LANGUAGE-DEPENDENT CONSIDERATIONS FOR COMPUTER APPLICATIONS IN FARSI AND ARABIC SPEAKING COUNTRIES

B. Parhami

Department of Mathematics and Computer Science,
Sharif University of Technology, Tehran, Iran

Abstract. A prerequisite to economic viability and social acceptance of computer-based systems, particularly in developing countries, is their ability to accept and present information in the natural language of the user community. This paper identifies problems relating to the use of Farsi and Arabic languages in computer-based systems and outlines known approaches to their solutions. The problems discussed fall into four main categories: (1) Information representation and standardized information interchange codes; (2) Input of information through devices such as keyboards and optical readers; (3) Processing functions such as sorting, text formating and data compression; and (4) Output of information on hard-copy devices and various types of displays.

Keywords. Character recognition; Computer applications; Computer input and output; Computer peripheral equipment; Data processing; Display systems; Human factors; Printers; Standards; Text editing.

INTRODUCTION

Computers are information processing machines and much of the information we deal with in our everyday lives is generated, maintained, and used in a natural language. Furthermore, for computers to be useful, they must interact with human beings and the most convenient way of doing this, at least as far as ordinary users are concerned, is by utilizing a natural language as the medium of communication. It follows that the ability of computers to deal with information presented in a natural language is essential for their successful utilization in most environments.

Even though language considerations are not critical in activities such as programming, there are clearly instances where information storage (e.g., in the case of names and addresses), manipulation (e.g., in text processing), and presentation (e.g., for questic.ı-answering systems) can be most effectively handled in the users' native language. It is, therefore, quite natural that for a number of years, we in Iran have been dealing with language-related problems created by the rapid expansion of computer applications (Parhami and Mavaddat, 1977).

In this paper, some basic problems relating to the encoding, input, processing and output of information in Farsi and Arabic languages are identified and approaches to their solutions are outlined. Our discussion will be in terms of the Farsi language, with the understanding that almost all of these considerations apply to Arabic and Urdu as well.

ENCODING OF INFORMATION

The recording of information on computer storage media and its transmission from one place to another (be they circuits on the same board or geographically distant computer systems) usually requires some form of digital encoding. The advantages of using standard codes for these purposes are well known and have resulted in the adoption of national and international standards for various natural languages. National standards of this type must be designed according to specific guidelines, put forth by the International Standards Organization (ISO, 1973), if they are to be mutually compatible.

Following relative maturity in various application areas, the need for a standard represention of Farsi symbols was felt in Iran (Mavaddat and Parhami, 1977). In February of 1978, the Iranian Plan and Budget Organization invited a number of computer and language experts to form the "Study Committee for Standard Farsi Information Code." Even though the Committee's final report was published in August of 1979 (SCSFIC, 1979), it has not yet been adopted as a national standard.

Based on ISO's code extension guidelines, the committee set out to define a 94-symbol G1 graphic set which can either replace the standard G0 set in a seven-bit code or used in conjunction with it in an eight-bit environment (Figure 1). The following five points were taken into account in designing the Farsi G1 graphic set:

1. One-to-one correspondence between code and output symbols. This point, which proved quite important in practice, precluded the use of a reduced code set (Hyder, 1972), as originally envisaged in the Committee's preliminary proposal.

2. Inclusion of all common Farsi alphabetic and special symbols. The proposed set includes all needed numeric, alphabetic, punctuation, mathematical and commercial symbols as well as vowels which may be needed in special linguistic applications.

3. Proper alphabetic ordering of symbols for ease of sorting. Even though it has been shown that the Farsi sorting problem cannot be solved by code assignment alone (Parhami, 1978c), care has been taken to avoid complications whenever possible.

4. Maintaining the correspondence between Farsi symbols and equivalent or similar symbols in the G0 set to the extent possible. The particular set of alphabetic symbols used turned out to contain 52 members, thus exactly replacing the upper and lower case letters in G0 (Figure 1).

