INFORMATION PROCESSING 77, B. GILCHRIST, EDITOR

COMPUTERS AND THE FARSI LANGUAGE—A SURVEY OF PROBLEM AREAS

BEHROOZ PARHAMI and FARHAD MAVADDAT
Arya-Mehr University of Technology
Tehran, Iran

As users of computer systems produced by the West, we face unique problems brought about by the interplay between the Farsi language and the field of computing. It is felt that a satisfactory solution of these problems is essential if computers are to play an important role in the development of our country. In this paper, we review these problems by dividing them into five categories; namely, education, codes, input, output, and programming. For each category, we present an introduction to the problems in that particular area and discuss past approaches and promising techniques.

1. INTRODUCTION

The application of computers in Iran started in 1962 with the installation of an IBM 1620. [1] Since then, the number of computers has increased roughly by a factor of ten every five years. The introduction of computers in Iran brought about many problems, the most important being a serious shortage of specialists to run these systems, a lack of knowledge on the part of executives and public officials as to their potentials and limitations, and finally, the interplay between Farsi language and the field of computing. Whereas the first two problems are shared by most developing countries, the third one is more unique.

The purpose of this paper is to survey problems in the field of computing which are related to the Farsi language. Even though computer and peripheral manufacturers have attempted to alleviate some of these problems, we do not consider the proposed solutions satisfactory. This is mainly due to time constraints in such development efforts and the unavoidable goal of cost minimization.

Before proceeding further, a brief introduction to the Farsi alphabet appears to be necessary. Farsi symbols are juxtaposed from right to left to form a Farsi text. These symbols normally have varying widths and heights in printed and handwritten texts and can be connected to adjacent symbols in the same word through the center line. In typewritten texts, however, usually only two different widths (of one unit and two units) and occasionally three different widths are utilized. In all Farsi computer printers constructed to date, constant-width Farsi symbols are utilized for implementation reasons, with some of the wider ones formed by decomposition into two adjacent symbols.

Table 1 shows a constant-width representation of the 32 Farsi letters and the 10 decimal numerals. As shown in table 1, each letter has up to four variations which are used depending on whether it is to be connected to the preceding symbol on its right (A_r=1) or to the succeeding symbol on its left (A_ℓ=1). It is possible to use the delayed decision algorithm of Hyder [2] for automatically converting a text stored using only 32 letter codes (plus blank) into a readable printed text. The 114 alphabetic symbols of table 1 may be reduced to about 60 by combining those with only minor differences, as normally done in Farsi typewriters and computer terminals. (Some manufacturers, forced by particular hardware constraints, have even smaller character sets).

It is interesting to note that Arabic alphabetic symbols and numerals are a subset of those in table 1. Many of the problems discussed in this paper are, therefore, directly applicable to Arabic as well. Clearly, joint ventures with interested organizations in Arab countries, in particular with regard to standardization efforts discussed in section 3, are desirable.

TABLE 1
Constant-width representation of Farsi alphanumeric symbols

Letter No.	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
English Equivalents	SH	S	ZH	Z	R	Z	D	KH	H	CH	J	S	T	P	B	AA
A_ℓ A_r 0 0	ش	س	ژ	ز	ر	ذ	د	خ	ح	چ	ج	ث	ت	پ	ب	ا
0 1	ش	س	ژ	ز	ر	ذ	د	خ	ح	چ	ج	ث	ت	پ	ب	ا
1 0	ـش	ـس	ـژ	ـز	ـر	ـذ	ـد	ـخ	ـح	ـچ	ـج	ـث	ـت	ـپ	ـب	ـ
1 1	ـشـ	ـسـ	ـژـ	ـزـ	ـرـ	ـذـ	ـدـ	ـخـ	ـحـ	ـچـ	ـجـ	ـثـ	ـتـ	ـپـ	ـبـ	ـ

