However, after the Arabic conquest in 651, the Persians adopted a unified Arabic script for writing. As a result, Persian language is written in Arabic script in Iran and Afghanistan. However, Tajik uses Cyrillic alphabet for the transliteration of the language. Despite Arabic and Persian's shared transliteration, they are belonging to separate genetic language families, namely, Indo-European and Afro-Asiatic and have different phonology and grammar. Arabic is a Semitic language with template based morphology, however Persian language uses agglutination to form new words.

With the expansion of Islam in the course of history, Arabic script was forced as a system of writing also for other languages like Persian. As in many of these languages, among them Persian, Urdu, and Sindhi, have phonemes different to Arabic ones, the repertoire of Arabic characters was extended. The original Arabic language alphabet consists of 28 characters. Persian writing system uses the Arabic alphabet, but with the addition of four letters which do not occur in Arabic. These are: "گ", /gâf/, "چ", /če/, "پ", /pe/, and "ژ", /že/. Additionally, it changes the shape of another two i.e. "ی", /ye/, and "ک", /kâf/. As a matter of fact, not all of the sounds in the Arabic alphabet exist in

[*] The statistics are quoted from BBC World Service, who launched a new Persian TV channel on January 14th, 2009.

the Persian language; as a result, more than one letter may represent one sound. For example, there are two letters in Persian for the sound /t/, i.e. "ط", and "ت".

Salient characteristics of Arabic script are: existence of various connecting letters, varying graphic forms for many letters depending on their position in a word (figure1), varying letter width, absence of full size characters for vowels (vowels are represented as particular signs above and below characters), existence of a number of digraphs and composite letters, writing direction from right to left, and absence of upper case and lower case letters.

Figure 1. Arabic letters may have up to four visual representation (glyph) based on their position in a word. Above shows glyphs for the letter /p/.

The use of Arabic script for transliteration of Persian language brings difficulties in computational analysis of the Persian electronic texts. The difficulties usually take place in the form of ambiguity in encoding of Persian electronic texts. This paper addresses problems when manipulating Persian e-texts and introduces e-orthography concept as a guideline for solving these problems.

The paper is organized as follow. The next section, Section 2, describes common problems of transliteration of Persian language in Arabic script in the paper based system, then it describes how the "Iranian Academy of Persian Language and Literature" tries to solve the problem by introducing *Persian Script Orthography*. The section 3 discusses the problems highlighted in Section 2 in the context of computational analysis of Persian e-texts; further more it introduces the e-orthography as a solution. Finally, we conclude in section 4.

2 Transliteration of Persian language in Arabic Script

As mentioned earlier, Persian language uses Arabic script as its transliteration system in Iran and Afghanistan. Since Arabic is a cursive script, the number of possible shapes that letters actually can adopt exceeds the number of these letters. Letters attach to each other to represent a word. Since Arabic has a template based morphology, it is obvious that how letters must be attached to each other to form a word. In Persian, however, due to the fact that it is an agglutinative language, there could be ambiguity in what letters should be written attached together or detached. For instance, the plural form of the word "کتاب"- /ketâb/, which means book - may be written as "کتاب ها" -/ketâb hâ/ which means books, or "کتابها"-/ketâbhâ/ with the same meaning and pronunciation. One of the problems of such variation in written forms is the ambiguity in word boundaries. In addition, the fact that more than one letter represent one sound cause confusion when transliteration of words with this sound, e.g. both the word "طهران", and "تهران" - pronounced /tehrân/ - and one may use any of these transliteration to write "Tehran".

Persian benefits from a case system (Bakhtiari, 2003), and words may be pronounced in different way in different morphosyntactic situations. Cases usually are presented as short vowels at the end of words, for example genitive case of a Persian word is composed of the word in addition to the short vowel /e/ at the end of the word. The facts that short vowels are not full letters in Arabic transcription, and they are not usually written cause the loss of information about the case of words in the Persian written texts. Moreover, as many words in the Persian language are different only in pronunciation, omitting the short vowels in the written form of words exceeds the number of homographs in the Arabic transliteration of the Persian language.