5. Simple recognition of numeric and alphabetic symbols and efficient packing and unpacking of numeric information. This is accomplished by assigning numeric and alphabetic sybmols to positions where G0 contains symbols of the same types.

A complete presentation of the tradeoffs involved in designing the Iranian National Standard Information Code (INSIC) and justification of the particular decomposed character set used is beyond the scope of this paper. Interested readers are referred to the Study Committee's final report for a discussion of these points as well as extensive application notes.

INPUT OF INFORMATION

The existing problems with Farsi input and some of the approaches for dealing with them have been enumerated by Parhami and Mavaddat (1977). With respect to the input of natural language information, we can visualize three distinct possibilities at the present:

1. Keyboard data entry.
2. Document data entry.
3. Voice data entry.

The third alternative is in the research and experimental development stage even for languages of technologically advanced countries. The Farsi language will undoubtedly present its own unique problems in this respect.

The second alternative, though in practical use for languages utilizing the Latin alphabet, is relatively new in the case of Farsi. The difficulties in the automatic recognition of Farsi texts are depicted in Figure 2. Parhami and Taraghi (1980) have developed algorithms for separating the symbols in connected subwords of printed Farsi texts and for recognizing the symbols from their geometric properties. Figure 3 shows the result of their algorithm applied to a three-letter subword.

The insight gained from the above-mentioned study will be helpful in the design of type fonts for various document input devices. However, the development of standards in this area does not seem urgent at present.

For most applications, the initial data entry is, and will be for the foreseeable future, through the first alternative: namely, keyboards. Even though there is no technical reason for using the present typewriter keyboard layout for computer applications, the practical factor of familiarity has prompted most manufacturers to adopt the same layout and may continue to influence keyboard designs for years to come. Nevertheless, a total reorganization may be considered desirable or even necessary in future, based on one or both of the following considerations:

1. The known inefficiency of the present typewriter layout for Farsi data entry.

2. Large scale replacement of Farsi typewriters by word-processing machines.

Figure 4 shows a proposed keyboard layout for use with the Iranian National Standard Information Code in Farsi and bilingual data entry. This layout has been proposed by the "Study Committee for Standard Farsi Information Code" (SCSFIC, 1979) based on the following considerations:

1. Maintaining the typewriter key positions to the extent possible, unless where this is in conflict with other goals described below. A total of eight changes turned out to be necessary. These changes made the keyboard layout more systematic, more efficient and easier to remember and also enabled the numerals to be placed on lower case.

2. Inclusion of all the standard symbols, and only these symbols. The only exceptions are Farsi letters which are represented by two symbols in the standard code due to the requirement for one-to-one correspondence between code and output symbols. These are entered by a single key stroke, with the needed tail symbols added automatically.

3. Placing equivalent and similar Farsi and Latin symbols in corresponding positions, to the extent possible, in order to reduce the amount of information appearing on each key and to facilitate operator training for bilingual data entry. This was accomplished quite nicely with minimal changes to conventional Farsi and Latin layouts (Figure 4).

PROCESSING OF INFORMATION

The three processing functions of sorting, text formating and data compression are briefly discussed in this section. With respect to sorting, it has been shown that the problem cannot be solved by code assignment alone, even if a reduced code set is used (Parhami, 1978c). Therefore, we either have to adjust ourselves to a new notion of sorting, which might appears somewhat unnatural, or resort to more sophisticated algorithms.

In the area of text formating, two complicating factors are observed:

1. Hyphenation is not allowed for breaking long words.
2. Words are not consistently separated by blanks.

The second problem arises since some words appear naturally disconnected when juxtaposed, enabling a human reader to determine word boundaries by paying limited attention to the context and/or meaning.

