ON THE USE OF FARSI AND ARABIC LANGUAGES IN COMPUTER-BASED INFORMATION SYSTEMS*

BEHROOZ PARHAMI †

Arya-Mehr University of Technology
Tehran, Iran

Abstract: 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.

In this paper, some basic problems concerning the use of Farsi and Arabic languages in computer-based information systems are identified and guidlines for their solutions are presented. The most immediate problems are in the area of information representation and standardized information interchange codes. Thus we present a short analysis of the important factors and tradeoffs in Farsi code design. We next turn to the input problem for Farsi information and discuss both conventional keyboard data entry and automatic recognition of Farsi texts needed for document data entry. Alternatives for keyboard layout standardization are presented. In the area of information processing, the problems with sorting, text editing and data compression are touched upon. This section is concluded with a discussion of the appropriateness of Farsi programming languages. Finally, in the area of output devices, we review the problems with available high-speed printers and propose a method for dealing with the representation of vowels and supplementary symbols in conventional devices. This discussion is followed by a brief summary of work on line-segment Farsi displays and optically weighted representation of Farsi numerals.

Index Terms: Arabic, Computer Applications, Display Devices, Document Data Entry, Farsi, Information Codes, Input/Output, Keyboards, Language Processing, Programming Languages, Reduced Code Set, Sorting, Standards, Text Editing.

* Presented to the Symposium on Linguistic Implications of Computer-Based Information Systems, New Delhi, India, Nov. 10-12, 1978.

† Assistant Professor of Computer Science and Vice Chairman, Dept. of Mathematics and Computer Science, AMUT, P.O. Box 3406, Tehran, Iran

1. BACKGROUND

The widespread use of computers in varied fields of human activity makes it mandatory for computer-based information systems to be capable of dealing effectively with information presented to them in the natural language of the user community and to present the resulting output in a manner which is comprehensible to users with minimal computer knowledge and training. 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 question-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. In this undertaking, we have been aided by a number of manufacturers who, due to market requirements in Iran and also in some of the Arab countries, have developed Farsi and Arabic input/output capability for their computer systems. Even though because of time constraints and economic factors the proposed solutions are seldom satisfactory [1] , they still constitute important first steps on which more fundamental solutions can be based.

In this paper, some basic problems concerning the use of Farsi and Arabic languages in computer-based information systems are identified and guidelines for their solutions are presented. Our subsequent discussion will be in terms of the Farsi language, with the understanding that almost all of these considerations apply to Arabic (and possibly to Urdu) as well.

The requirement for handling of Farsi information can cause difficulties in each of the four basic phases of data processing: input, encoding (for storage and transmission), processing, and output. Each of these phases will be discussed separately. The encoding phase is presented first since some of its concepts are needed in the subsequent discussions.

2. THE ENCODING PHASE

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, if they are to be mutually compatible.

The difficulties and waste of system and manpower resources, caused by the current incompatibility of computer systems in dealing with Farsi information, prompted the formation of the "Study Committee for Standard Farsi Information Code" in February of this year [2] .

The first step in code design is to select a set of symbols with general utility. As each Farsi alphabetic symbol can appear in several forms depending on context, an immediate question is whether to include in the standard set the full alphabet (each letter having up to four different shapes, for a total of about 120 symbols), the typewriter-style alphabet (most symbols having only two different shapes, for a total of about 65 symbols), or a reduced alphabet (each symbol having a single code, for a total of about 35 symbols, including the pseudo-space character which forces the separation of adjacent non-blank characters and the pseudo-connection character which causes the connected form of a letter to appear next to a blank); the approximate values given reflect the fact that it is not yet known exactly which symbols constitute the useful Farsi alphabet for computer applications.

The advantages of a reduced code set [3] are well-knwon:

1. Smaller code size or, alternatively, greater flexibility for the use of special or user-defined symbols, given a fixed code size.
2. Total separation of the internal storage method from the actual symbol set used for producing output script of desired quality (i.e., technology independence).
3. Considerable increase in data entry speed due to the smaller symbol set size (less frequent use of the shift key).
4. Ease of processing (in particular editing) due to the fact that internal symbol representations are context-independent.

