IFAC

International Federation of Automatic Control

SYSTEMS APPROACH FOR DEVELOPMENT

Edited by
N.A. GHONAIMY
Ain Shams University, Abbasia, Cairo

PERGAMON PRESS

International Federation
of Automatic Control

IFAC Conference on

SYSTEMS APPROACH FOR DEVELOPMENT

Cairo, The A.R.E.
26-29 November 1977

FINAL
PROGRAM

Education and Transfer of Technology 2 (Oxy Room)
Chairman: A.H. Rashwan
Co-Chairman: M. Najim
Reporter: Mohsen Tawfik

Dynamic power system simulator for design and training purposes
S.M. Minkley and M.J. Short

What does a developing country expect from intergovernmental co-operation in information technology?
E.K. Chapman-Mundy

Two models for the development of transnational education systems: redesigned transfer and controlled multiplication
G. Peterlongo

Optimal planning in education to meet social demand and manpower requirements
A. Deif

Remarks about printing and displaying some non-Latin characters
M.C. Vanwormhoudt and M.A. El - Hamalawy

Power Systems 1 (Thoron Room)
Chairman: M. Cuenod
Co-Chairman: T. El Tablawy
Reporter: Moh. Badr

Optimal operation of the Algerian electric power system
H. Ait-Ouyahia, F. Falgarone and P. Lederer

Stability and control of large scale power systems using decomposition-aggregation techniques
M. Darwish and J. Fantin

Nuclear power planning with reference to developing countries
S.M. Fadillah

Controllability of power systems with long transmission lines
A.R. Abu-El-Wafa

Regional electric power system planning using mixed integer linear programming
T. Backlund and J.A. Bubenko

1400-1700 Industrial Application 2 (Oriental Hall)
Chairman: Youssef Ismaiel
Co-Chairman: Ashraf Hamdy
Reporter: Moh. El Nahas

Forecasting cement and steel needs in developing countries
P. Coldan and M. Cuenod

On-line moisture measurement in phosphate process control
M. Najim, F. Najim, M. Ayoute and T. Ouazzani

Process-computers in cement plants
J. Gravgaard

Optimal multilevel control of hot rolling steel mills
G.M. Aly, M.M. Aziz and M.A.R. Ghonaimy

Some economic and technical aspects of mini-mills
H. Stiekler

REMARKS ABOUT PRINTING AND DISPLAYING SOME NON-LATIN CHARACTERS

M.C. Vanwormhoudt* and Mohamed A. El Hamalaway**

**Laboratory of Electronics, University of Ghent, St. Pietersnieustraat 41, 9000 Ghent, Belgium*
***National Semiconductor GmbH, Industriestrasse 10, D8080 Fürstenfeldbruck, West Germany*

Abstract. Most typewriters, electronic printers and display systems were designed initially, specifically for Latin characters. This has delayed the introduction of conversational and business terminals in regions where the written language calls for using characters of an essentially different nature. A typical case of this situation occurs for the Arabic group of languages upon which the paper will place special emphasis.

It is possible to adapt typewriters and printers into devices, useful for the given purpose, but such adaptations are always "ad hoc", that is for a particular given language, and often involve using trade-offs that are sometimes inconvenient or aesthetically unacceptable. For most electronic display systems such an adaptation even proves to be impossible in practice. The problems that arise are due to the fundamental structural differences between the typing of Latin and non-Lation texts. These differences relate mainly to (a) the necessity of joining some or all of the letters in a word, (b) the impossibility of incorporating all characters in a module of a fixed length, (c) the variations a character undergoes, depending on its place in a word and (d) the need for composing, in some cases, the characters out of different elementary signs.

Most of these difficulties can and have been overcome to some degree by generating the characters by a dot-matrix. The matrix of 5 x 7 dots commonly used for Latin script is insufficient for the representation of Arabic characters. An example is presented where using matrices of 7 x 7 and 14 x 7 dots leads to quite satisfactory representations. Patterns of 14 x 7 dots can of course be treated as a double impression with a 7 x 7 matrix. Other solutions have been used, that call for the typist to compose some characters out of several, partial impressions.

A simple, elegant and convenient solution is proposed for printing as well as for displaying. It is achieved by using a single column of dots as the basic representational unit. Since such a configuration does not imply any modularity or periodicity foreign to the language to be represented, it can be easily sequenced so as to achieve any required sequence of printed characters, as long as the vertical resolution of 7 dots is sufficient. Additional resolution can of course be provided.

It is concluded that on the basis of the proposed procedure printers and display units can be developed, that would be controlled by a microprocessor and could be used with identical hardware for representing many different scripts or for different styles of typing belonging to the same language. The specificity of the display unit or the printer with respect to its character font would reside in a set of data (character

file) and a driver program stored in a ROM-memory. The character file could be identical for a display unit as for the printer.