Although in the long run this problem can be rectified by proper training, the current difficulties will be with us for some time. Fortunately, however, complete decomposition is not needed in text formating applications, as long as the word boundaries that are actually recognized are not too far apart. This is usually true in practice (Nasseri, 1978). On the positive side, the possibility of extending connected Farsi letters is at times helpful for text justification.

In the area of data compression, a recent study (Tafaghodi Jami, 1980) has shown promising results. With suitable text fragment selection algorithms, a compression ratio of 45 percent was achieved. When a reduced code set enabling the encoding of more text fragments was used, the compression ratio improved to 48 percent. These results are somewhat better than those achieved for English texts.

The final aspect of processing discussed here is that of algorithm specification. There have been suggestions that we should be thinking about a high-level programming language (similar to COBOL) based on Farsi. Even though on the surface it may appear that the English-language orientation of many high-level programming languages is a deterrent to Farsi speaking programmers, there is no evidence that a Farsi-based language will make the programmer's task any easier or that our computing community will be better off in general as a result.

In response to the argument that, in the near future, interaction with computers will be done mostly by non-programmers which might be aided by the availability of a Farsi programming language, it can be stated that non-programmers are not likely to interact with computers in languages like FORTRAN and COBOL. They will probably perform simple computations using a mathematically oriented language (or simply function keys on a private, pre-programmed keyboard) or define a system by filling tables and answering multiple-choice questions.

Any new language proposed will need to be initially developed and documented for various computer systems and subsequently maintained and updated periodically. Clearly, our limited resources are better spent on developing more effective computer applications than on designing new programming languages and systems. Furthermore, for the exact reason that so many programming languages are in use today (i.e., the diversity of application areas), no single Farsi-based language can satisfy all needs. How many of these languages should we develop? The most logical answer appears to be zero!

OUTPUT OF INFORMATION

With respect to the output of natural language information, three areas need to be investigated:

1. Hard copy output of information.
2. Visual display of information.
3. Special methods (voice, braille, etc.).

We will not discuss the third area here except for saying that its importance will certainly draw some of our attention once the pressing problems in the other two areas are satisfactorily solved.

While the difficulties with Farsi codes, input and processing are of concern to computer professionals and a limited number of direct users, Farsi output is a problem of broad concern. In view of the fact that over the centuries of Islamic influence, writing of Farsi has become something of an art, extensive modification of the Farsi script for adaptation to machine printing or display is highly undesirable and has already caused some damage in the public attitude towards computers (Parhami and Mavaddat, 1977).

Most high-speed Farsi line printers produce low quality output due to the following:

1. Inter-symbol gaps make the technology of most high-speed printers unsuitable for producing the connected symbols of the Farsi script.

2. The varying widths of Farsi alphabetic symbols are either ignored for simplicity or dealt with by decomposing wider characters into two parts. The first approach reduces the readability and aesthetic quality of the resulting output while the second approach compounds the problems of connectivity and horizontal alignment.

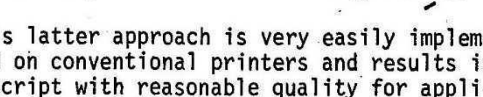
3. Due to the large symbol set needed for printing Farsi, frequently some of the special symbols or variants of letters are eliminated. The need for mixed Farsi/Latin print in some applications has compounded this problem.

4. The similarity of many Farsi symbols, differing only in the number or places of small dots, necessitates a higher quality printout for readability. The smearing effect of most high-speed impact printers is, therefore, a serious drawback in this respect.

Farsi printer output samples in Figure 5 illustrate the above problems. Many of these difficulties can be easily overcome with character printers. Even though quite valuable in word processing applications, such devices are relatively slow and thus only of limited interest in most other environments.

For high-speed Farsi output, the new generation of non-impact printers appear to be promising, since by their method of operation (e.g., thermal, electrostatic, or xerographic processes), such devices produce sharp and highly readable output with no inherent inter-symbol gap. The only adjustment is for the need of a larger frame in the case of dot-matrix printers, since five by seven or even seven by nine matrices are inadequate for Farsi.