With the advent of "intelligent" peripherals, it becomes less and less meaningful to waste a considerable portion of the code space for storing minor variations of symbols which are easily deducible through a simple algorithm. The disadvantages, which in the author's opinion are relatively insignificant, are as follows:

1. Waste of storage space and data transmission time due to the use of the pseudo-space character.
2. Need for a somewhat higher degree of intelligence in all input/output devices and/or controllers.
3. Unfamiliarity of the concept and, hence, need for re-educating data entry operators and other personnel.

Taking the reduced Farsi alphabet into consideration, the following set of symbols may be considered for standardization:

GROUP	DESCRIPTION	SYMBOLS	NUMBER
1	Letters of the alphabet	ـ آ ا ء ب پ • • • • ی	35
2	Space, pseudo-space, and pseudo-connection	SP PS PC	3
3	Vowels (not normally written)	َ ُ ِ	3
4	Supplementary symbols	ّ ً ٔ ..	4
5	Digits and decimal "point"	٠ ١ ٢ ٣ ٤ • • • • ٩ ٫	11
6	Commercial symbols	ℛ (ریال) ∓ (تومان) ق	3
7	Common punctuation marks and special symbols	. : ! ' " = ≠ # _ % / | \ + - * < > ( ) [ ] { }	≈ 24
8	Farsi punctuation marks	؟ ، ؛ « »	5
		TOTAL	≈ 88

Of these, 63 are Farsi symbols which can be easily incorporated into the extension of an existing standard code set.

To summarize, two factors contributed to difficulties in the Committee's task, as opposed to the relatively straightforward assignment of codes to symbols:

1. Tradeoff between potential advantages of a reduced code set and

the need for "intelligence" in peripheral devices to convert back and forth between the reduced code set and the conventional written representation of Farsi information.

2. Determination of what actually constitutes a reasonable symbol set for representing Farsi information in "typical" applications, with the attendant tradeoff between usefulness of various symbols and code expandability in future.

Even though a reduced code set was finally agreed upon as the main proposal, a more conventional set was also proposed for temporary usage in anticipation of technical difficulites in the immediate adoption of this code by some manufacturers [2] . Both of the proposed codes are standard extensions of ASCII [4] .

3. THE INPUT PHASE

The existing problems with Farsi input and output and some of the approaches for dealing with them have been enumerated elsewhere [1] . 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 societies. 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 caused by one or more of the following properties of Farsi:

1. Possible connectivity of adjacent letters.
2. Varying widths ~~of~~ of symbols (see Figure 1).
3. Minor differences in symbol shapes (e.g., number of dots).

Despite these problems, the results of an initial study in the recognition of printed Farsi texts [5] are quite encouraging and clearly point to the possibility of practical systems in the near future.

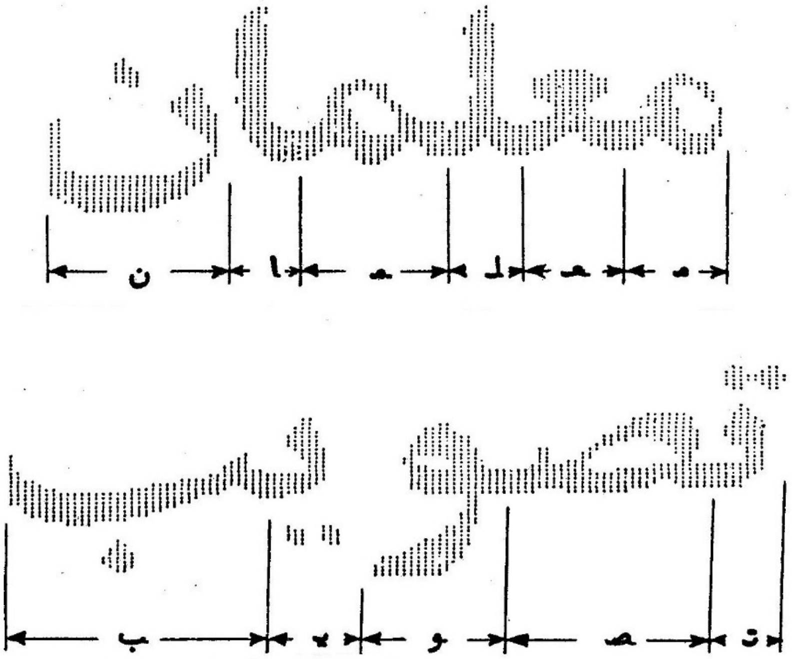

Figure 1. Example of Digitized Printed Farsi Words Showing the Connectivity and Variable Width of Symbols.