Additional advantages of the proposed procedure are: (a) The presence of a microprocessor in the system and the"intelligence" it provides can be used to alleviate some of the tasks of the typist and to achieve a better performance as a terminal. (b) The printed text would be well adapted for treatment with automatic text-readers.

Keywords. Printers; display systems; text editing; digital systems; intelligent data terminals.

INTRODUCTION

Because many written languages have a structure that is rather different from those employing the Latin character set, some problems arise in typing and displaying non-Latin texts. Some scripts, like the Chinese call for composing words out of different signs, while others like the Arabic require joining some letters to each other in a word and are written out from right to left.

Adapting typewriters, printers and display systems that were originally designed for Latin text, to handling non-Latin text usually involves making allawances and trade-offs. These are mostly inconvenient and sometimes aesthetically unacceptable. The Arabic language needs e.g. eighty two different symbols for letters, signs and numerals in order to obtain an acceptable representation. Reducing that number to sixty-four, as used for the Latin character set, therefore seems most inconvenient.

Some maufacturers which have introduced machines to be used for the Arabic group of languages have reduced the number of symbols by simply omitting some letter forms. Others try to circumvent the problem by composing some characters out of different common parts so that most of the characters are incorporated into the available sixty-four codes.

The above mentioned problems, as well as some other questions, are dealt with in the present paper. The approach taken is not to forse the written text into a font that can be handled using the limited possibilities of a present-day modified system for printing and displaying Latin text. Instead, it will be shown that by incorporating a modest amount of "intelligence" one can obtain a printing or display system that can handle a given language or even a set of different languages quite satisfactorily. The feasibility of this proposal is guaranteed by some new advances in printing and display technology and by making use of the possibilities of microcomputers.

It is believed that doing so can result in printing and display systems that would not only be technically superior but would offer economical advantages as well, especially because of their easy adaptability to several languages.

FUNDAMENTAL STRUCTURAL DIFFERENCES BETWEEN TYPING LATIN AND NON-LATIN CHARACTERS

A study of Latin and non-Latin printed text reveals a number of fundamental differences. The most obvious difference is of course related to the direction of printing. This difference calls, however, for only a minor modification of the available machines. It should also be noted that some languages are quite flexible in this respect. Like the Chinese language, they can be written in different directions. Other languages, such as the Arabic language, can only be written from right to left.

Other, more fundamental differences are summarised in the following:

The Need to Join Some Letters In a Word

While the Latin group of languages are readable in both joined and unjoined forms, this is not the case for other languages, such as in Arabic, where the unjoined form is unacceptable. For matrix printers and displays, this difficulty can be overcome by modifications in the electronic circuitry involved. For other types of printers, such as chain printers, this problem requires extensive modification and redesign of mechanical parts. An example of an Arabic text in which the spaces between different letters are reduced

wherever necessary,is shown in fig.1.

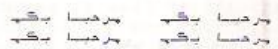
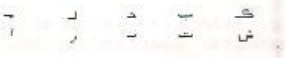


Fig. 1. Dot Matrix Typed Arabic Text

The Impossibility of Incorporating All Characters in a Module of Fixed Length

Although high-quality Latin text also uses letters with different lengths, the problem is usually ignored in typewriters and this results none the less in quite acceptable written text. There even exist some typewriters for Latin characters that use a variable space for the different characters. In Arabic, using different lengths for some characters is almost unavoidable (see fig. 2).



(a) Arabic Letters

i l e w m
W Q R X M

(b) Latin Letters

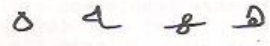
Fig. 2. Some Arabic and Latin Letters

The Variations a Character Undergoes Depending on its Place in a Word

To a lesser degree, this problem also exists in normal Latin text, where sentences are required to start with a capital letter or where, such as in German, every substantive is written starting with a capital.

In the Arabic language, the problem is more complex. Some characters can take up to four variations while others occur always in the same form (see fig. 3). Such morphological changes do not imply any variation in the information contents of the character, but depend only on the position of the given letter in the word. The form to be used first depends on whether it lies at the begining, in the middle or at the end of a word. It also depends on whether it can or cannot be joined to the previos letter. It follows that the choice of the required form can be made, once the characters preceding and following the given letter are known.

(a) A Character Having Only One Shape



(b) A Character which Undergoes Four Different Shapes

Fig. 3. Two Arabic Characters

The Need for Composing in Some Cases, a Character out of Several Elementary Symbols

This problem is most important for languages such as Chinese or Japanese. To a lesser extent, it also exists for the Arabic group of languages when printing or displaying a text, using phonetic signs. Using such phonetic signs and indications are a necessity in some cases. The process of composing with impact printers is cumbersome, as it calls for multiple impacts separated by bachspaces. With dot matrix printers the problem is much easier to solve, especially when some form of intelligence is available. The composition must then not be done on paper, but can be accomplished in the random access memory of the system,a few microseconds before anything is printed or displayed.