In the actual implementation of output devices, a great deal of flexibility can be provided. As long as the output device accepts a standard code set, it can use as many different forms for each letter as deemed necessary for a reasonable output script quality. Vowels and similar special symbols can either be overprinted on the preceding letter or form an independent symbol, with connected and separate forms, as shown in the following examples:



This latter approach is very easily implemented on conventional printers and results in a script with reasonable quality for applications where vowels and supplementary symbols need to be printed without cost or speed penalties.

In the visual display of Farsi information, we essentially face the same difficulties as in hard copy output. In the case of visual display units associated with intelligent terminals, the needed changes for adaptation to the special requirements of Farsi are usually not very complex. The comments made on dot-matrix printers also hold true for dot-matrix display units.

The representation of Farsi numeric and alphanumeric information on line-segment displays is also of some interest (Parhami, 1976). A prototype display unit for a single 20-character line of Farsi output is at the final stages of construction at Sharif University of Technology. The display uses 18-segment character displays as shown in Figure 6.

Even though purely numeric information can be displayed with as few as seven segments (Parhami, 1976), no design with fewer than 18 segements is known for Farsi alphabetic symbols.

In some applications, it may be desirable to present large quantities of numeric information in combined digital/analog form. This can be achieved through the use of an optically weighted numeral font. Figure 7 shows an example for decimal Farsi numerals in a seven by five matrix. The area covered by a numeral x is ax+b (i.e., linearly proportional to x) with a=3 and b=4 (b=3 if the points marked with an "o" are deleted). General guidelines for the design of such numeral fonts as well as several specific designs have been presented elsewhere (Parhami, 1978a).

CONCLUSION

Satisfactory solutions to the problems enumerated in this paper are important for the successful development of informatics technology in Iran as well as in the Arab world. In particular, immediate action on the standardization of information interchange codes is needed to assure compatibility of systems and to avoid costly replication of effort.

In the area of input devices, standard keyboard layouts need to be developed. The design of better character recognition algorithms as well as OCR type fonts is also needed, though it is less urgent.

For effective processing of Farsi information, the problems of sorting, text formating and data compression need to be further studied. The inital results in these areas are quite encouraging and point to the possibility of significant advances in the near future.

A considerable amount of effort has so far been placed on the Farsi output problem as the most immediate user concern. Hard copy devices (both impact and non-impact types) are rapidly improving in this respect. Visual display units have generally relied on large dot matrices (say with more than 150 points) to generate the Farsi symbols. This situation is not likely to improve soon and thus line-segment displays may become a viable alternative when cost is a prime factor.

As we solve the fundamental language-related problems facing us in the field of computing, we may in future attack less crucial, but perhaps more interesting, problems of generating classic Farsi scripts and analyzing the rich heritage of Farsi literature with the aid of computers. The artistic and scholarly possibilities of such undertakings are almost limitless (Parhami, 1978b).

REFERENCES

Hyder, S.S. (1972). A system for generating Urdu/Farsi/Arabic script. Information Processing 71 (Proc. of IFIP Congress). North-Holland, Amsterdam, pp. 1144-1149.

Institute of Standards and Industrial Research of Iran (1976). Character arrangement on keyboards of Persian typewriters. Document No. ISIRI 820.

International Standards Organization (1973). Code extension techniques for use with the ISO seven-bit coded character set. Document No. ISO 2022.

Mavaddat, F. and B. Parhami (1977). Informatics in Iran: problems and prospects. Proc. of the International Conf. on Computer Applications in Developing Countries, Bangkok, Thailand, pp. 121-133.

Nasseri, P. (1978). Automatic processing of Farsi texts. Master of Science Thesis, Sharif University of Technology, Tehran, Iran (in Farsi).