Standardization activities are important for character sets suitable for various text input devices, although they do 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 application, 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 more of the following considerations:

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

2. Adoption and widespread acceptance of a reduced Farsi code set.
3. Large scale replacement of Farsi typewriters by word-processing machines.

Presently, the problem of standardized keyboard layout(s) for computer applications is receiving a great deal of attention from the "Study Committee for Standard Farsi Information Code" as a side activity.

Experience with the use of reduced keyboards (directly corresponding to the reduced code set) has shown an important disadvantage; namely, that the keyboard operator has to think in order to insert the pseudo-space character in appropriate places. On the other hand, the existence of compound words and grammatical rules dictating the separation of certain prefixes and suffixes makes the inclusion of a PS (pseudo-space) character mandatory. The following are only a few of the examples:

دانش آموز　دیوانه وار　حیله گر　بی ادب　خانه اش　رفته ایم
هم منزل　کم درآمد　هم اکنون　کم کم　آتش بس　چوب پنبه
نخست وزیر　گه گاه　لاك پشت　پس انداز　خط کش　غلط انداز

Three approaches are possible for dealing with this problem. On one extreme, one may select the reduced keyboard and hope to overcome the PS-problem by proper training. On the other extreme, one may use a conventional typewriter layout for the keyboard (of course, data transmission may still be in the reduced alphabet). An intermediate approach is to use a special keyboard where the shift key is used to disconnect a letter from the following one (actually when the shift key is pressed, a PS is inserted automatically if needed). This approach has the advantage of not requiring any thinking and the disadvantage of forcing a considerable amount of redundant shifting. It remains to be verified if the advantage of mechanical operation outweighs the disadvantage of more work.

The symbols in Groups 3 and 4 are entered immediately following the letter to which they apply. Thus as an example, one enters:

ه ُ ت َ ک َ ل ّ ِ ه - —

Special provisions are needed for entering alphanumeric Farsi or mixed Farsi/ Latin information, due to the opposite direction of text scanning. Again with some local intelligence, this is relatively easy to handle.

4. THE PROCESSING PHASE

We first consider two main problems of immediate interest briefly; namely sorting and text processing. For the purpose of sorting, the alphabet symbols may be divided into three disjoint subsets: The significant symbols (SP plus Group 1, except for "ـ"), the non-significant symbols (PS, PC and "ـ"), and optionally significant symbols (Groups 3 and 4). Such an optional treatment of symbols in Groups 3 and 4 will enable sorting of information both with and without taking vowels and supplementary symbols into account. Another contributor to the difficulties in sorting (and word matching; e.g., directory search) in Farsi is the possibility of writing certain composite Farsi words in several forms; with separate or connected subwords.

To illustrate some of the problems in sorting [6] , we consider the symbols "ی" and "ﻴ" in the Farsi alphabet. If the sorting sequence of symbols is such that "ی" precedes "ﻴ", then the name " قائمیان " will follow "قائمی‌زاده" in a sorted list. On the other hand, if "ی" follows "ﻴ", then the name "شاهین " will appear before " شاهی " . Clearly, both of the above alternatives violate our intuitive notion of sorting in Farsi. Therefore, we either have to adjust ourselves to a new notion of sorting or resort to more sophisticated algorithms for this purpose.

One may think that the use of a reduced code set can alleviate this problems. However, the following example shows that the problem will still exist. Consider these names and their reduced form representations:

علیائی : ع ل ی ا ء ی
علی‌زاده : ع ل ی PS ز ا د ه
علیزاده : ع ل ی ز ا د ه
علیشاهی : ع ل ی ش ا ه ی

Now, whether in the sorting sequence the character "PS" precedes or succeeds all alphabetic symbols, the two names " علی زاده " and "علیزاده" will be separated by one of the other names in the sorted list; this clearly contradicts our common notion of sorting.

Problems in the automatic processing of Farsi texts (e.g., page formating and left-end justification) are caused by the following:

1. Long words cannot be broken at the end of the lines by hyphenation.
2. Punctuation rules are not followed consistently in Farsi (educational problem).
3. Blanks are not consistently used to separate words, 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.

To elaborate on the third point, we present the following exagerated example which is impossible to decompose into words without some kind of semantic processing:

پرویزد رکنارد ارا وآذ رایستاد وازآنها چند د فترسبزگرفت.

