Iranian Standard Code for Information Interchange (ISCII) *

Mohammad SANATI

Computer Science Department, Worcester Polytechnic Institute, Worcester, Massachusetts 01609, U.S.A.

Mohammad DADASHZADEH

Department of Management Science and Information Systems, University of Detroit, Detroit, Michigan 48221, U.S.A.

and

Mohammad B. DADFAR

Department of Computer Science, Bowling Green State University, Bowling Green, Ohio 43403, U.S.A.

The need for a standard code for computer data representation is well understood. Such standards for Latin speaking countries have been in existence for decades and have facilitated information interchange and the advancement of computer peripheral technology. In this paper we present the proposed Iranian Standard Code for Information Interchange (ISCII) and the proposed Iranian Standard Keyboard Layout. The Iranian standards are specified to accomodate the difficult goals of maintaining the 7-bit ASCII subset for bilingual applications and a well defined collating sequence while allowing for the variety of shapes that letters of the Farsi alphabet assume depending on their position in a written word, a problem that is shared by another script language – Arabic. The proposed ISCII, which is compatible with ISO-IRV, is shown to satisfy the two criteria of sufficiency and efficiency for computer data representation and display purposes. Furthermore, it is independent of the strengths or limitations of any particular computer output display technology.

Keywords and Phrases: Arabic Script Design; ASMO-449; Bilingual Keyboard Layout; Dual Code; Farsi Information Processing; Intelligent Output Devices; ISO-IRV; Lexical Analysis; Pseudo-space; Script Languages; Shifted Code; Standard Code.

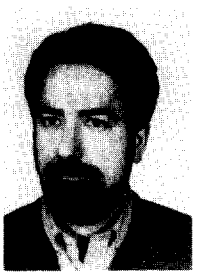
* This work was initiated by the High Council of Informatics of Iran in an effort to establish a standard code for Farsi language. It was supported in part by the High Council of Informatics of Iran and Computer Science department at Worcester Polytechnic Institute. Requests for reprints should be addressed to M.B. Dadfar.

North-Holland
Computer Standards & Interfaces 6 (1987) 427–432

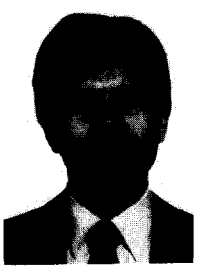
1. Introduction

A standard code for information interchange is required in order to provide a uniform means of communication among data processing equipments. Such standard codes have been long in existence for Latin languages. Unfortunately, establishing similar standard codes for non-Latin languages has been slow and difficult. Two such languages are Arabic and Farsi, the latter being the language of countries such as Iran, Afghanistan, and as Urdu language in Pakistan and India.

Mohammad Sanati was born in 1953 in Ardabil, Iran. He received his B.Sc. in Physics from University of Tehran. In 1977 and 1980 he received his M.Sc. and Ph.D. in Computer Science from State University of New York at Binghamton. From 1980-1983 he was a faculty of Computer Science Department at Bowling Green State University. In 1983 he joined the Computer Science Department at Worcester Polytechnic Institute. He is currently president of SinaSoft, Inc.



Mohammad Dadashzadeh was born in 1954 in Tehran, Iran. He received his B.Sc. in Electrical Engineering, and his M.Sc. in Computer Science both from Massachusetts Institute of Technology. Having completed in his M.B.A. training at the American International College, he joined the Computer and Information Science department of University of Massachusetts at Amherst from which he received his Ph.D. in 1985. Since August 1985, he has been with University of Detroit where he serves as an assistant professor of management information systems.



Mohammad B. Dadfar is a faculty member in the Computer Science Department at Bowling Green State University which he joined as an assistant professor in January, 1982. He received his B.Sc. in Physics from University of Tehran in 1975, and his M. Sc. and Ph.D. in Computer Science from the State University of New York at Binghamton in 1978 and 1982, respectively. His research interests are Computer Extension and Analysis of Perturbation Series, Computer Algebra in Applied Mathematics, Bilingual Applications of Microcomputers, and Operating Systems and Communications Networks.