Letter No.	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17
English Equivalents	Y	H	V	N	M	L	G	K	Q	F	CH	'	Z	T	Z	S
A_ℓ A_r 0 0	ی	ه	و	ن	م	ل	گ	ک	ق	ف	غ	ع	ظ	ط	ض	ص
0 1	ی	ه	و	ن	م	ل	گ	ک	ق	ف	غ	ع	ظ	ط	ض	ص
1 0	ـی	ـه	ـو	ـن	ـم	ـل	ـگ	ـک	ـق	ـف	ـغ	ـع	ـظ	ـط	ـض	ـص
1 1	ـیـ	ـهـ	ـوـ	ـنـ	ـمـ	ـلـ	ـگـ	ـکـ	ـقـ	ـفـ	ـغـ	ـعـ	ـظـ	ـطـ	ـضـ	ـصـ

Numerals	0	1	2	3	4	5	6	7	8	9
Farsi Forms	۰	۱	۲	۳	۴ ٤	۵ ٥	۶ ٦	۷	۸	۹

2. COMPUTER SCIENCE EDUCATION

Computer education in Iran, particularly at the more elementary levels, is hindered by a lack of adequate educational material and the needed terminology in Farsi. The bulk of computer literature is published in English. What we are experiencing in terms of "data processing jargon" is shared by many countries around the world. Every day, new words are created and old terms die out. The pace is so rapid that it is extremely difficult to invent words in Farsi for describing the new concepts. This is, of course, part of the more general problem caused by the fact that we are consumers of Western technology.

The weakness of Farsi in technical fields has caused the so called "invasion" problem whereby many foreign words have found their way into the language. There is much discussion these days as to whether we should

accept these words as part of Farsi or attempt to find a cure. [3] On the two extremes of this argument, are those supporting the absolute purity of the Farsi language and those who advocate abandoning Farsi altogether in favor of a more technically oriented language.

Neither of the above views is practical and the final solution will probably be based upon a compromise; e.g., forming new words from foreign roots using Farsi grammatical constructs. In the computer field, in particular, the question is not whether to accept foreign words but how much of them to use. The answer, we believe, lies in finding equivalents only for the more frequently used basic words.

In the area of educational material, the current problems will be with us for some time. The number of computer texts in Farsi is small and only a handful of them are of reasonable quality. Our resources are clearly too limited for writing books on all aspects of computer science. Translation of foreign sources is also out of question because of the dynamic nature of our field. Any translated computer book will probably be out of date by the time it is published. Besides, the question of which books to translate is a difficult one to answer. Only a few computer science texts are generally accepted as the "best" in their respective fields.

The solution is again obvious. Our limited resources should be directed toward preparing educational material for elementary computer science education. Our advanced students must continue using English sources and should, therefore, be required to have a good working knowledge of the language.

An important starting point for improving the quality of computer science education is teaching the general public about the potentials and limitations of computers. This effort, if linked with some degree of control over the activities of private institutes engaged in teaching computer skills, in order to assure reasonable quality of education and to prevent unrealistic and misleading advertisement of their services, will undoubtedly result in a better environment for educational programs to proceed.

3. ON CODES AND SYMBOLS

As mentioned previously, a minimum set of about 60 different symbols are needed in order to produce an easily readable Farsi script. The symbols actually selected for this purpose and their internal representations are different from one system to the other. This lack of compatibility is clearly undesirable. The need for, and advantages of, standardization cannot be overemphasized. However, premature standardization must also be avoided at all cost.

Ironically, even a standard set of symbols for written Farsi has not been defined. There are differences of opinion as to the existence of variations of some letters (e.g., ﺗ and ﺘ). Although such questions must be settled by the Language Academy, extensive effort and lobbying of computer people for an early decision is essential. Only we know how much effort will be wasted if hardware components are developed without provisions for needed symbols.

For the near future, peripheral devices dealing with Farsi information will probably be based on technologies developed for Western countries. Therefore, subsets of Farsi symbols need to be defined in order to enable Farsi communication through devices with inherent technological limitations. It is reasonable to have three sets of such symbols (minimum, adequate, and desirable) defined for computer applications. Once such sets of symbols are defined, implementation details should be left to the ingenuity of designers.

