STANDARD FARSI INFORMATION INTERCHANGE CODE AND KEYBOARD LAYOUT : A UNIFIED PROPOSAL

BEHROOZ PARHAMI

Department of Mathematics & Computer Science, Sharif University of Technology, PB 3406 Tehran, Iran

Computer applications with a need for Farsi Language information processing suffer from the lack of several fundamental standards, of which the most pressing and far-reaching ones relate to an information representation and interchange code and a computer keyboard layout. In this paper, systematic extensions of the ISO (International) Standards Organization) 7-bit and 8-bit standard codes are proposed in connection with the Farsi language. The proposed standard codes possess a number of desirable properties which systematize and simplify the processing of Farsi and mixed (Farsi/Latin) information. Design considerations and technical justifications for the codeword selection problem are presented. A standard bilingual keyboard layout, which maintains the conventional typewriter format, to the extent possible, while eliminating its short-comings and inconsistencies, is proposed. Justifications for the various design decisions in arriving at the proposed layout are also given.

THE importance of standards for achieving compatibility between different computer systems cannot be overemphasized. Computer manufacturers, users, and professionals share the social and economic benefits of standardization, provided that standards are logical, systematic, mutually compatible, mature, practically realizable, and technically sound.

With respect to the computerized processing of the Farsi language, the most fundamental standard relates to an information representation and interchange code. Use of a standard code in all applications requiring the transfer of Farsi information between different systems will make unique interpretation of such information possible, without a need to know the characteristics of the sending and receiving systems. This type of inter-system communication is presently carried out through special translation routines.

Proliferation of computer systems, particularly in view of current trends in the utilization of low-cost office automation systems, word processors, and personal computers, makes such an approach highly inefficient in terms of system and human resources, since each computer must have a special routine for each of the different systems with which it might communicate. Even in systems with no need for external data exchange, a standard code must be used whenever possible for the sake of compatibility and ease of future modifications.

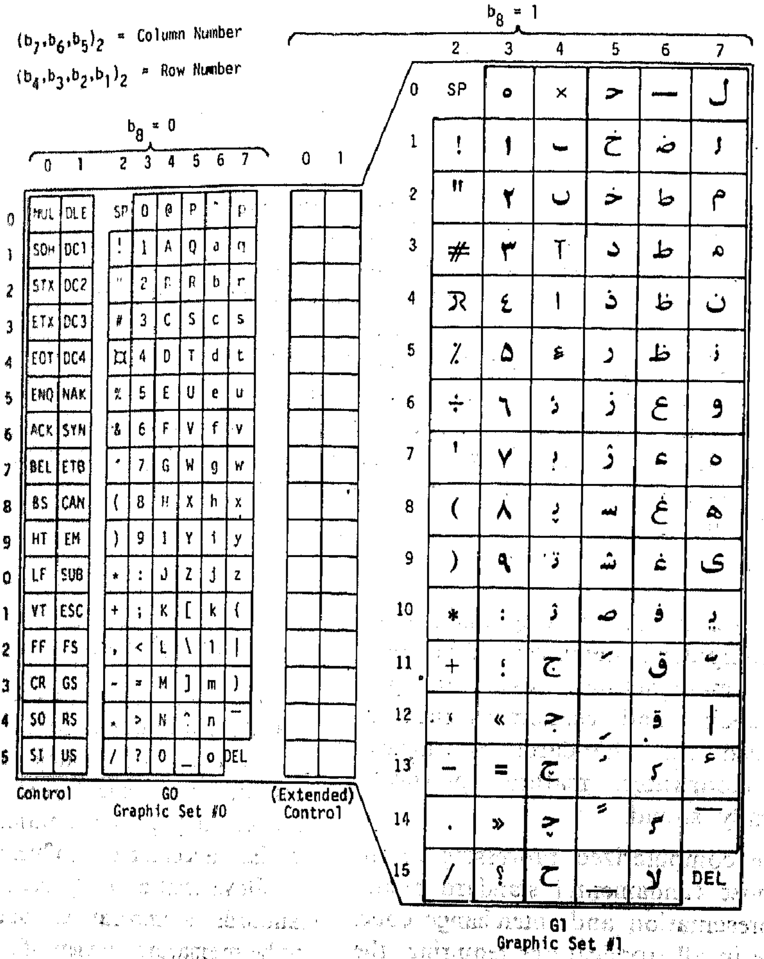
Of equal importance, is the adoption of a standard keyboard layout for the Farsi language. The lack of such a standard layout has resulted in costly replication of design efforts, retraining, and data entry inefficiencies. Computer manufacturers have selected the key positions in accordance with typewriter layouts (which are not necessarily optimal for computer applications) or with the objective of minimizing cost. In the latter case, some grossly illogical and inconvenient layouts have resulted. For example, one encounters situations which are in effect equivalent to assigning "A" and "a" to different keys on a Latin keyboard.