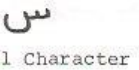
THE DOT MATRIX SOLUTION TO THE REPRESENTATION OF DIFFERENT CHARACTERS

Most of the difficulties discussed above have been overcome to some degree by generating the characters using a dot matrix. The 5 x 7 dot matrix is commonly used for Latin group of languages as it does not result in an aesthetically alternative alphabet. For the Arabic language it was found to be totally unacceptable. A character set was designed using two matrix sizes, namely of 7 x 7 and 14 x 7 dots (see fig.2). This character set was based on the Kufie script, used extensively in

engineering texts, as shown in fig.1. The same Kufie script is also useful for other languages belonging to the Arabic group, such as Farsie, Urdu, Jawi and Swahili. Although using these matrices was found to be convenient, much more flexibility and freedom could be obtained if one would be allowed to use matrices of arbitrary length for different characters. The fundamental printing or display element would then consist of a single column of dots. The resulting text would then imply no periodicity foreign to the language represented. In many cases a column of seven dots seems acceptable but more resolution could be obtained by using more dots. In some cases other solutions such as multiple line printings could be acceptable. Additional research in these directions is being carried out by us.

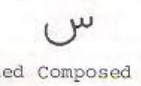
It is interesting to contrast the processes of composing characters out of a set of basic symbols, in case the composition is carried out by an operator versus what happens when the composition is done by an electronic device such as a microprocessor. In the first case, one will try to hold the number of superpositions as small as possible, most likely limited to two superpositions at the most. This will require more basic symbols to be available and will encourage making simplifications and trade-offs. In a dot matrix system backed up by an intelligent processor unit, a smaller number of basic symbols will tend to be more important than the requirement of a reduced number of superpositions. The time required to perform the superpositions will be so short that it can legitimately be neglected. The compositions shown in fig.4, and requiring three basic symbols will then be favored over those shown in fig.5, although the latter require only two symbols but imply some compromising. Moreover the required superpositions will be executed automatically, relieving the burden of the operator. Each typed character will act as a call for a small program, performing the compositions and commanding the print-out of the required dot columns.

Because in Arabic, the form of the character can only be determined once the next character is known, it will be necessary to delay the print out over one character. For the convenience of the operator, a display of a single character might be useful, in order to indicate the last character keyed in by the operator.

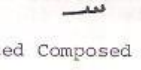


(a) Original Character

(b) Three Common Signs



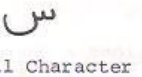
(c) A Printed Composed Character



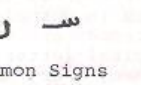
(d) A Printed Composed Character

Fig. 4.An Example of an Arabic Character Composed by the System

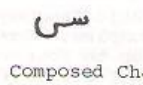
With many dot printers a complete line is printed in one, undeconposable operation. In such a case a display of several characters could prove to be a useful complement to a printer, were it only to facilitate corrections.



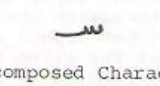
(a) Original Character



(b) Two Common Signs



(c) Printed Composed Character



(d) An Uncomposed Character

Fig. 5. An Example of an Arabic Character Composed by Operator

Many display procedures are compatible with the dot matrix or the dot column system, in that sense that almost the same programs, stored data and system organization can be used for printing as well as for the display. As such we mention e.g.cathode ray tubes (CRT) displays or displays that use light emitting diodes (LED) or small glow discharges as dots. It has also the advantage that the same system can be used for printing and displaying characters belonging to different groups of languages upon receiving the appropriate command signals. This can indeed be done without any mechanical modifications, as long as the number of dots in the column is sufficient to obtain an acceptable resolution for the different languages. In fig. 6 we show e.g. a mixed print-out of Arabic and Latin characters.

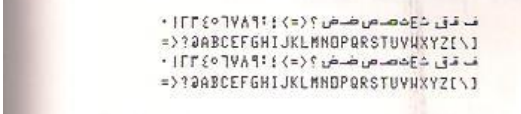


Fig. 6. An Example of a Mixed Print-Out of Arabic and Latin

One might also suspect that a dot matrix print-out will prove advantageous in the future, when automatic text reading will become widely used, because the characters formed in such a procedure are already in a coded form on paper.

THE INTELLIGENT TERMINAL AND ITS ADVANTAGES

It has been shown that adding some form of intelligence to a typewriter greatly enhances its possibilities and convenience, especially for printing non-Latin characters. The use of a dot matrix or column printer deserves special emphasis in this context, because it leads to a very flexible system. The combination of such a printer and of a microprocessor leads to a print-out of better quality and appearance and relieves the operator from the selection of the approbriate character form to be used and also relieves him from some tedious work related to character composition. Such a system has also the advantage of allowing character sets of greatly increased size. It then becomes possible to print texts formed out of characters belonging to more than one script, or in several styles of typing belonging to the same language (see fig. 6). Advantageous is also that printers and display units for different scripts could all be implemented using the same mechanical and electronic hardware. The specificity of the machine with respect to its character font would reside in a set of data (character file) and a driver program, both stored in ROM-memory.