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 editing applications, as long as the boundaries that are actually recognized are not too far apart. This is usually true in practice [7] . On the other hand, the possibility of extending some Farsi letters is at times helpful for text justification; e.g., "حمـــید" or "حمیـــد".

Another aspect of Farsi text processing has to do with efficient storage methods for natural-language data bases. This is currently under investigation [8] with particular attention to the implications of the reduced code set. The extra code words resulting from the use of a reduced code set enable the encoding of a larger number of text fragments as single characters, which should more than offset the redundant usage of PS and PC in most applications.

The final aspect of processing discussed here is that of algorithm specification [6] . 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 may be 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. Some of the reasons for this pessimism are:

1. Programmers should know some English anyway. A great deal of computing literature is produced in English and there is no hope that we can become self-sufficient in this respect in the near future.
2. The amount of English knowledge required for efficient programming is relatively low, even in the case of languages such as COBOL. What is difficult in teaching programming is certainly not the fact that "GO TO" means "برو بـــه"!
3. If, for example, through the use of a preprocessor we simply replace each COBOL word by its Farsi equivalent in our programs, the resulting structures will be so foreign to Farsi as to make their learning equally, if not more, difficult.
4. Any new language proposed will need to be initially developed and documented for various computer systems and subsequently maintained and updated periodically. It is felt that our limited manpower resources are better spent on developing more effective computer applications than on designing new programming languages and systems.
5. We cannot and should not duplicate the thousands of person-years of effort that have gone into developing the popular programming languages and related systems and material (compilers, interpreters, software packages, textbooks, etc.) for a variety of purposes, on almost every type of hardware.

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 fuction keys on a private, pre-programmed keyboard) or define a system by filling tables and answering multiple-choice questions. 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!

5. THE OUTPUT PHASE

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 existing problems in the other two areas, which are of more immediate interest, are satisfactorily solved.

Most high-speed Farsi line printers are knwon to be of poor quality. this is due to the following :

1. The technology of most high-speed printers is unsuitable for printing the connected symbols of the Farsi script (they have inter-symbol gaps).
2. The varying widths of Farsi alphabetic symbols are either ignored for simplification or simulated by decomposing wider characters into two parts, whenever possible, thus compounding the problems of connectivity and horizontal alignment.
3. The large symbol set needed for printing Farsi has sometimes caused the elimination of needed special symbols and/or some of the letter forms (e.g., " ـهـ" and " ع") due to limitations

in symbol set size. The need for mixed Farsi/Latin print in some applications has compounded this problem.

4. The similarity of various Farsi letters (e.g., "ـبـ" and "ـثـ") necessitates a higher quality printout for readability. The smearing effect of most high-speed impact printers can, therefore, cause problems in this respect.

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 (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 can accept a standard code set, it can use as many different forms for each letter as deemed necessary for a reasonable output script quality. Symbols in Groups 3 and 4 can either be overprinted on the preceding letter or form an independent symbol, with connected and separate forms as shown in the following examples:

آنـَـا ً   مـُـتـَـكـَـلـِـ ّـم

ا ُ ر د َ ك   مـِـشـكـوه ٌ

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.

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 [9] . The irregular shapes of Farsi alphabetic symbols requires the use of at least 18 line segments per symbol in order to obtain an acceptable output script (Figure 2a). For numerals, a seven-segment display can be designed, but the segments have highly irregular shapes (Figure 2b).