The selection of codes for Farsi symbols is an even worse case of negligence. Computer manufacturers have adopted different codes, with the result that programs with Farsi input and output are highly machine-dependent. This has caused serious problems and can easily lead to one manufacturer's dominance due to incompatibility of different systems.

In the design of standard codes for Farsi symbols, a number of decisions have to be made. The first and probably the most important question is whether to use distinct codes to represent variations of letters. If distinct codes are used, care must be taken to make the variations of each letter functionally transformable. However, it appears that considerable storage efficiency can be gained by having a single internal code for each letter. Then, a "dead space" character can be used to force the separation of adjacent letters in the output, while it is itself ignored as a non-printing symbol.

Several other points must be taken into account in the design of a standard code for information interchange in Farsi: (1) The need for mixed Farsi and Latin symbols in some applications. (2) Lexicographic ordering of letters and numerals, in relation to Latin alphanumeric symbols. (3) The possibility of storing vowels, even though most of them are deleted in written Farsi. (4) Specification of the full standard with a sufficient number of symbols for an aesthetically pleasing script. (5) Recommendation of subsets of the standard code for applications where the full standard is not needed or cannot be implemented. (6) Simplification of the existing problems in sorting of Farsi information, as discussed in section 6.

4. FARSI INPUT PROBLEMS

Conventional input media (e.g., punched cards, punched paper tape, terminal keyboards) do not present any problem unique to Farsi, since with such media, symbol codes are read into the computer rather than the symbols themselves. However, since the initial data entry for all such media is (and will be for the foreseeable future) through keyboards, it is important to reach a decision regarding standardized Farsi keyboards.

The current practice is to use keyboards in which the key positions follow closely those of typewriters. Since typewriter keyboards were originally developed for Arabic, the key positions are not necessarily optimal for Farsi. In addition, the large number of Farsi symbols necessitates frequent use of the SHIFT operation, which in turn contributes to reduced data entry speed.

As an alternative to the use of the existing keyboard layout, one may standardize a "reduced" Farsi keyboard, made possible by recent developments [2], [4]. Considerable improvement in operating speed and reduction in keyboard cost can be expected from these techniques. The main objection to such an approach would be the large number of available personnel trained in using conventional Farsi keyboards. Nevertheless, the advantages of reduced keyboards may justify the cost of a national re-training program. In any case, the decision has to be made once and for all and it should be made before further damage is done.

Another problem arises in dealing with mixed alphabetic and numeric Farsi information, as numbers are written from left to right. The same problem is present if mixed Farsi and Latin symbols are to be dealt with. The current practice on Farsi typewriters is to enter numeric information in reverse order, from right to left. This approach is clearly even more inconvenient when dealing with Latin alphabetic symbols in Farsi texts. In the case of key-

boards with no hard-copy display and keypunch equipment which buffer the information before actual punching, additional hardware may be used for recognizing Farsi numeric and Latin alphanumeric symbols and placing them in proper order within Farsi text.

An existing problem with card punch equipment is that they punch and print from left to right. As the printing is not even done in Farsi, it is extremely difficult to read the information punched on cards; thus increasing the possibility of input errors. Minor modification in existing equipment can alleviate this problem.

The more challenging problems in the area of Farsi input are related to less conventional methods, such as optical and magnetic character recognition. The difficulties are caused by one or more of the following properties of Farsi, in order of importance: (1) Possible connectivity of adjacent letters in Farsi words. (2) Varying widths and heights of Farsi symbols, even in typewritten text. (3) Existence of many symbols with only minor differences; e.g., in number and place of dots. (4) Radical differences in symbol shapes in existing common typefaces.

5. FARSI OUTPUT PROBLEMS

While the problem of Farsi codes and input, as discussed in sections 3 and 4, are of concern to computer professionals and a limited number of advanced users (and therefore technological constraints can be compensated for by proper training), Farsi output is a universal problem. Over the centuries of Islamic influence, writing of Farsi has become something of an art. In view of this fact, extensive modification of Farsi script to adapt it to machine printing is highly undesirable. The present inadequacy of computers (these supremely capable machines!) in generating an acceptable Farsi printout has alrady caused some damage in the public attitude toward them.