The principal problem in defining a standard code for Arabic and Farsi stems from the script nature of these languages. Unlike Latin languages, the shape of a particular letter in Arabic or Farsi is dependent upon its poisition in the word. As such, in defining a standard code for these languages one is faced with the decision to either assign a single code for each letter of the alphabet regardless of its shape, or to assign different codes for the different shapes of each letter. The first approach assumes the existence of an algorithmic process by means of which the correct shape of a letter can be deduced from its adjacent characters (lexical analysis), and the incorporation of such intelligence in the output devices. The second approach increases the number of required codes by approximately a factor of four while introducing major difficulties in such common data processing tasks as sorting and searching (see section 3).

These problems notwithstanding, the increased urgency to introduce and take advantage of information systems' technology in the middle east countries resulted in the adoption of ASMO-449 [1] in October 1982 as the standard code for Arabic by the Arab Standard and Metrology Organization. In the following section we show that an approach similar to ASMO-449 for Farsi proves to be inadequate. The proposed Iranian Standard Code for Information Interchange is presented in section three. In section four we present the proposed Iranian Standard Keyboard Layout. We conclude in section five with a summary and some implementation considerations.

2. Characteristics of Farsi

Farsi is the official language of Iran and it is also spoken in other neighboring countries. Farsi is a script language with an alphabet consisting of 35 letters. As in Arabic each letter has up to four different shapes depending on its position in the word and the root of the word itself. Figure 1a shows an example of how the position of a letter in a word affects its shape. In most cases the letters are connected together in a word, although there exist situations where a letter should be separated from the previous and/or the next letter. (Ten letters in the Farsi alphabet appear only as separated or connected last.)

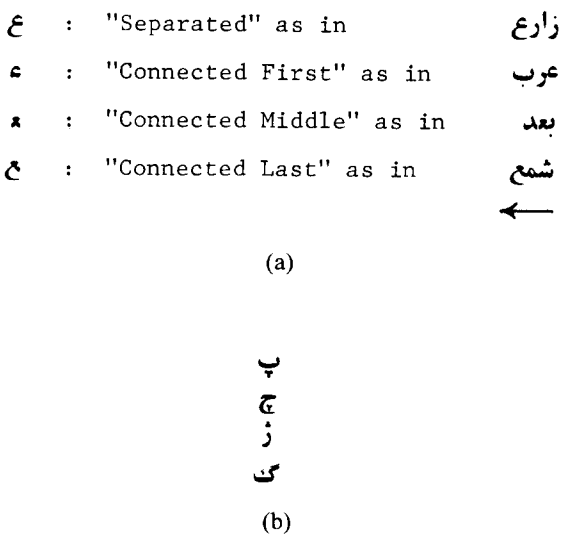
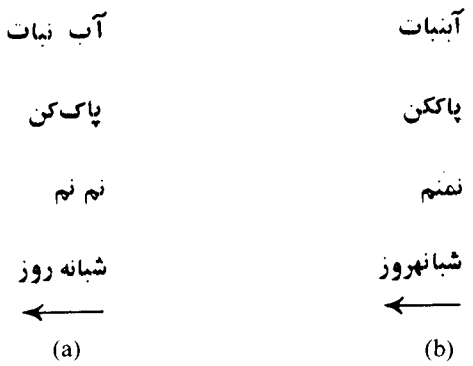


Fig. 1. (a) Different shapes of the letter "ain". (b) Four letters in the Farsi alphabet not included in Arabic.

In Farsi text is written from right to left while numeric information, as in Latin languages, is written from left to right. In many ways Farsi is similar to Arabic. The two languages share a subset of letters with Farsi including an additional four letters (Fig. 1b). Both languages incorporate a Tashdid (Shadda) form for each letter which is disregarded in alphabetization. Both languages include a hamzah form for the letters aleph, waw, and ya, although they differ in alphabetization significance of the hamzah forms. A major difference between Farsi and Arabic is in the absence of vowel signs that are so prominent in Arabic text.