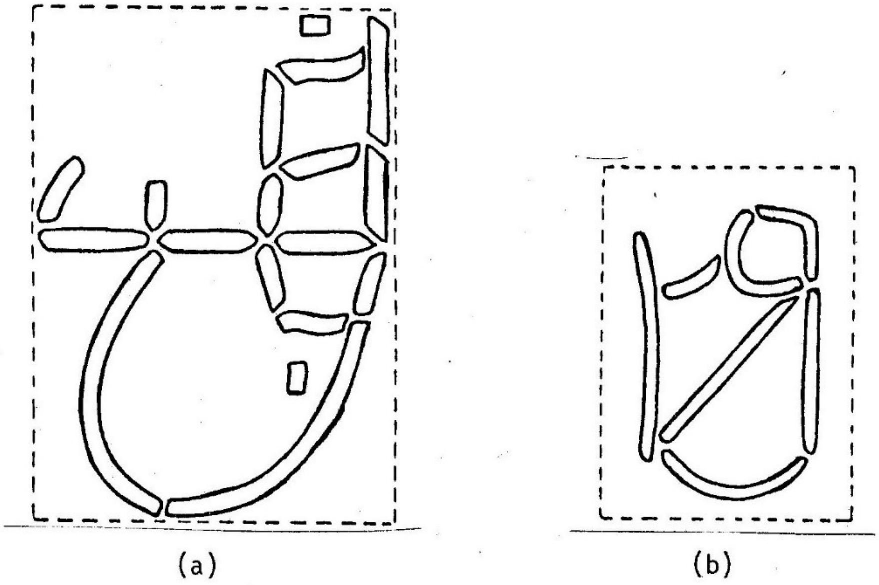

Figure 2. Line-Segment Displays for Alphanumeric and Numeric Farsi Information.

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 [10] . Figure 3 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 [10] .

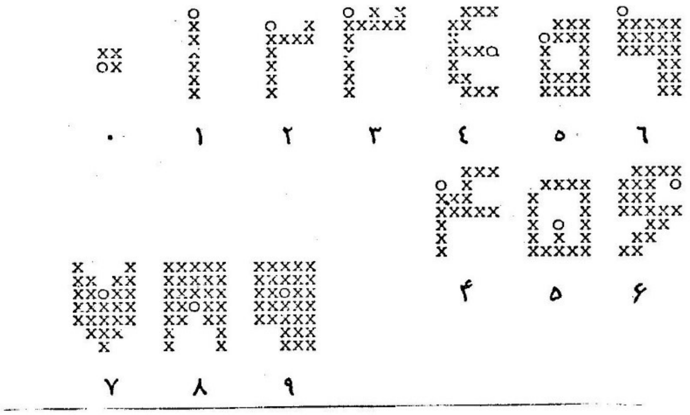

Figure 3. Optically Weighted Decimal Farsi and Arabic Numerals in a Seven by Five Matrix.

6. 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. Fortunately, these problems are receiving a good deal of attention from governmental organizations, manufacturers, and academic institutions. Despite the progress made so far [11] , there appears to be much room for systematic research in this area.

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.

7. REFERENCES

[1] Parhami, B. and F. Mavaddat, "Computers and the Farsi Language: A Survey of Problem Areas," Information Processing 77 (Proc. of IFIP Congress, Toronto, August 1977), North-Holland, Amsterdam, 1977,pp. 673-676.

[2] "Preliminary Proposal for the Iranian Standard Information Code," Operations Research Bureau, Informatics Division, Plan and Budget Organization of Iran, Tehran, May 1978.

[3] Hyder, S.S., "A System for Generating Urdu/Farsi/Arabic Script," *Information Processing 71* (Proc. of IFIP Congress), North-Holland, Amsterdam, 1972, pp. 1144-1149.

[4] "American National Standard Code Extension Techniques for Use with the Seven-Bit Coded Character Set of American National Standard Code for Information Interchange," American National Standards Institute, Document ANSI X3.41, 1974.

[5] Taraghi, M., "Automatic Recognition of Printed Farsi Texts," M.S. Thesis in Computer Science, Arya-Mehr Univ. of Technology, Tehran, July 1978 (in Farsi).

[6] Parhami, B., "Impact of the Farsi Language on Computing in Iran," *Middle East Computer* , to appear.

[7] Nasseri, P., "Automatic Processing of Farsi Texts," M.S. Thesis in Computer Science, Arya-Mehr Univ. of Technology, Tehran, Jan. 1978 (in Farsi).

[8] Tafaghodi Jaami, A., "Techniques for Increasing the Efficiency in the Storage and Retrieval of Farsi Texts," M.S. Thesis in Computer Science, Arya-Mehr Univ. of Technology, Tehran, in progress.

[9] Parhami, B., "Low-Cost Output Displays for Microcomputer Applications," *Proc. of the Second India Symp. on Computer Architecture and System Design*, New Delhi, Nov. 1976, pp. 111-119.

[10] Parhami, B., "Optically Weighted Dot Matrix Farsi and Arabic Numerals, *Proc. of the Third Jerusalem Conf. on Information Technology*, Jerusalem, Israel, Aug. 1978, North-Holland, Amsterdam, 1978, pp. 207-210.

[11] Ashjaee, M.J., F. Mavaddat and B. Parhami, "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.