Farsi e-Orthography: An Example of e-Orthography Concept

Behrang Qasemizadeh
Text and Speech Ltd
Tehran, Iran
qasemizadeh@comp.iust.ac.ir

ABSTRACT
Farsi, also known as Persian, is the official language of Iran and Tajikistan and one of the two main languages spoken in Afghanistan. Farsi enjoys a unified Arabic script as its writing system. The fact of using Arabic scripts, a Semitic Language, for representation of Farsi, an Indo-European Language, leads to problems when analyzing, and retrieving Farsi e-text. In this paper we briefly introduce Farsi writing system, and highlight problems when analyzing Farsi electronic texts especially during retrieving Farsi e-texts. Then we introduce the concept of e-orthography. We discuss how e-orthography could be used to improve search results while using keyword based search engines.

Keywords
E-Orthography, Farsi.

1. INTRODUCTION
People in different countries use different characters to represent the words of their native languages. With library automation and the development of networked information structures, the problem of finding a unique way to show information has become much more complex [1][2]. Unicode [4] was devised so that one unique code is used to represent each character, even if that character is used in multiple languages [3]. In this paper, we describe Farsi language transcription in Unicode framework and we discuss challenges that someone would face when processing and retrieving Farsi e-texts.

Farsi is a member of the Indo-Iranian family of the Indo-European languages. Farsi has the properties of agglutinative languages. [5][6] The majority of affixes in Farsi are suffix with limited prefixes as well. After the Arab's conquest in 651 A.D., the Persians adopted an extension of unified Arabic script for writing. Salient characteristics of Arabic script are: existence of various connecting letters, varying graphic forms for many letters depending on their position in a word, varying letter width, absence of full size characters for vowels (vowels are represented with 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.

SIGIR'07 iNEWS07 workshop, July 27, 2007, Amsterdam, The Netherlands.

Editors: Fotis Lazarinis and Jesus Vilares and John I. Tait

General rules of Arabic writing system are followed by the writing system of Farsi.

Since Arabic is a cursive script, the number of possible shapes that letters actually can adopt exceeds the number of these letters [8]. Letters attach to each other to represent a word. Since Arabic is a Semitic language, it is obvious that how letters must be attached to each other to represent a word. In Farsi, 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/ (book) may be written as 'کتابها' /ketâbhâ/ or 'کتاب ها' /ketâb hâ/ (books). This results in some difficulties in Farsi text analysis as cited in [7][8][9], i.e. tokenization of Farsi e-text since word boundaries are not clear. Also, the fact that short vowels usually are not written and capitalization is not used will result in ambiguities that impede computational analysis of the texts. Since these various representations of Farsi are encoded in different manner, then in many cases a search engine can not retrieve Farsi texts.

In the following, after a brief introduction to Farsi encoding, we will introduce the concept of e-orthography and we discuss how it may be used to tackle the problems when analyzing and retrieving Farsi e-texts. The rest of paper is organized as follows: section 2 introduces Farsi transcription and encoding. Section 3, describes the e-orthography concept and its application to Farsi. Finally, we conclude in section 4.

2. Farsi Transcription and Encoding in Digital Environments
"Iranian Academy of Persian Language and Literature", which is a governmental body presiding over the use of the Farsi language, has created an official orthography of the Farsi language, entitled "Dastur-e Xatt-e Fârsi" (Farsi Script Orthography) [10], 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 attached to each other to present a Farsi Word. For example, it specifies how affixes should be attached to words.

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 in the previous section, Arabic letters can have up to four different positional forms 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, two special characters namely Zero Width Joiner (0x200D) and Zero Width Non Joiner (0x200C) are added to the character codes, either before or after them. The use of these special characters after a code means that a ZWJ or a ZWNJ should be added after the character if the character is not followed by a "right-join causing" character, or a "non-joining character" respectively.

The ISIRI 6219:2002 (Information Technology – Farsi Information Interchange and Display Mechanism, using Unicode) [11] has been proposed as the Farsi standard for using Unicode in digital environment. This standard indicates a subset of Arabic character set in Unicode to be used by Farsi users. Despite this standard, Farsi keyboard layouts are using different codes and therefore, many of Farsi users do not follow this standard. Moreover, the ISIRI 6219:2002 standard does not enlighten how Farsi Orthography can be obeyed in this standard.

The mentioned fact imposes difficulties when retrieving Persian texts, since characters, and therefore words are represented with different codes and search engines do not cover this problem. For example, a word like 'اتمی' /atomi/ which means "Atomic" can be represented in two different coding string since the last character has two encoding option. So, if you search for documents which contain this word, you may miss number of actual results since you have searched just for one of the forms of the word depending on the keyboard layout of your system. The problem is getting more complex when an affix is used to change morphosyntactic features of words. Usually affixes can be written in three different forms regarding the word, attached to the word, detached and with a space between word and affix, detached but with a ZWNJ character between them.