A number of characteristics are desirable for a device if it is to generate Farsi output of reasonable quality: (1) The generated symbols should have varying widths; acceptable results are possible with a minimum of three or four different symbol widths. (2) A large set of symbols need to be generated; especially if mixed Farsi and Latin output is desired. (3) Due to similarity of many symbols (differing only in the number or place of dots), very sharp, high quality output is needed in order to form an easily readable script. (4) The need for connectivity of adjacent symbols reduces the desirability of output devices with inherent inter-symbol gaps; e.g., chain and drum printers.

Conventional drum and chain printers, which are currently the only ones widely available for hard-copy Farsi output, constitute the worst possible choices; they have inherent shortcomings due to constant-width symbols, smeared printout, inter-symbol gaps, and small symbol sets. Of these two, drum printers are more limited by the size of their symbol sets and difficulty in horizontal alignment of symbols which is essential for the connectivity in Farsi script.

Attempts at solving these problems have been of limited success. To overcome the multiple-width difficulty, wider symbols have been decomposed on some printers into two parts. [5] Even though the decomposition technique can be extended to provide three or four different widths, reduced printing efficiency (amount of information per printed line) and increased alignment problems (which are bad enough in existing systems) render the approach impractical.

The vertical drum printer of Dataproducts [6] appears to be the first output device designed with the particular needs of Farsi in mind. It effectively alleviates the previously mentioned problems of inter-symbol gaps and alignment of connected symbols by printing the lines vertically. This is of course done at the cost of increased storage requirement and reduced average printing speed. Although the first drawback may soon become insignificant because of the dramatic reductions in memory cost, the second one is inherent in this approach.

Many of these difficulties can be easily overcome with character printers. However, since such devices are relatively slow, they are only of limited interest. For high-speed output, the new generation on nonimpact printers [7] appear to be promising. By their method of operation (e.g., thermal, electrostatic, or xerographic processes), such devices produce sharp, highly readable output with no inherent inter-symbol gap; some of them are actually used for plotting as well as printing.

Most of the above difficulties are also present in the design of Farsi output displays. Much work is needed for the implementation of output displays (both of the CRT type [8] and dot-matrix or line-segment types [9]) with easily readable Farsi symbols. In the case of dot-matrix displays, Farsi symbols need a relatively large matrix for readability. It is important to determine smallest possible matrix sizes for generating a minimum, adequate, and desirable set of Farsi symbols, as discussed in section 3.

6. ON PROGRAMMING AND FARSI

Whereas it makes little sense to talk about a Farsi assembly language, or even procedural high-level language for that matter, it is clearly desirable to have Farsi equivalents for the so called "English-like," very high-level languages. It is not clear at what level this transition from insensibility to desirability takes place. A Farsi-speaking individual can master FORTRAN rules in a few weeks, even if he knows little or no English. But the same person will have a harder time learning an English-like query language.

Another fruitful area of investigation is in the formal description of the Farsi language itself. This is a difficult task, since even a generally accepted Farsi grammer book does not exist. [3] Nevertheless, with a reasonably limited subset of Farsi, the problem is not impossible to solve. This effort should provide us with valuable insight into the structure of the language which will then not only benefit computer scientists pursuing problems enumerated subsequently, but also provide linguists with a new dimension of the Farsi language.

We now turn to problems concerning the automatic processing of Farsi texts with very little or no attention to their meaning. In addition to problems encountered in the input and output phases (outlined in sections 4 and 5), one main obstacle exists; the automatic recognition of word boundaries in texts. The difficulty arises because of the fact that some adjacent pairs of Farsi symbols appear disconnected and, therefore, blanks are not consistently used to separate adjacent words in printed and typewritten Farsi texts.