At first glance, the similarities between Farsi and Arabic would suggest that a standard code for Farsi could be obtained by extending ASMO-449 to include the four additional Farsi letters. This would create a compatibility at the data processing level between Farsi and Arabic similar to that existing amongst Latin languages. Unfortunately, such a approach cannot provide a *sufficient* coded character set for Farsi. The reason lies with the fact that while in Arabic the particular shape of a letter in a word can *always* be unambiguously determined by lexical analysis, the same is not true for Farsi. That is, in Farsi, sometimes a letter in the middle of a word is displayed in its connected last form rather than its connected middle shape. For example, assuming a single code for



(a) (b)

Fig. 2. (a) Examples of words where a letter located in the middle of the word is displayed as connected last. (b) By using lexical analysis, these words are not displayed properly.

each letter as in ASMO-449, the words in Fig. 2a, reconstructed by lexical analysis, would be displayed as in Fig. 2b which is undesirable at best. To remedy this situation, the use of a space or "pseudo-space" following the letter to be displayed in its connected last form in the middle of the word has been suggested [5]. This solution, however, suffers from inefficiencies both in lengthening the text and in introducing difficulties in sorting applications [6].

Thus, in closer examination, an approach similar to ASMO-449, which utilizes a single code for each letter in the alphabet, proves to be inadequate for Farsi. It seems unlikely that a single coded character set could efficiently meet the needs of both languages.

3. The proposed Iranian Standard Code for Information Interchange

Although the objective of a one-to-one correspondence between codes and displayable char-

Row	b4 b3 b2 b1	0 (000)	1 (001)	2 (010)	3 (011)	4 (100)	5 (101)	6 (110)	7 (111)
0	0 0 0 0	NUL	DLE	SP	۰	ب		ض	ك
1	0 0 0 1	SOH	DC1	!	۱	بـ	د	ضـ	كـ
2	0 0 1 0	STX	DC2	"	۲	پ		ط	ل
3	0 0 1 1	ETX	DC3	×	۳	پـ	ذ	طـ	لـ
4	0 1 0 0	EOT	DC4	ﷺ	۴	ت		ظ	م
5	0 1 0 1	ENQ	NAK	/	۵	تـ	ر	ظـ	مـ
6	0 1 1 0	ACK	SYN	:	۶	ث		ع	ن
7	0 1 1 1	BEL	ETB	؟	۷	ثـ	ز	عـ	نـ
8	1 0 0 0	BS	CAN	(	۸	ج		غ	و
9	1 0 0 1	HT	EM	)	۹	جـ	ژ	غـ	ؤ
10	1 0 1 0	LF	SUB	=	ء	چ	س	ف	ه
11	1 0 1 1	VT	ESC	+	ئ	چـ	سـ	فـ	هـ
12	1 1 0 0	FF	FS	،	آ	ح	ش	ق	ی
13	1 1 0 1	CR	GS	–	أ	حـ	شـ	قـ	یـ
14	1 1 1 0	SO	RS	.	إ	خ	ص	ک	ۀ
15	1 1 1 1	SI	US	/	ا	خـ	صـ	کـ	DEL

Fig. 3. The proposed Iranian standard code in 7-bit environments.

acters is an attractive one, it would be both difficult to attain and inefficient for Farsi. On one hand, to allow for the different shapes of each letter of the Farsi alphabet we would require 120 codes. On the other hand, because the different shapes of each letter are considered equivalent for alphabetization purposes and since not each letter has the four possible shapes, we would require additional processing to map the different shapes of each letter to its unique colatting sequence value.

The proposed Iranian Standard Code for Information Interchange solves this dilemma in the following manner. We remove the requirement of one-to-one correspondence between codes and displayable characters by placing the burden of deciding the appropriate shape of each letter on intelligent output devices. This means that, for example, on receiving a letter in its input stream, the logical output device (a video driver or a printer driver) would decide the correct shape of the letter by lexical analysis and would cause the associated physical output device to display the desired shape (perhaps by means of an escape sequence). Of course, as we have already explained, lexical analysis alone cannot always decide the correct shape of a letter and a *dual code*, basic and shifted, would be required. The shifted code is used to override lexical analysis whenever a letter in the middle of a word should appear in its connected last form.