For the Arabic language, the driver program would roughly have to perform the following tasks: (a) Decode the command received from the keyboard. (b) Determine which variation of the typed character to use, considering the previous and the following character. (c) Compose the required variation out of some elementary symbols stored in ROM. (d) Store the composed character in the "current line file" in the RAM-memory of the processor. (e) Issue some warning message to the operator, e.g. indicating the nearing of the end of a line. (f) Once the "print" or "new line" command is received, adjust the spaces in order to justify the line at both margins. (g) Display and/or print the justified line and move the printing head to the first position of a new line. Additional research in this direction is also being carried out by us.

Because of its possibilities for composing characters there would be no great difficulty in adding pronunciation signs, provided the vertical resolution is sufficient so as to accommodate both character and phonetic signs.

It has already been mentioned that the operator's task could be much simplified. The fact that the machine would have a rather simple keyboard with fewer keys also leads to a simpler operation and to a higher work load. We think however that, once an intelligent microprocessor is included in the system, one could still exploit its possibilities to a greater extent. One could e.g. think of loading the machine with programs specifically designed for the most frequently requested jobs, such as form filling or bookkeeping, for storing the typed text on magnetic tape or disk or for transmitting it. We are convinced that such machines will find wide spread use as computer terminals, for book and text setting, as electronic composers for office work and even as a replacement for the present-day typewriters because of their higher turnout and better

performance. All these applications will of course only become economically feasable, once microprocessors and memory blocks will reach still lower prices than nowadays. One can be confident that this will happen, because typewriters for non-Latin as well as for Latin characters represent a considerable market volume and with a still important potential for growth.

CONCLUSION

A dot matrix solution for printing and displaying Arabic characters has been designed using two matrix sizes, namely 7 x 7 and 14 x 7 dots. Because of the usefulness of having a matrix of variable length, the advantages of using a dot column printer were discussed. The incorporation of a microprocessor in such a printer overcomes the different problems encountered in printing non-Latin characters. The advantages of doing so are mainly derived from the extensive flexibility obtained. Some uncommon possibilities of such a machine are discussed.

ACKNOWLEDGEMENT

Part of the work covered in this paper was supported by Wang Europe, Belgium.

Egyptian High Commitee
of Automatic Control

5th Conference on

AUTOMATIC CONTROL AND SYSTEMS ENGINEERING

Cairo, The A.R.E.
November 17-19, 1979

A LOW COST BILINGUAL MICROCOMPUTER DISPLAY SYSTEM

Dr. Salwa El Ramly* and Dr. Mohamed A. El Hamalaway**

* Communications Section, Electrical Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt.
** Systems and computers Engineering Department, Faculty of Engineering, Al Azhar University, Cairo, Egypt.

Abstract. Many of the existing microcomputer systems suffer from the high cost of the peripherals.

In the present work a standard TV receiver is used as a display unit in a microcomputer configuration. This has led to a reduction in the system cost.

Both Arabic and English texts are displayed on the system.

The data to be displayed are treated to correspond to a specific V.H.F. channel in order to be fed externally to the antenna jack, without any access to the internal circuitry of the TV.

The different modules forming the system are described, showing how the data are treated in each module.

1-Introduction

One of the determining factors of the cost of a microcomputer system is the display unit. However a standard TV set (which is generally of much lower price than the existing display units) can be used for this purpose. Doing so releases the user from the obligation of purchasing a special display unit.

The use of a TV set requires an interface between it and the keyboard of the microcomputer system.

2-System Description

The system described here is used as an interface between a bilingual keyboard and a standard TV set. The keyboard comprises Arabic characters besides the usual English alphanumeric characters.

While the English letters have only two shapes (capital and small), the Arabic ones have different shapes depending on their position in the word (at the begining, middle or at the end of the word). The number of letter shapes vary from only one such as the letter (ﺩ) to four shapes such as the letter (ﺪ)[1].

When inputting data to the computer it is recommended to have on the keyboard only one shape for the Arabic letter in order to speed up the inputting operation and to minimize the probability of error (by reducing the number of keys on

the keyboard). To do so, it is necessary to insert an intelligence to study the status of the written chracter to choose its proper shape according to its place in the word. It have also to write the chosen shape in a RAM which is continuously scanned and displayed on the TV screen.

The scanning operation is established by a hardware interfacing the RAM to the TV set. The data to be displayed are treated to correspond to V.H.F. channel number 2 on the commercially available TV sets in Egypt, in order to be fed externally to the antenna jack, without any access to the internal circuitry of the TV. However a further reduction in the cost of the display system would be by choosing a certain type of TV sets, and eleminating the modulation block and feeding the output of the system to the TV video section input (with necessary conditioning of the signal).

The whole system can be summerized in the block diagram of figure (1).

3-The Keyboard

Our keyboard is a bilingual one i.e. both English and Arabic characters are considered. There are character keys and control keys on the keyboard.