We should consider that the policy of text encoding, tokenization, orthography, and text processing are in interaction with each other. As a real example, consider we would like to define a tag set for Farsi Corpus tagging. As mentioned, in Farsi it is possible that a bound morpheme appears detached from its stem 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 [7] [9].

3. Farsi e-Orthography

Unfortunately there exists no standard format for Farsi orthography in the digital environment. As mentioned above, the encoding standard is not sufficient to represent a consistent representation for Farsi. For this reason, we have suggested an approach to represent Farsi electronic texts, or e-orthography. In other words, the 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.

As to Persian, according to the proposed paper-based orthography by the Academy, Farsi affixes must be written attached to their stem. In some cases when the stem ends in a letter which is a "right-join causing character", the affix must attach to the stem with a short space character before it. In order to reach this objective in electronic texts, ZWNJ character has been used as the short space. Also a character set based on the proposed standard in [11] has been used. This way, space characters represent unambiguous word boundaries and the orthography of Farsi e-texts remains consistent with the one proposed in [10]. Also, this transcription results in Farsi e-texts which are more consistent with the e-texts of other languages.

In a keyword based search engine, the e-orthography with the proposed definition influences the effect of search engines in two ways. First of all, the index terms may be changed since the tokenization policy may be varied. Moreover, the user query can be described in other forms which are consistent with proposed e-orthographies. To have an idea, as to Farsi, if we search for a word like 'کتاب‌ها' /ketâbhâ/, a search engine may just retrieve 10% of documents containing this term, considering that first of all character may be represented by different codes, the suffix is written in other forms, characters represented with different lengths, and short vowels may be written or not. An application of proposed e-orthography may be viewed in the development of '1984 corpus' for Farsi [12].

4. Conclusion

This paper introduces the concept of e-orthography and its important role in the efficiency of keyword based search engines. e-orthography tells us how the orthography of a language can be followed in an encoding system, what character codes should be used, how they attach to each other to form a word, and which tokenization policy must be taken in document processing.

E-Orthography can be a guideline for both systems that generate e-text, as well systems which are used to retrieve and manage e-texts. As to the keyword based search engines, the e-orthography can describe how the input query of the users should be refined to retrieve documents. Also e-orthography can change the indices and keywords which are used to retrieve documents.

Although the paper concerns Farsi, the concept of e-orthography can be expanded to other languages as well. Including the e-orthography concept as part of search engines' design can enhance recall and precision parameters. Moreover, the e-orthography concept can be used in other domains like natural language processing and corpus tagging. The mentioned fact indicates that the present standards for text encoding are not sufficient for proper representation, as well as retrieving e-texts.

The concept of e-orthography is getting more important while analyzing languages such as Farsi and Kurdish; languages that have problems in their representation because of the language nature and their writing system.

5. REFERENCES

[1] Erickson J.C. Options For Presentation of Multi-Lingual Text: Use Of the Unicode Standard, *Library Hi Tech, Vol. 15, No. 3-4, 1997.*

[2] Lutz W. Unicode and Arabic Script, *Workshop "Unicode Und Mehrschriftlichkeit In Katalogen", Sbb Pk, Berlin, 2003.*

[3] Wells J.C. Orthographic Diacritics and Multilingual Computing, Language Problems and Language Planning. *Vol. 24, No. 3, 2000.*

[4] The Unicode Standard At Http://www.Unicode.org/.

[5] Samare I. Typological Features Of Farsi. *Journal Of Linguistics, Iran University Press, No. 7, pp 61-80, 1990.*

[6] 7. Keshani, K. Suffix Derivation in Contemporary Farsi. *First Edition, Iran University Press, 1992.*

[7] Karine M., and Zajac R. Processing Farsi Text: Tokenization In The Shiraz Project..*Nmsu, Crl, Memoranda In Computer And Cognitive Scienc, 2000.*

[8] Qasemizadeh, B. and Rahimi, S. Farsi Morphology. *11th Computer Society of Iran Computer Conference, IPM, Tehran, Iran, 2006.*

[9] Rezaie S. Tokenizing an Arabic Script Language. *Arabic Language Processing: Status and Prospects, Acl/Eacl, 2001.*

[10] Iran's Academy Of Farsi Language and Literature. Official Farsi Orthography. *ISBN: 964-7531-13-3, 3rd Edition, 2005.*

[11] Isiri 6219:2002. Information Technology - Farsi Information Interchange and Display Mechanism Using Unicode. *2002.*

[12] QasemiZadeh B., and Rahimi S. Persian in MULTEXT-East Framework. *FinTAL 2006: 541-551, Springer Publisher, Lecture Notes in Computer Science, Vol. 4139, 2006.*