Parhami, B. (1976). Low-cost output displays for microcomputer applications. Proc. of the Second India Symp. on Computer Architecture and System Design, New Delhi, pp. 111-119.

Parhami, B. (1978a). Optically weighted dot matrix Farsi and Arabic numerals. Proc. of the Third Jerusalem Conf. on Information Technology. North-Holland, Amsterdam, pp. 207-210.

Parhami, B. (1978b). Impact of the Farsi language on computing in Iran. Mideast Computer, Vol. 1, No. 1 (Sep. 18), pp. 6-7.

Parhami, B. (1978c). On the use of Farsi and Arabic languages in computer-based information systems. Proc. of the Symp. on Linguistic Implications of Computer-Based Information Systems, New Delhi, India.

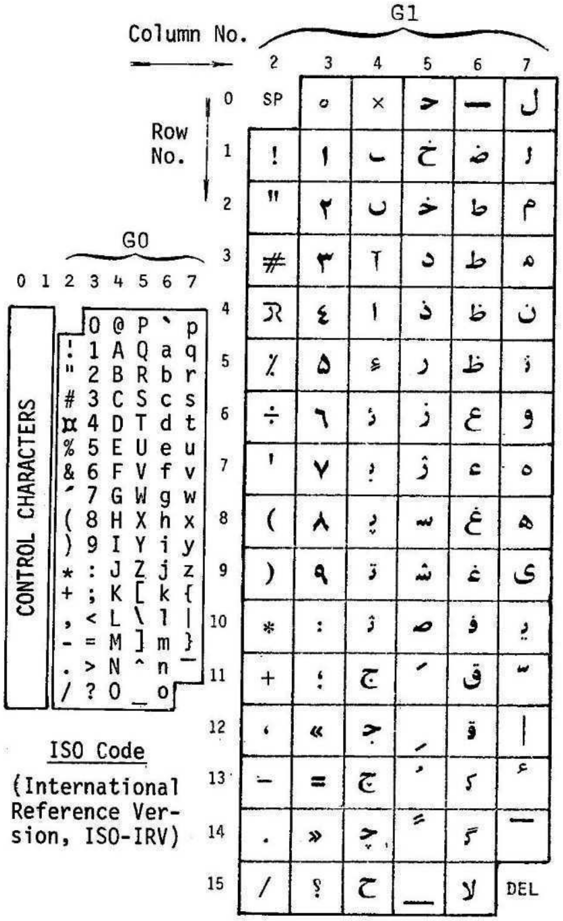
Parhami, B., M.J. Ashjaee, and F. Mavaddat (1977). The Farsi language and computers. Reports prepared for Plan and Budget Organization of Iran under Research Contract No. 10908201/9/120. Vol. 1, Nov. 1977. Vol. 2, June 1978 (in Farsi).

Parhami, B. and F. Mavaddat (1977). Computers and the Farsi language: survey of problem areas. Information Processing 77 (Proc. of IFIP Congress). North-Holland, Amsterdam, pp. 673-676.

Parhami, B. and M. Taraghi (1980). Automatic recognition of printed Farsi texts. Pattern Recognition (to appear).

Study Committee for Standard Farsi Information Code (1979). Final proposal for the Iranian National Standard Information Code (INSIC). Informatics Division, Plan and Budget Organization of Iran, Tehran (Farsi and English versions).

Tafaghodi Jami, A. (1980). Efficiency considerations in the storage and retrieval of Farsi texts. Master of Science Thesis, Sharif Univ. of Technology, Tehran, Iran (in Farsi).



Column No.

Row No.