"Iranian Academy of Persian Language and Literature" -which is a governmental body presiding over the use of the Persian language - has created an official orthography of the Farsi language, entitled "Dastur-e Xatt-e Fârsi" (Farsi Script Orthography), for the proper representation of texts in the paper based system of writing. This orthography is the common orthography widely used by the Persian speakers and indicates how characters must be attached to each other to present a Persian Word. For example, it specifies how affixes should be attached to words. In addition it provides a set of guideline for the use of Arabic letters, and the dictation of words in Arabic transliteration of the Persian words e.g. "تهران" instead of "طهران". The proposed orthography also suggests the use of short vowels as an option when omitting the short vowel results in ambiguity.

3 Persian in the electronic format

In this section we describe the encoding of the Persian language in the Unicode framework. We discuss how the current standard is insufficient for encoding a language like the Persian language. Unicode was devised so that one unique code is used to represent each character, even if that character is used in multiple languages. Unicode standard version 4.0 reserves the range 0600 to 06FF for Arabic characters. The important design principles observed in the Unicode standard and relevant to the representation of Arabic script are characters not glyphs. As mentioned earlier, Arabic letters can have up to four different positional forms (figure 1) depending on their position relative to other letters or spaces. According to the design principle "characters, not glyphs", there is no individual code for each visual form (glyph) that an Arabic character can take in varying contexts; but there exists only one code for each actual letter. The correct glyphs to be displayed for a particular sequence of Arabic characters can be determined by an algorithm. In order to display the characters properly, special characters such as "Zero Width Joiner (ZWJ)", "Zero Width Non Joiner (ZWNJ)", and "Right-to-Left Override (RLO)" are added to the character codes. For example, the use of ZWNJ characters after a code means that the character before ZWNJ character should be appeared in one of its final form glyphs, and character after ZWNJ should be represented in one of its initial glyphs, similarly, characters after RLO should to be treated as strong right-to-left characters.

The ISIRI 6219:2002 (Information Technology – Farsi Information Interchange and Display Mechanism, using Unicode) has been proposed as the standard for using Unicode in encoding Persian language by The Institute of Standards & Industrial Research of Iran. This standard indicates a subset of Arabic character set in Unicode to be used by Persian users. Despite this standard, Persian keyboard layouts are likely to use different codes and therefore, many of Persian users do not follow this standard. Moreover, the ISIRI 6219:2002 standard does not enlighten how Persian language orthography can be obeyed in this standard. To explain the latter fact we continue with an example; assuming that Persian Orthography asks users to write inflection suffixes in detached from, a user can represent this visual form by use of either ZWNJ character, or the white space character as a delimiter between a root and suffixes (figure 2).

کتاب + SPACE + ها = کتاب ها

کتاب + ZWNJ + ها = کتاب‌ها

کتاب + ها = کتابها

Figure 2. An example of different Arabic transliteration for a Persian inflected word. In the given example, the suffix /hâ/ is attaching to the root /ketâb/ to form a new word /ketâbhâ/. /ketâbhâ/ is the plural form of the noun /ketâb/ and means books. As figure shows, the word /ketâbhâ/ can be transliterated in three different ways, and with different logical encoding strings in the Unicode standard.

The consequences of the lack of guidelines for Persian language representation in electronic format are different encoding strings for the conceptually the same word, and problems in proper visual representation of Persian e-texts, especially when an e-text contains number or left to right characters. While the latter problem is important in word concordance view (Rychly, 2007), the former fact is important when providing frequency profiles of words, searching keywords, and dealing with *precision* and *recall* measures in corpus based linguistics. Get back to the example of books; pronounce /ketâbhâ/ in the Persian language (figure1), is composed of two morphemes, a free morpheme, the root /ketâb/ and a bound morpheme /hâ/, one of plural morphemes, this word can be transliterated in Arabic in three different ways: with a white space between root and bound morpheme: "کتاب ها", with a ZWNJ between the root and the bound morpheme: "کتابها", and the bound morpheme attached to the root: "کتابها"; needless to say each of these three forms may be written with two different Unicode character code for the letter "ک", /kâf/ and an arbitrary number of the letter '-', i.e. "Tatweel" - a letter to let verity of width in the visual form of a word- between any two attached letter, and absence or presence of short vowels. In this way, one written form of a word in a paper based system, can have several number of logical representation in en electronic format. The following table shows the number of results in a search for different transliteration of the word /ketâbhâ/ in Arabic, and English where the latter known as Pinglish between the Persian users.