Needless to say that standardization efforts encounter great difficulties due to heavy investments in non-standard systems and their related technologies. Developers of standards are at times forced to ignore certain theoretically or practically proven advantages in order to make the standard acceptable to influential users and manufacturers. It follows that an early decision on adopting the needed standards is imperative. Such standards will definitely not be premature, in view of at least a decade of experience with the problems of Farsi language processing. Accounts of this experience have been presented [1-4]. Continued proliferation of non-standard systems will only complicate the situation and increase the number of compromises to be made for adopting a standard in future.

THE PROPOSED FARSI INFORMATION CODE

Standard information interchange codes have been in existence for many years. The most important such code, from the point of view of universality and international recognition, is the ISO-IRV code; *i.e.*, the International Reference Version of the code proposed by the International Standards Organization [5]. This code differs from ASCII [6] in only two positions: The currency sign (column 2, row 4) and the overline (column 7, row 14). The standard code proposed here is derived from ISO-IRV using standard extension techniques of the International Standard Organization [7].

The Farsi G1 set (graphic set 1) for the proposed standard code is shown in Fig. 1, along with the G0 set of the ISO-IRV code. The extended code, henceforth called "proposed INSIC", can be used in both seven-bit and eight-bit environments. In an eight-bit environment, the G1 set is either implied, by simply using $b_8=1$, or explicitly defined through an escape sequence. These



$(b_7,b_6,b_5)_2$ = Column Number

$(b_4,b_3,b_2,b_1)_2$ = Row Number

$b_8 = 0$

	0	1	2	3	4	5	6	7	0	1
0	NUL	DLE	SP	0	@	P	`	p		
1	SOH	DC1	!	1	A	Q	a	q		
2	STX	DC2	"	2	B	R	b	r		
3	ETX	DC3	#	3	C	S	c	s		
4	EOT	DC4	¤	4	D	T	d	t		
5	ENQ	NAK	%	5	E	U	e	u		
6	ACK	SYN	&	6	F	V	f	v		
7	BEL	ETB	'	7	G	W	g	w		
8	BS	CAN	(	8	H	X	h	x		
9	HT	EM	)	9	I	Y	i	y		
10	LF	SUB	*	:	J	Z	j	z		
11	VT	ESC	+	;	K	[	k	{		
12	FF	FS	,	<	L	\	l	|		
13	CR	GS	-	=	M	]	m	}		
14	SO	RS	.	>	N	^	n	‾		
15	SI	US	/	?	O	_	o	DEL		

Control GO Graphic Set #0 (Extended) Control

$b_8 = 1$

	2	3	4	5	6	7
0	SP	٠	×	حـ	ـ	ل
1	!	۱	ـ	خ	ضـ	لـ
2	"	۲	ب	خـ	ط	م
3	#	۳	آ	د	ط	مـ
4	ℛ	٤	ا	ذ	ظ	ن
5	٪	۵	ء	ر	ظ	نـ
6	÷	٦	ئـ	ز	ع	و
7	'	۷	بـ	ژ	ـعـ	ه
8	(	۸	پـ	سـ	غ	هـ
9	)	۹	تـ	شـ	غـ	ی
10	*	:	ثـ	صـ	فـ	یـ
11	+	؛	ج	َ	ق	ّ
12	،	«	جـ	ِ	قـ	|
13	−	=	چ	ُ	ک	ٔ
14	.	»	چـ	ً	گ	‾
15	/	؟	ح	_	لا	DEL

G1 Graphic Set #1

Fig 1 The Farsi G1 set for the proposed Iranian National Standard Information Code (INSIC), shown along with the G0 set of the ISO-IRV code for comparison

also apply to a seven-bit environment, where G0 and G1 are used, one at a time, by means of control characters S0 (Shift out, for mode change to G1) and SI (Shift In, for returning to G0).

The following is a list of major design issues for the proposed INSIC. Explanations are kept to a minimum for the sake of brevity. A report published by the Iranian Plan and Budget Organization [8] offers more detail.