3-1-Character Keys

There are 44 keys of them. They form the alphabets group, the special signs group and the numerals group. The same arrangement of character keys in both Arabic and English standard keyboards is kept in order not to confuse the operator, although some of the Arabic characters have more than one shape.

This is considered to be an intermediate stage in the development of the system as having more than one key for the same character represents a redundancy in the system hardware, as the system generates one code from these different key positions and process that code using the intelligence of the machine to choose the appropriate shape. The development of a new Arabic/English keyboard is an area for further work.

3-2-Control Keys

3-2-1-Carriage Return: Is used when it is required to start writting a new line.

3-2-2-Master Mode: It indicates the mode of character, is it Arabic, English or graphic signs.

3-2-3-Clear Page: Is used when it is required to clear the written page to print another one.

3-2-4-Current Mode: The operator may need to print some English characters although he is in the process of printing Arabic text or vice versa. Such key is used to indicate such situation.

3-2-5-Stop Intelligence: It may be required not to use the intelligence of the machine, e.g. when writting mathematical formulas in Arabic.

3-2-6-Space: Used to insert spaces between characters and words.

3-2-7-Cursor: A cursor indicates the position at which the next character will be inserted. It appears as a straight line below the character and is of variable length. It can be moved to the right, to the left, upwards, downwords or to the home position.

3-2-8-Line/Shape: These are 4 keys used to indicate which of the four lines will be used.

Specially when writting mathematical formula we have to use some forms of the Arabic letters not following the normal writting rules. This is managed by stopping the intelligence of the machine and then choosing the appropriate character form using these four keys.

In the normal operational mode of the machine we use these keys in addition to the main character keys to select the required character.

3-3-Reading the Keyboard

It is managed using the 8255 programmable peripheral interface chip.[2] Two I/O operations would determine which key is being pressed and whether we have key rollover or not.

4-Intelligence Module

This software module is the heart of the whole system. It is based on the 8080 microprocessor which control all the I/O operations and the selection of the proper shape of the character.

The Arabic characters are devided into three groups to facilitate the selection process. Group (1) contains the characters that have only one shape independent of its position in the word e.g. the Arabic (ﺩ) character. Group (2) contains the characters that have only two shapes one of which is used only when the character comes at the end of the word. The character (ﺱ) is an example of this group. Group (3) contains the rest of the characters. These characters depend on both previous and following letters in the word. The (ﻫ) is a good example of this group.

The machine performs in addition to the selection of the proper shape of the character the following tasks:

4-1-Automatic initiation of a new line once the available space in any line is utilized.

4-2-Adjusting the text to justify the line at both margins by generating the extension character (-) in some of the words in Arabic texts and by breaking the last word in the

Latin text.

4-3-Managing the display of Latin characters in an Arabic text and vise versa, and adjusting the direction of letters flow on the screen for each language.

4-4-Generating the cursor underneath the character to be printed.

4-5-Using the cursor control block, any given character can be changed or deleted. The machine would thus make the necessary modifications, as for the surrounding characters and justifying the modified line at both margins.

The machine writes the data to be displayed in the current page file in a form ready for display on the TV screen.

5-TV Set And RAM Interfacing

This section is a determining factor in the design. All the timing and the sequencing of operations is determined by the standard synchronization operation of the TV set.

As is adopted in Egypt, the TV picture is displayed 25 times per second by two frames each of them consisting of 312.5 lines summing up to 625 lines/picture. Horizontal sync pulses which are repeated each 64 u sec together with vertical and equalization pulses must be supplied to the TV set in order to have a stable pattern on the screen. Thus a sync generator derived by a master clock is indespensible for such a system.

The writting operation on the TV screen is also to be planned. Vertical and horizontal margins must be thought of. For the present system 7 scanning lines are devoted to the display of one written line (to be named one row in order not to confuse with a scanning line). Also a space between two rows consists of 14 scanning lines (thus we have a double space separation between two rows). It is intended to write only 24 rows (thus having 23 spacings), thus a total of $24 \times 7 + 23 \times 14 = 490$ scanning lines are being the effective lines; all the rest lines ($625 - 490 = 135$) are to be blanked. Thus in addition to the standard value of 20 vertical blanking lines in each field, also a number of scanning lines are blanked each field. As it is known in interlaced scanning, one field being with a complete horizontal line and ends with a half one, while the other begins with half a scanning line and ends with a complete one. As a consequence the 47.5 additional blanking lines are devided into 23.5 and 24 lines, placed at the beginning and at the end field in accordance with its case. This can be seen in figure (2).

It is seen that a sequence of 245 active lines are separated by 68 and 67 blanking lines alternatively well timed with the ordinary 20H vertical blanking as shown in figure (2).

Not all the 245 horizontal lines will be used fully. As it is known a time of 0.16H is devoted for horizontal blanking (for fly back of the beam horizontally), the rest of 0.84H must contain left and right margins. Here only 0.75H is used for writting and a margin of 0.045H is left from both sides.