Transliteration	Number of occurrence acording to a keyword based search engine
کتاب ها	12,600,000
کتابها	2,580,000
کتاب‌ها	2,490,000
Ketabha	12,900
Ketab ha	1,650

Table 1. The right column shows different transliteration of the Persian word /ketâbhâ/ in the electronic form; the right column shows the number of result in web pages for that transliteration as a measure of popularity between users

However, the importance of encoding goes beyond that. The policy of text encoding, tokenization, orthography, and corpus tagging are in interaction with each other. For example, as mentioned earlier in Persian it is possible that a bound morpheme appears detached from its root with an intervening space; if we assume space as a delimiter in the tokenization process according to the used orthography, either we have to consider a tag for these bound morphemes during corpus tagging or, we have to consider a more complicated tokenization process as it is cited in (Megerdoomian, 2000).

The concept of e-Orthography for Persian language has been introduced by (Qasemizadeh, 2007). E-orthography tries to fill the gaps in electronic encoding systems; the gaps are consequences of the lack of enough guideline for encoding and representing electronic texts in the current standard frameworks such as Unicode. The e-orthography indicates how the orthography of a language can be followed within an encoding system. A simple e-orthography guideline for the Persian language can be as follows: The character set based on the proposed standard in (ISIRI 6219:2002), ZWNJ character as the short space between bound morphemes and free morphemes, and space characters as unambiguous word boundaries.

The e-orthography can be used in two different ways. First, it is to design keyboard layouts in a way that they support all necessarily needed characters, e.g. adding “Alt Gr” key to Persian keyboards, and supporting RLO, ZWNJ, etc. in a keyboard layout, along with an education for the Persian users to have more consistent representation of Persian e-texts. Second, the e-orthography can be used as a guideline when indexing and manipulating Persian e-texts by corpus query systems. A corpus query system may provide a set of normalization processes on the raw data to represent a standard form of e-texts, and at the same time provides users of the system about the e-orthography that has been used for indexing e-texts. Lastly, the e-orthography can be used as a functional requirement when developing graphical representation of character codes.

4 Conclusion

This paper described problems in encoding of the Persian language in the Unicode framework. It was explained why the current encoding standards, like the Unicode, is not sufficient to represent a consistent encoding of the Persian language. We suggested e-orthography as a solution to these problems. E-orthography indicates how the orthography of a language can be followed within an encoding system. Therefore, e-orthography should notice what character codes must be used, how they attach to each other to form a word, and finally which tokenization policy must be taken.

Important measures in text retrieval, such as *Precision* and *Recall* measures, indexing terms, frequency lists, even statistics for collocation are subject to our definition for logical encoding of texts. To be in line with the current language computation technology, a new standard for encoding of electronic texts in a more semantic way looks necessary. The expansion of web, the huge amount of electronic texts generates every hour, and the need for more effective way of text manipulation and retrieval, in addition to the mission of preserving languages on the web, might be enough reasons to review current standards such as Unicode in the outlook of World Wide Web.

References

Behrrooz M. Bakhtiari, 2003. *Case Systems in West Iranian Languages: A typological Study*, PhD Dissertation, Allame Tabataba'i University.

Iran Kalbasi, 2001. *The Derivational Structure of Word in Modern Persian*, ISBN 964-426-128-3, Institute for Humanities and Cultural Studies.

The Unicode Standard, Http://www.Unicode.org/.

Iranians Academy of Persian Language and Literature, *Official Persian Orthography*, http://www.persianacademy.ir/UserFiles/Image/Dastoor-e%20khat/d02.pdf.

Institute of Standards & Industrial Research of Iran, 2002. *Isiri 6219:2002, Information Technology – Persian Information Interchange And Display Mechanism, Using Unicode*, Http://www.shci.Ir/Download/Unicode%20finalversion.pdf.

Karine Megerdoomian, and Remi Zajac, 2000. *Processing Persian Text: Tokenization In The Shiraz Project* .Nmsu, Crl, Memoranda In Computer and Cognitive Scienc.

Behrang Qasemizadeh, 2007. *Farsi e-Orthography: an Example of e-Orthography Concept*, In: F. Lazarinis, J. Vilares, J. Tait (eds) Improving Non-English Web Searching (iNEWS07) SIGIR07 Workshop, pp. 62--64.

Pavel Rychly, and Vojtěch Kovar, 2007. *Displaying Bidirectional Text Concordances in KWIC format.* In Proceedings of 5th Biennial Conference of the Asian Association for Lexicography. Chennai, India : University of Madras, pp. 96-100.