Correspondence between code and output symbols

Here, the choice is between one-to-one correspondence of a full-code set (greatly simplifying output logic, record formatting, and tabulation procedures) and one-to-many correspondence of a reduced code set (increasing storage efficiency, and/or number of characters representable, by assigning a single code to the various written forms of each Farsi letter). The use of context for determining the correct written form of each letter and the necessity of defining "pseudo space" and "pseudo connection" characters to over-ride contextual information for representing "separated" and "connected" Farsi letters, respectively, are well-known. In the proposed INSIC, there is a one-to-one correspondence between coded characters and printed or displayed symbols. The main reason for this choice is that application of the standard code (at least in the near future) will be limited to the interchange of information between various computer systems. Manufacturers will continue to use special internal representations suited to the capabilities and limitations of their particular hardware subsystems. For most such internal representations, there is a one-to-one correspondence (with respect to the number of positions taken) between code symbols and symbols actually printed or displayed. At present, no method is known for converting information between one-to-many and one-to-one representations without disturbing record formats or deranging tabulations.

Completeness

An attempt has been made to include all common Farsi symbols, plus special characters needed for various applications, in the proposed INSIC. Thus, all Farsi punctuation marks (including open and closed Farsi quotation marks, "≪" and "≫"), common mathematical symbols, several international special symbols which may also be used in Farsi (*e.g.*, the symbol in column 2, row 7, usable both as a separator for grouping digits in a numeric field and as the Arabic "short aleph"), commercial "Rial" and "per cent" signs (column 2, rows 4 and 5), vowels and other special symbols for Farsi texts (heretofore totally ignored in computer-based information systems), and finally, the extension bar (used in Farsi for elongating connected letter forms) have been included. That several of the special symbols from G0 have also been included in G1 may seem inappropriate for an eight-bit environment, where all symbols in G0 and G1 are directly available to the user. However, since in dealing with Farsi information fields; one often has to invert them for storage and output purposes, special symbols within a Farsi field entered through a mode change to Latin would make the recognition of Farsi field boundaries extremely difficult. In addition, such duplication enables independent usage of the proposed Farsi code, without a need for bilingual capability in computer systems.

Compatibility

The assignment of G1 symbols is completely compatible with that of G0. This is at least partly due to good luck. The fact that exactly 52 symbols are needed for representing all the variants of the 33 Farsi alphabetic characters (by using a lengthwise decomposition in some cases), makes the assignment of these to positions corresponding to uppercase and lowercase Latin alphabetic symbols a natural choice. Positions of Farsi digits in G1 are dictated by packing and unpacking efficiency, ease of performing arithmetic, and software portability considerations. Other than alphabetic symbols, digits, and punctuation marks (which despite differing shapes, appear in the same positions of G0 and G1), only the following changes have been applied to the ISO-IRV code table: The "Rial" sign for currency symbol, division symbol for "&" (in Farsi, "slash" is used as the Latin "decimal point"), open and closed Farsi quotation marks for "<" and ">", multiplication symbol for "@", Farsi vowels for "[", "/"and "]", "tanvin" for the symbol in column 5, row 14, extension bar for the symbol in column 6, row 0, and finally, "tashdid" and superscript "hamza" for "{" and "}". An important by-product of this compatible code assignment shows up for systems normally working in Latin mode and occasionally changing over to Farsi mode due to the small volume of Farsi applications (*e.g.*, changing the printer chain once a day). The proposed code allows programmers to perform a substantial amount of their system development and debugging with no need for Farsi output capability, since many of the essential features of their outputs (*e.g.*, numeric and non-numeric fields, most special symbols, table frames, and other general information) show up on the Latin printout or display.

Sorting of data

It has been shown [2,3] that existing problems in the sorting of Farsi data can not be solved merely by suitable encoding of symbols. Therefore, the best that can be done is to maintain the natural ordering of the alphabetic symbols and to assign the "flat tail" and "round tail" symbols (column 4, rows 1 and 2) to positions where they would cause fewest problems. The temptation to eliminate the composite "la" (column 6, row 15) one of the sources of difficulties in sorting, should be resisted due to this symbol's fundamental role in generating aesthetically pleasing Farsi script and the importance of one-to-one correspondence between code and output symbols (as discussed previously).