Now in the 0.75H equalling 48 u sec. the reading pulses must be applied to the RAM in order to search for the text to be read. As the TV has a standard of 5 MHz bandwidth, so we make use of 240 dots/horizontal line.

The necessary reading pulses are generated using the configuration shown in the block diagram of figure (3).

The scanning takes place as follows; the first 240 dots read the written information in the RAM and displays it on the TV screen from left to right. The display then corresponds to the first line of the first row. The following 240 dots read the written text from the RAM and display it on the third scanning line which correspond to the third line of the first row. The fifth then the seventh lines of the first row will be displayed in a similar way. After that, follows the second line of the first space then the fourth line, then the sixth etc... . In the following field i.e. after the elapse of 312.5 scanning lines from the beginning then comes the scanning of the second line of the first row, then the fourth line of the first row, then the sixth line of the first row follows.

Between the occurrence of the reading pulses and the response of the RAM to it giving the data to be displayed there ellapses some time delay (τ). Thus the sync pulses must be delayed by the same amount before being added to the RAM output before being applied to the vestigial side band modulation section, whose output is applied to the input of the TV set. This is shown in figure (4).

6-Reading From The RAM

Each character is represented in a dot matrix form. The matrix consists of colums or seven dots height and of a variable length for Arabic letters and fixed for English. The characters after being processed by the microprocessor and the proper shape is chosen, are stored in the RAM memory. Only seven bits of the byte is used, the eighth bit indicates end of character, and is not displayed.

The TV display can be seen as multiple rows each consisting of 240 x 7 bits.

The scanning process begines with the first address of page, which is to be incremented by 240 after the scanning of the first row is accomplished (1,3,5,7 or 2,4,6) depending on which field is being scanned). The scanning process is managed totally by hardware, during a hold status of the

microprocessor, letting the reading operation to be synchronized with the TV signals.(3)

The same process is repeated for each row. Counters A & B in figure (5) manages such operations, while the output selection circuit in the same figure manages the selection of the correct line (bit in the row to be scanned).

7-Conclution

The system described represents a universal approach to low cost Arabic/English display systems utilizing a standard TV set.

The data to be displayed are treated to correspond to a specific V.H.F. channel in order to be fed externally to the antenna jack, without any access to the internal circuitry of the TV.

The used keyboard comprises the keys found on both Arabic and English keyboards.

8-References

1-M.C. Vanwormhoudt and Mohamed A. El Hamalaway, Remarks about Printing and Displaying some non-Latin Characters; International Federation of Automatic Control (IFAC) Conference on Systems Approach for development; Cairo, Egypt; November 26-28 th, 1977; pp. 185-190.

2-Intel Component Data Catalog, Intel Corp., Santa Clara, CA., U.S.A.; 1978; pp. 12.76 - 12.94 .

3-Michael T. Gray; Microprocessors in CRT Terminal Applications: Hardware/Software Tradeoffs; Computer, Vol. 8, No 10, pp. 53-59; October 1975.

Fig.(1) System Block Diagram.

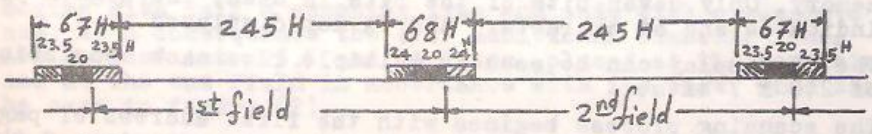


Fig.(2) Sequence of Required Vertical Blanking.