Two other factors also contribute to the difficulties. The first one is the possibility of writing many composite Farsi words in several forms; with separate or connected components. This complicates the problem of word matching in many applications and is the most important contributor to the difficulties in sorting of Farsi information; a serious problem which we now face. The second factor relates to the fact that punctuation rules are seldom followed in Farsi texts. The result is Farsi texts in which the words, phrases and sentences are difficult to recognize, even by human readers.

Applications in which the meaning of a Farsi text needs to be taken into account are numerous and quite varied in their degree of difficulty (depending

on the depth of meaning considered). Most such problems have been studied for other languages; notably English. Special effort must be devoted to the identification of simplifying properties of and special problems created by the Farsi language in this respect.

Many benefits can be gained from computer-aided analysis of the rich heritage of Farsi literature. Of course, the help of linguists is needed in such an undertaking. The results can also be expected to clear up many ambiguities in the Farsi language itself and to provide a basis for a clear and concise description of its grammar.

7. CONCLUSION

In this paper, we have surveyed problems in the field of computing which are related to or caused by the Farsi language. The basic problems enumerated here cannot be pushed aside for long if computers are to play an important role in the development of our country. Any long-term plan for our computing needs should, therefore, include attempts at finding satisfactory solutions for these problems. (A longer version of this paper, [10] which is available upon request, includes an appendix in which the problems discussed here have been broken into 47 project topics, many of which can be dealt with at the level of a master's thesis).

As pointed out elsewhere, [11] many of the problems enumerated in this paper are part of the difficulties caused by the rapid growth in the number of computer-based systems in Iran [12] without an adequate plan to build the needed foundation in terms of manpower training and standardization. We feel that, due to our severe shortage of expert manpower, these problems cannot be solved satisfactorily unless immediate action is taken to establish a national framework for informatics development in Iran.

Undoubtedly, as these problems are solved, many new ones will be created in the process or as a result of advances in the state of computing in Iran. Therefore, we should be looking beyond the immediate problems of today in order to prepare ourselves to deal with the more challenging problems of tomorrow. With this goal in mind, our computer science programs must be based not on what we face now but on what we wish to accomplish in the future.

REFERENCES

[1] Statistical center of Iran, The computer in Iran, June 1971, 55 p. (in Farsi).

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

[3] P. Rajabi, The invasion of foreign words in Farsi: Problem statement, Tamaashaa, vol. 6, No. 281, 2 October 1976, p. 14 (in Farsi).

[4] A. Nahapetian and F. Mavaddat, Decomposition of Farsi alphabetic characters for computer input and output, Patent no. 14151, Iranian Patent Office, 26 January 1976.

[5] F. Mavaddat and A. Nahapetian, Decomposition methods applied to the design of character output devices, Proc. of the Eleventh Conf. on Statistics and Computational Science, Cairo Univ., Egypt, April 1975.

[6] Model 2235 Page Printer, Specification sheet, Dataproducts Corp., Woodland Hills, Calif., USA.

[7] Computer, Special Issue on Printer Technology, vol. 8, no. 9, September 1975.

[8] F. Mavaddat, Character decomposition applied to the design of more ecomomical character generators, Computer Systems Lab., Arya-Mehr Univ. of Technology, Tehran, Iran, Technical Report CSL-76-001, January 1976.

[9] B. Parhami, Low-cost output displays for microcomputer applications, Proc. of the Second India Symp. on Computer Architecture and System Design, New Delhi, November 1976.

[10] B. Parhami and F. Mavaddat, Computers and the Farsi language: A survey of problem areas. Computer Systems Lab., Arya-Mehr Univ. of Technology, Tehran, Iran, Technical Report CSL-76-007, December 1976.

[11] F. Mavaddat and B. Parhami, Informatics in Iran: Problems and prospects, Proc. of the International Conf. on Computer Applications in Developing Countries, Bangkok, Thailand, August 1977.

[12] Informatics in the Sixth Plan, Report of the Joint Committee for Planning in Statistics and Informatics, Informatics Division, Plan and Budget Organization, Tehran, Iran, February 1977 (in Farsi).