Logical consistency

Logical consistency is desirable both to simplify Farsi processing algorithms and to achieve the technology independence which is so important in view of the rapid pace of change in the computer field. Specifically, two changes in the conventional treatment of Farsi alphabetic symbols have been found necessary to achieve these goals. The first change is to provide "separated" and "connected" forms for the letters "ta" and "za" (column 6, rows 2 through 5). Heretofore, Farsi typewriters and computer output devices have used a single "connected" form for each of these letters, with the "separated" form represented by the single form followed by a blank. However, this is grammatically incorrect and also causes difficulties in text processing applications where blanks indicate word boundaries. The second change is to provide single "separated" and "connected" forms (rather than two of each) for the letters "ain" and "ghain" (column 6, rows 6 through 9). Thus, the following rule has been observed consistently: A letter which is connectable to the next letter has two forms, while one that is not connectable has a single form. This same rule precluded the use of two forms for "aleph", as done on some systems for aesthetic effect. Note that the proposed INSIC applies to represention codes and not symbol shapes. Thus, despite the fact that only two codes have been assigned to "ain", an intelligent output device capable of producing all four shapes of this letter can deduce the correct shape based on the preceding character (with no delay). At the same time, by accepting slightly less pleasant results, very simple low-cost output devices can be used for generating the Farsi script.

THE PROPOSED KEYBOARD LAYOUT

Just as typewriter keyboard standards were developed to facilitate the training of typists, the need for relative uniformity in keyboard layout for terminals and other computer input devices has resulted in the adoption of similar standards. Due to the lack of a Farsi computer keyboard standard, manufacturers have taken their pick. Some have followed the "ease of implementation" path (changing key caps of a Latin keyboard to Farsi ones), thus sprinkling Farsi symbols on the keyboard in a seemingly random pattern. Others have been more logical, either selecting the typewriter keyboard layout proposed

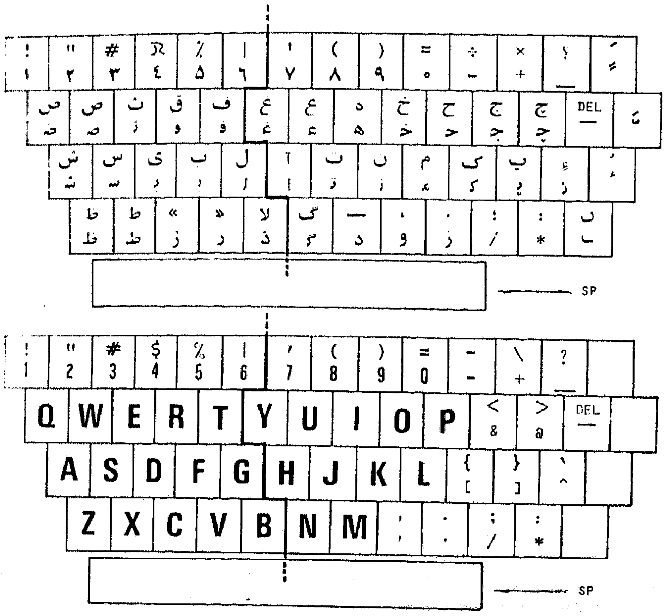


Fig 2 Keyboard layout for Farsi and bilingual data entry in connection with the proposed information interchange code.

by the Institute of Standards and Industrial Research of Iran [9] or opting for a particular typewriter manufacturer's design. But typewriter layouts are not necessarily optimal for computer applications; in fact there is strong evidence that they are not even optimal for typing!

The Farsi keyboard layout, designed for use with the proposed INSIC of Fig 1, is shown in Fig 2. A list of major design considerations for the proposed layout follows.

Familiarity

Standard Farsi typewriter keyboard layout has been chosen as the basis of the design, with symbols kept in their original positions, except when this would seriously compromise the other design goals. Due to the eminent proliferation of word processors, typing and computer data entry will be considered identical activities in the near future. Having a single keyboard layout standard for typewriters and computer input devices will ease the transition. A total re-arrangement of typewriter key positions to obtain the most efficient single standard is not warranted, in view of the potential speed gain and the estimated cost of a massive retraining effort. Pronouncing the first six Farsi alphabetic symbols on the bottom row (à la "qwerty"), the proposed layout may be called "zoodgozar"; a Farsi word which means "passing" (as a fad). Well, I hope there is no truth to this coincidental meaning!

Ease of use

This factor has affected the design in two ways. First, the letters on fourth-row (top) keys have been moved to the other rows and placed where they can be most easily remembered. Digits have been assigned to the more convenient unshifted positions and arranged in natural left-to-right order. A number of simple rules, shown graphically in Fig 3, help in memorizing the changes for the transition period. Figure 3 also shows the greater logical consistency of the proposed layout compared to those of present typewriters. Second, each Farsi letter can be entered by a single key depression, in spite of the fact that some letters are represented by two symbols (the second one being a flat or round tail) according to the standard code. Automatic transmission of the needed tail symbol is something that can easily be achieved with present-day keyboards. The logical consistency and three-row arrangement of alphabetic keys along with unshifted digits should contribute to ease of use and data entry speed.