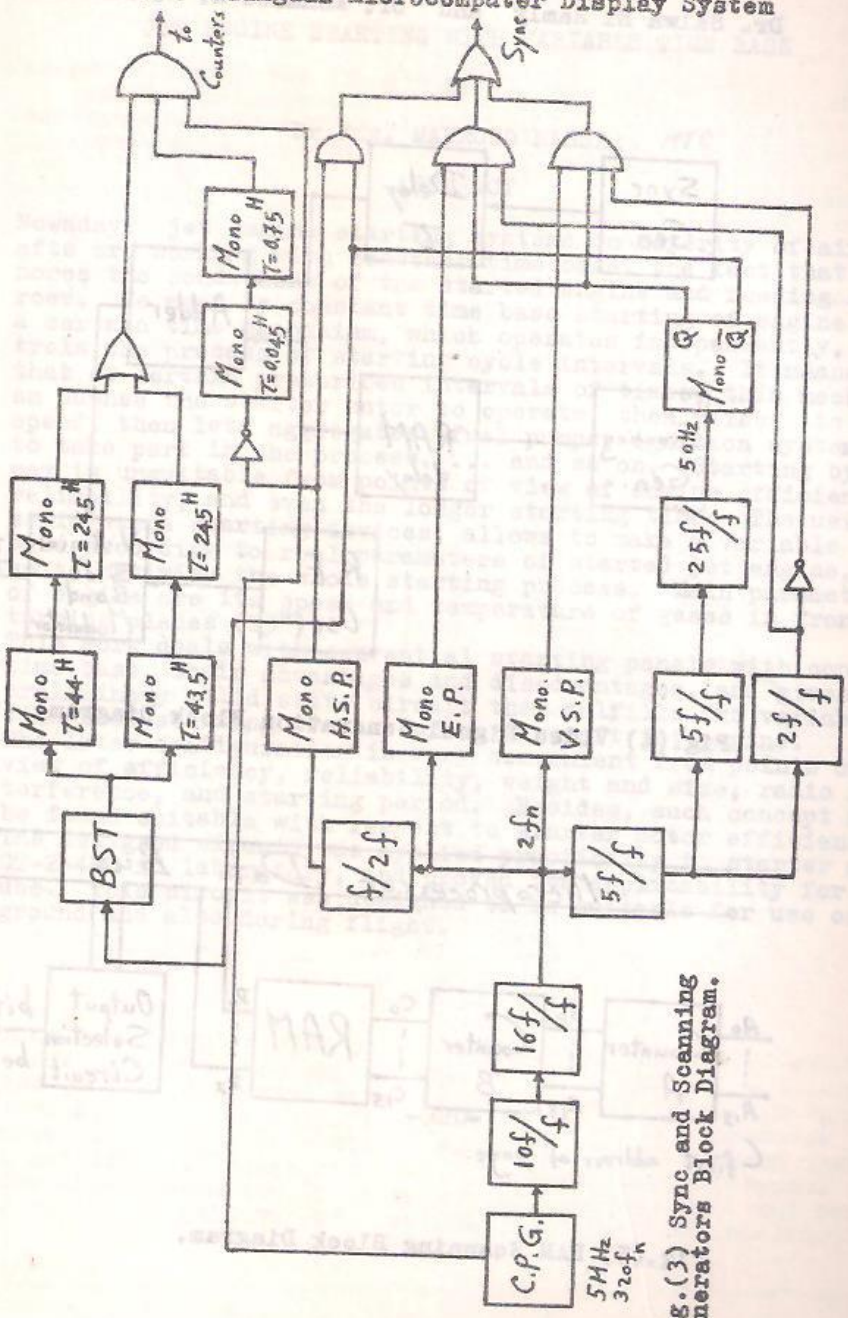


Fig.(3) Sync and Scanning Generators Block Diagram.

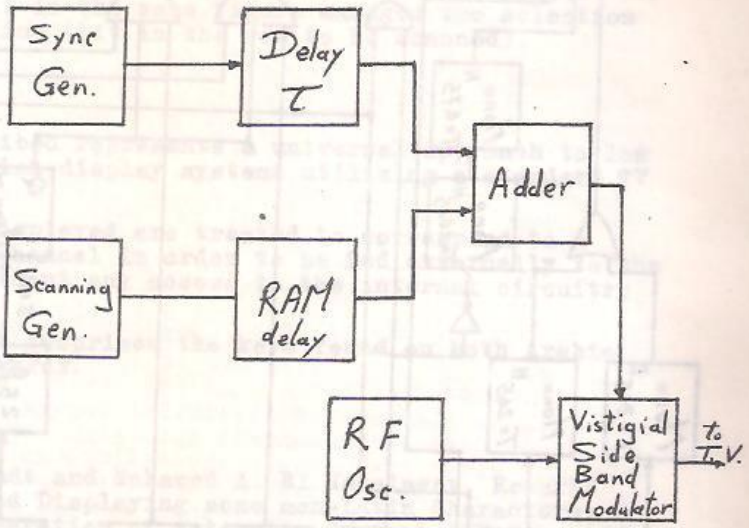


Fig.(4) Video Signal Generation Block Diagram.

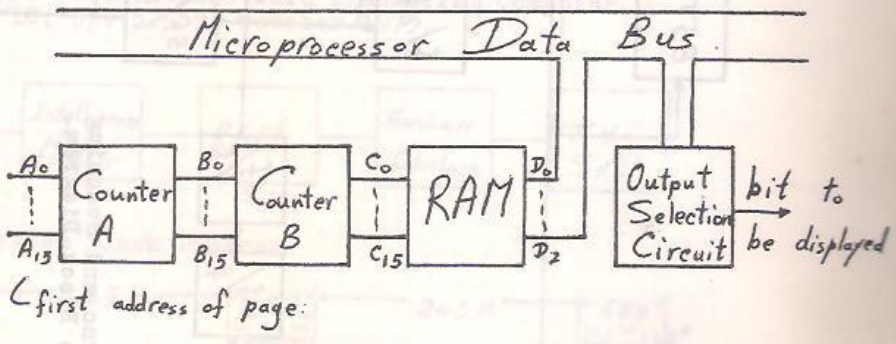


Fig.(5) RAM Scanning Block Diagram.