G0

Row	0	1	2	3	4	5	6	7
0	CONTROL CHARACTERS			0	@	P	`	p
1			!	1	A	Q	a	q
2			"	2	B	R	b	r
3			#	3	C	S	c	s
4			¤	4	D	T	d	t
5			%	5	E	U	e	u
6			&	6	F	V	f	v
7			'	7	G	W	g	w
8			(	8	H	X	h	x
9			)	9	I	Y	i	y
10			*	:	J	Z	j	z
11			+	;	K	[	k	{
12			,	<	L	\	l	|
13			-	=	M	]	m	}
14			.	>	N	^	n	‾
15			/	?	O	_	o	

ISO Code

(International Reference Version, ISO-IRV)

G1

Row	2	3	4	5	6	7
0	SP	٠	×	ﺣ	ـ	ل
1	!	١	ﺑ	خ	ﺿ	ﻟ
2	"	٢	ں	ﺧ	ط	م
3	#	٣	آ	د	ﻃ	ﻣ
4	ℛ	٤	ا	ذ	ظ	ن
5	%	٥	ء	ر	ﻇ	ﻧ
6	÷	٦	ﺋ	ز	ع	و
7	'	٧	ﺑ	ژ	ﻌ	ه
8	(	٨	ﭘ	ﺳ	غ	ﻫ
9	)	٩	ﺗ	ﺷ	ﻐ	ی
10	*	:	ﺛ	ﺻ	ﻓ	ﻳ
11	+	؛	ج	َ	ق	ّ
12	،	«	ﺟ	ِ	ﻗ	|
13	-	=	چ	ُ	ک	ٔ
14	.	»	ﭼ	ً	گ	‾
15	/	؟	ح	_	لا	DEL

Fig. 1. The Proposed Iranian National Standard Information Code (INSIC).

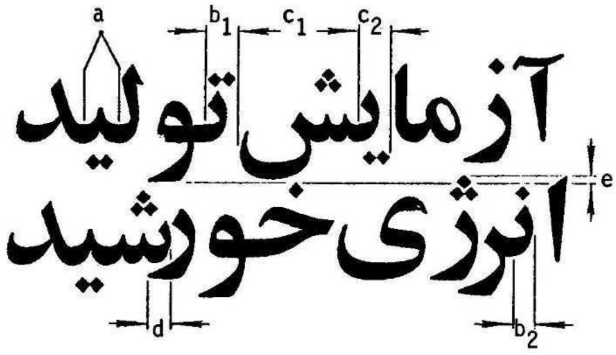


Fig. 2. Difficulties in the automatic recognition of printed Farsi texts:
(a) Connectivity of symbols,
(b) Similarity of symbols,
(c) Variable-width symbols,
(d) Overlapping subwords, and
(e) Overlapping lines of texts.

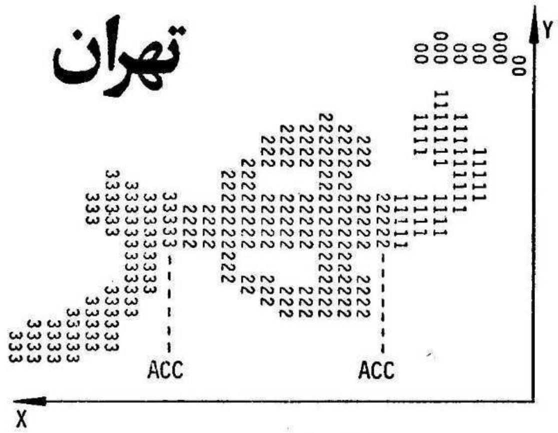


Fig. 3. Seperation of symbols within a subword in the digitized text. Actual connection columns (ACC's) are selected from a set of potential connection columns (PCC's). Dots are marked and dealt with separately.