Ten letters in the Farsi alphabet do not have a connected middle form and as such do not require a shifted code. However, as indicated above, by

Row	b4 b3 b2 b1	0 (1 0 0 0)	1 (1 0 0 1)	2 (1 0 1 0)	3 (1 0 1 1)	4 (1 1 0 0)	5 (1 1 0 1)	6 (1 1 1 0)	7 (1 1 1 1)
0	0 0 0 0	≡	>	SP	۰	ب		ض	گ
1	0 0 0 1	±	<	!	۱	بـ	د	ضـ	گـ
2	0 0 1 0	≥	⌐	"	۲	پ		ط	ل
3	0 0 1 1	≤	⌐	×	۳	پـ	ذ	طـ	لـ
4	0 1 0 0	∫	⌐		۴	ت		ظ	م
5	0 1 0 1	∫	⌐	/	۵	تـ	ر	ظـ	مـ
6	0 1 1 0	÷	⌐	:	۶	ث		ع	ن
7	0 1 1 1	≈	⌐	؟	۷	ثـ	ز	عـ	نـ
8	1 0 0 0	°	⌐	(	۸	ج		غ	و
9	1 0 0 1	√	⌐	)	۹	جـ	ژ	غـ	ؤ
10	1 0 1 0	∪	⌐	=	ء	چ	س	ف	ه
11	1 0 1 1	∩	⌐	+	ء	چـ	سـ	فـ	ه
12	1 1 0 0	⊂	⌐	،	آ	ح	ش	ق	ی
13	1 1 0 1	Σ	1/2	–	أ	حـ	شـ	قـ	یـ
14	1 1 1 0	Π	1/4	.	ا	خ	ص	ک	ۀ
15	1 1 1 1	∞	س	/	أ	خـ	صـ	کـ	

Fig. 4. The proposed Iranian standard code in 8-bit environments.

not assigning the same number of codes to each letter we would introduce the need for pre-processing whenever alphabetization is required. Because of the importance of sorting and searching in data processing applications a highly efficient means of establishing the collating sequence value of each code would be desired [4]. The proposed Iranian Standard Code for Information Interchange achieves this goal by assigning a dual code to each letter with the two codes differing only in the first bit. Therefore, conversion between the shifted code and the basic code for alphabetization purposes is obtained by a simple operation that forces the first bit to zero.

The proposed Iranian Standard Code for Information Interchange consists of 96 codes: 70 codes for the letters of the alphabet (a basic and a shifted code for each letter); 10 codes for the digits; 7 punctuation marks; and 9 special symbols. The organization of the ISCII table for 7-bit environments is shown in Fig. 3 where the first two columns are devoted to control codes. Fig. 4 shows the organization of the ISCII table for 8-bit environments where it can be used as an extension of ISO-IRV [3] for bilingual applications. The first two columns in this table are used for graphic symbols.

The following characteristics of the proposed ISCII should be noted:

(i) Codes are assigned to tanvin form of aleph as well as the hamzah forms of aleph and waw. Aleph is the only letter in the Farsi alphabet with the tanvin form. Aleph, waw, ha, and ya are the only Farsi letters with the hamzah form. The assigned codes are such that they maintain the alphabetization significance of these forms.

(ii) The Farsi dictionaries do not agree on the collating sequence value of the hamzah form of ya. Some place it at the beginning of the alphabet while others place it at the end. In the proposed ISCII, the hamzah form of ya appears at the end of the alphabet.

(iii) In Farsi typewriters, the composite "la" is located on a separate key and is displayed in a decidely different manner than the combination of its constituents "lam" and "aleph". Assigning a separate code to "la" would introduce complications in alphabetization. These complications are unnecessary in light of our assumption on the existence of intelligent output devices that could display the combination of "lam" and "aleph" either separately or in a composite form.

(iv) The hamzah form of "ha" appears only at the end of a word and is represented by two codes.

(v) All letters in the Farsi alphabet have a tashdid form. In the proposed ISCII the tashdid form of a letter is represented by the code for tashdid followed with the code for the letter. As such, text involving tashdid forms would require pre-processing for alphabetization purposes. Therefore, it would be more efficient to avoid tashdid for the majority of data processing tasks.