Completeness

Every single symbol of the proposed INSIC appears on the keyboard of Fig 2. The only exception is the substitution of "$" sign for the general "currency" sign, due to its usefulness for international banking and trade. The full keyboard has a 53-key format, as shown. It even

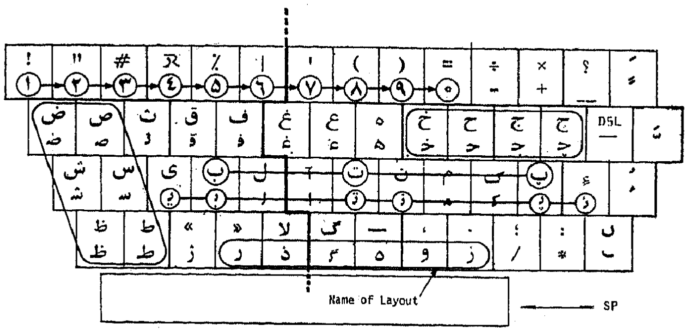


Fig 3 A graphical representation of the relationships between similar letters in the modified standard Farsi typewriter keyboard layout proposed for computer applications.

contains the two tail symbols for completeness, although they are not required for entering the letters represented in composite form. Thus each of the 96 symbols in G0 or G1 can be entered. The keys at the right end of each of the four rows are only used for G1 and not for G0. Deleting these four keys, results in a 49-key arrangement which is sufficient for most applications. Deleting the rightmost key of the 49-key arrangement (the one with "DEL" and "overbar" symbols), yields a standard 48-key format as the minimal standard subset.

Bilingual data entry

Many applications require bilingual data entry capability. This is facilitated by the proposed keyboard layout in two ways. First, equivalent and similar symbols of G0 and G1 have been assigned to identical key positions, while maintaining the standard format of Latin keyboards to the extent possible. This facilitates the memorizing of key positions by bilingual data entry operators and reduces the chance of confusion and error. Second, neater appearance and greater functionality results from the fact that each key requires fewer labels. Identical symbols of G0 and G1 need only be shown once. The 26 keys assigned to the two forms of single Farsi letters can be labelled with the "separated" (shifted) form, advising the users that connected forms are to be entered unshifted. This reduces the number of labels per key cap from about 3.5 in conventional designs to about 2.6 for the 48-key arrangement. This can be further reduced to about 2.3 if equivalent punctuation marks and digits appear in only one language. This will definitely be appreciated by those who have worked with crowded-looking bilingual keyboards. By using the front face of key caps and a different colour for symbols of the second or auxiliary language, confusion and error may be reduced considerably.

CONCLUSION

In this paper, the design of a standard Farsi information interchange code (proposed INSIC) and keyboard layout was given a unified treatment, emphasizing factors such as functionality, completeness, ease of use, compatibility, and logical consistency. Even though the discussion was in terms of the Farsi language, many of the issues raised apply to Arabic and Urdu as well. In fact, only with joint international effort is the full social and economic benefits of such standards realizable.

Time is running out on us. Word processors and personal computers are mushrooming everywhere. Such systems are, by their very nature, quite language dependent. Word processors deal with natural-language texts and personal computers are used, among other things, for maintaining mailing lists, forwarding electronic mail, and accessing natural-language data banks. A standardized Farsi information interchange code and keyboard layout is thus urgently needed.

REFERENCES

1. B. Parhami & F. Mavaddat, Computers and the Farsi language: a survey of problem areas. *Information Processing 77* (Proc of IFIP Congress). 1977, North-Holland, Amsterdam, pp 673-676.
2. B. Parhami, Impact of the Farsi language on computing in Iran. *Mideast Computer*, vol 1, pp 6-7, 18 September 1978.
3. B. Parhami, On the use of Farsi and Arabic languages in computer--based information systems. *Proc symp on linguistic implications of computer-based information systems*. 1978. New Delhi, India.
4. B. Parhami, Language-dependent considerations for computer applications in Farsi and Arabic speaking countries; in *System approach for development* (Proc Third IFAC Symp.), North-Holland, Amsterdam. pp 507-513 1980.
5. International standards organization, Seven-bit coded character set for information processing interchange, 1973, ISO 646.
6. ASCII—American national standard Code for information interchange, 1977, ANSI X3. 4.
7. International standards organization, Code extension techniques for use with the ISO seven-bit coded character set, 1973, ISO 2022.
8. Iranian plan and budget organization, Final proposal for the Iranian national standard information code (INSIC), 1980, Farsi and English versions.
9. Institute of standards and industrial research of Iran, character arrangement on keyboards of Persian typewriters (in Farsi), 1976, ISIRI 820.