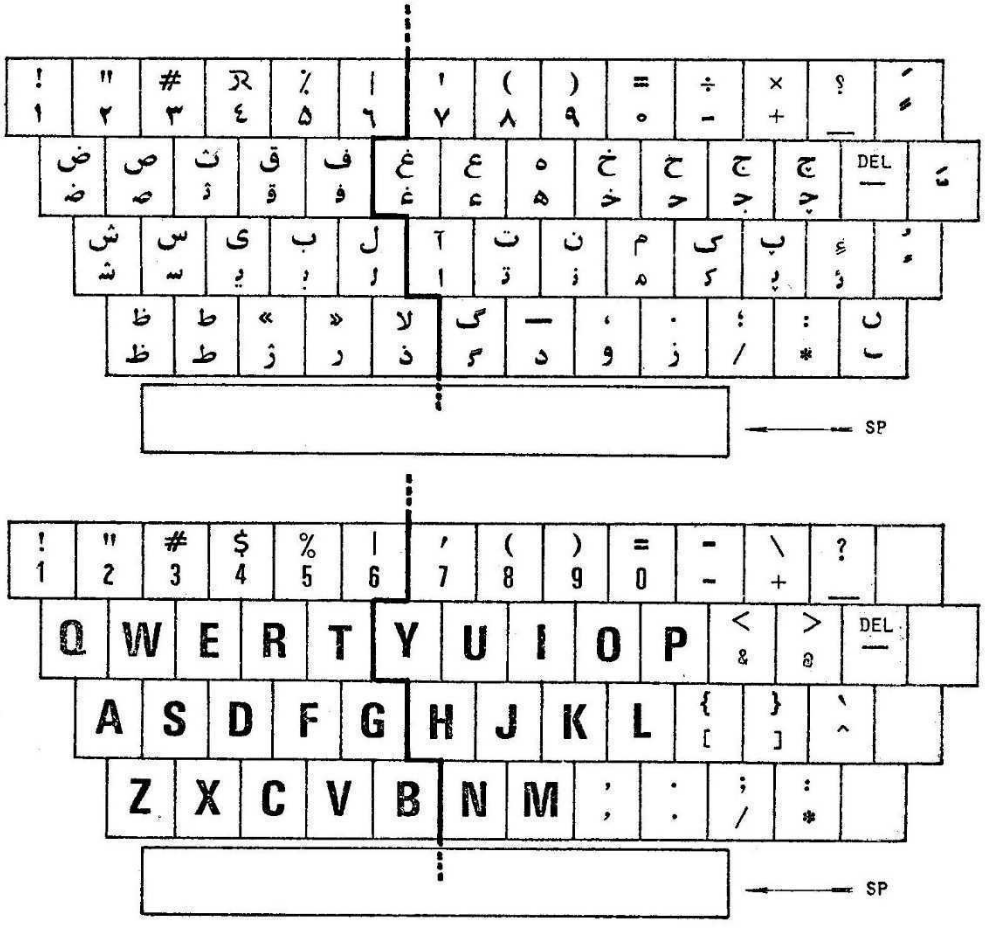
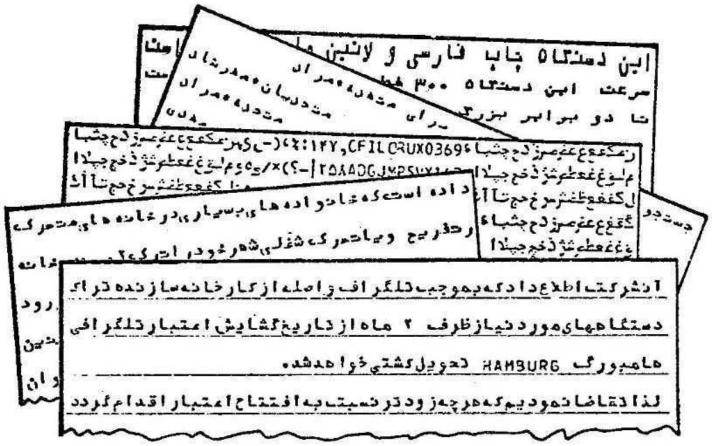


Fig. 4. The proposed layout for Farsi and bilingual (Farsi/Latin) keyboards.