5. The Proposed Iranian Standard Keyboard Layout

The proposed Iranian Standard Code for Information Interchange is independent of the keyboard layout. However, the existence of a standard keyboard layout would greatly facilitate the use of computers. As such, in Fig. 5 we present

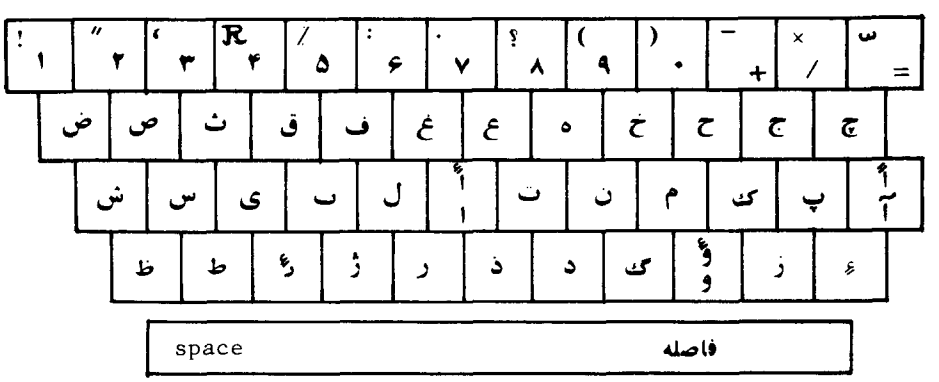


Fig. 5. The proposed Iranian Standard Keyboard Layout for computer applications.

our proposal for the Iranian Standard Keyboard Layout. In arriving at this particular layout, we have considered the following factors: similarity to Farsi typewriter keyboards; similarity to current computer keyboards in use; grouping of the alphabetical characters and other symbols; and assigning a separate key to each letter so that the shifted form is merely constructed by pressing that key while holding the shift key. The proposed Iranian Standard Keyboard Layout requires a total of 49 keys. Nearly twice as many keys may be found in most existing Latin computer keyboards, making it possible to overlay the Iranian Standard Keyboard Layout on such keyboards and thus allowing bilingual applications.

6. Concluding remarks

In this paper we have presented our proposals for the Iranian Standard Code for Information Interchange as well as the Iranian Standard Keyboard Layout. We have pointed out the special characteristics of the Farsi language that prevent the adoption of an approach similar to ASMO-449 which assigns a single code to each displayable character. We have emphasized the importance of defining a coded character set that lends itself easily to efficient implementation of alphabetization in view of the fact that multiple codes have the same alphabetization significance.

We have also argued for the existence of logical output devices that employ lexical analysis to decide on one of several possible shapes for each character code. This latter point implies that the displayable equivalent of a Farsi document in standard code may be several times longer since it would include control codes and escape sequences for a particular physical output device in order to cause it to display those shapes which do not have a unique character code and are identified through lexical analysis. It is important to recognize that while allowing for a variety of display technologies, the proposed Iranian Standard Code for Information Interchange is dependent on none.

References

[1] ASMO 449, "Data Processing – 7-bit Coded Arabic Character Set for Information Interchange", Arab Standard Specifications, No. 449-1982, 1982.

[2] International Standard Organization, "Code Extension Techniques for Use with the ISO Seven-bit Coded Character Set", ISO 2022, 1973.

[3] Iranian Plan and Budget Organization, "Final Proposal for the Iranian National Standard Information Code (INSIC)", (Farsi and English versions), 1981.

[4] C.E. Mackenzie, Coded Character Sets, History and Development, Addison-Wesley, New York, 1980.

[5] B. Parhami, "Standard Farsi Information Interchange Code and Keyboard Layout: A Unified Proposal", J. IETE, vol. 30, no. 6, pp. 179–183, 1984.

[6] M. Sanati, M. Dadashzadeh and M. Dadfar, "Iranian Standard Code for Information Interchange (ISCII)", Report to High Council of Informatics of Iran (in Farsi), 1986.