این دستگاه چاپ فارسی و لاتین

سرعت این دستگاه ۳۰۰ خط

تا دو برابر بزرگ

آنشرکت اطلاع داد که بموجب تلگراف واصله از کارخانه سازنده تراک

دستگاههای مورد نیاز ظرف ۲ ماه از تاریخ گشایش اعتبار تلگرافی

هامبورگ HAMBURG تحویل کشتی خواهد شد.

لذا تقاضا نمودیم که هرچه زودتر نسبت به افتتاح اعتبار اقدام گردد

CFILCRUX0369
DG JMPSVY

Fig. 5. Samples of hard-copy Farsi output, illustrating some of the difficulties in producing high-quality Farsi output on conventional printers.

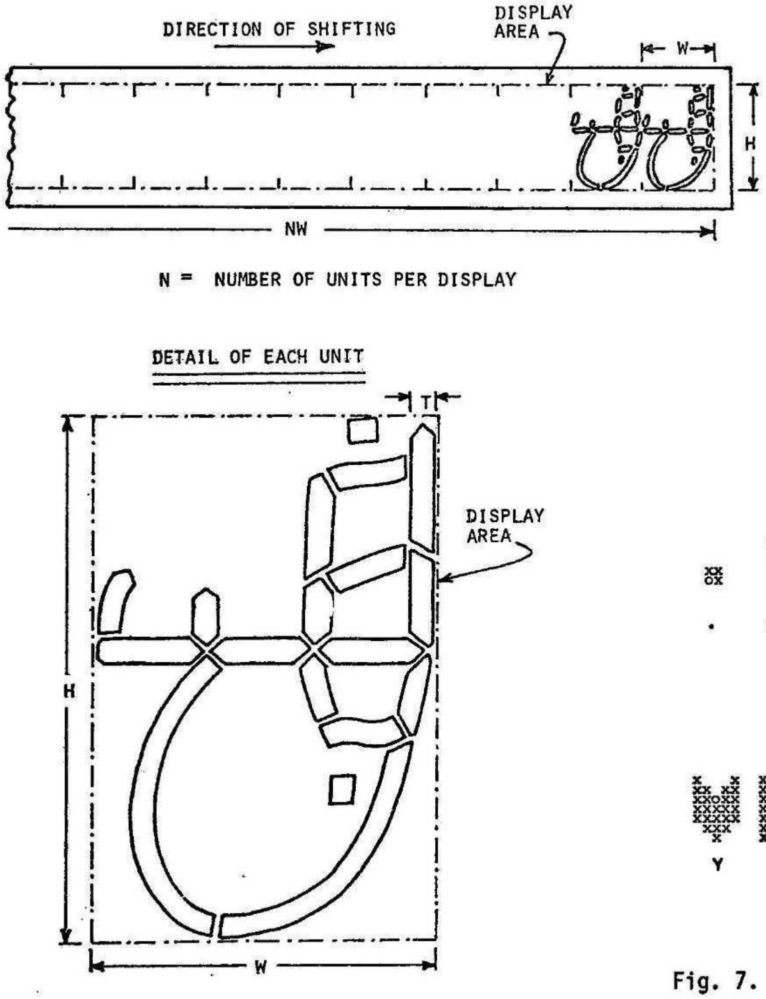


Fig. 6. The proposed line-segment display for Farsi alphanumeric information.

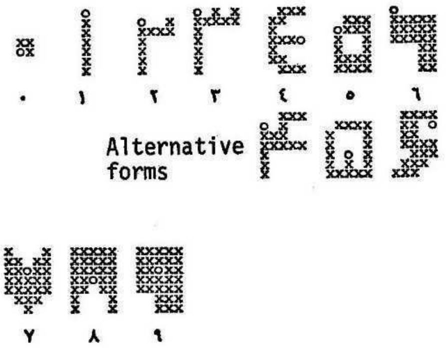


Fig. 7. Optically weighted decimal Farsi and Arabic numerals in a 7 by 5 matrix. The area covered by the numeral x is ax+b.