STATISTICAL DISTRIBUTION OF ARABIC LETTERS AIDS TO THE DESIGN OF A NEW KEYBOARD

S. H. El-Ramly* and M. A. El Hamalawy**

**Department of Electronic and Computer Engineering, Faculty of Engineering, Ain-Shams University, Cairo, Egypt*
***Systems and Computers Engineering Department, Faculty of Engineering, Al Azhar University, Cairo, Egypt*

Abstract. One of the difficulties encountered in treating arabic texts by computer is the low speed of inputing data due to the large number of keys on the keyboard(some of the arabic letters have more than one shape: it ranges from only one shape up to four distinct shapes depending on the previous and the following letters). There are many attempts to reduce the number of such keys. One method is to use an intelligence unit comprising a microprocessor, with a keyboard having only one shape for each letter, with a software program to take care of the determination of the proper letter shape to be typed. However the statistical distribution of letters enables a better arrangement of letters on the board to follow the different degrees of skillness of each finger. The present work treats all these problems, also a new alphabetic code is designed for minimum redundancy of the coded data.

Keywords. Printers; coding; statistics; artificial-intelligence; computer peripheral equipment; man-machine systems; computer interfaces optimal systems.

INTRODUCTION

Many applications of computers in the public survicing centers such as banks, insurance companies, etc... have shown the necessity to input the data in the native language of each country. The use of arabic language in communication with the computer is in growth and one of the problems encountered is the problem that some of the arabic letters have more than one shape; actually the number of different shapes can go to four of them. The large number of keys on a a keyboard increases the difficulty of feeding data by increasing the probability of error(due to confusion) and the reduction of the writing speed. This led to the need of a new keyboard having only one key for each letter, managed with an intelligence unit able to choose the proper shape of the character depending on its position in the word (El Ramly and El Hamalawy 1979). Such a problem and also others show the necessity of having the statistical

distribution of the arabic letters in written texts.

STATISTICAL DISTRIBUTION OF ARABIC LETTERS

When calculating the distribution of arabic letters one must notice the following:

1. The choice of essays from different sources, different topics and of different authers.
2. Essays of both scientific and litteral nature must be considered.
3. The amount of data to be taken is determined by the study of the amount of error encountered.

Notes (1) and (2) need no comment. As for note (3), we proceed as follows: the measurement of a probability P using a limited sample size N leads to an expected root mean square error given by (Bendat, 1971 and El-Ramly, 1976):

$$e = \frac{1}{\sqrt{NP}} \qquad (1)$$

Thus the root mean square error is a function of both the sample size N and the measured probability P. The error is not the same for all the letters; the least frequent of them are the most difficult to measure.

The probability distribution obtained is tabulated in Table 1 , together with the expected root mean square error encountered in the measuement. From the table it can be seen that the maximum value of error is 10% except for two letters (غ ، ظ) , actually it is less than 5% for half the letters.

CODING OF THE ARABIC LETTERS

The processing of arabic texts (storage, translation,etc...) necessitates the use of an efficient code to match the probability distribution of the arabic letters. In other words it is required that shorter codes be attached to more frequent letters and longer codes correspond to less frequent ones, (Reza,1961).

There are many methods of encoding a given probability distribution, the most efficient of them is the Huffman's code. However for encoding languages there exists alphabetical codes which preserves the alphabetical order of dictionaries in addition to the above recommendation for the length of coded messages; also it has the prefix property (i.e. no encoded letters can be obtained from each other by the addition of more symbols). The Gilbert-Moore alphapetical encoding method (Gilbert and Moore 1959) is used and the result is tabulated in Table 1 together with the result of using the Huffman's minimum redundancy code.

The average cost using Huffman's code is found to be 4.083681 digits/letter while that using the alphabetical code is found to be 4.240586 digits/letter which is so close to that of Huffman.

THE KEYBOARD

The knowledge of the arabic letters distribution led to the layout of a new keyboard. Figure 1 is an illustration of such a keyboard where the keys are ranged in a matrix form

Table 1 :Probability Distribution of Arabic Letters and their Codes.

Letter	Probability	e %	Huffman code	Alphabetic code
space	0.188318	1.4725	00	00
أ	0.151014	1.6475	101	010
ب	0.029286	3.7300	01111	01100
ت	0.059718	2.6150	1000	01101
ث	0.005071	8.9900	11001111	011100
ج	0.013252	5.5500	010101	011101
ح	0.018324	4.7250	110001	011110
خ	0.004908	9.13	11001110	011111
د	0,027814	3.83	01110	100000
ف	0.008835	6.80	1100110	100001
ر	0.039921	3.20	11011	10001
ز	0.005399	8.70	0110100	100100
س	0.018651	4.675	110010	100101
ش	0.007853	7.21	0110101	100110
ص	0.008835	6.80	1100001	100111
ض	0.004253	9.80	11000000	1010000
ط	0.011616	5.94	010100	1010001
ظ	0.001963	14.435	110000010	101001
ع	0.026014	3.965	01100	10101
غ	0.002617	12.50	110000011	101100
ف	0.022414	4.27	111001	101101
ق	0.014234	5.0355	011011	101110
ك	0.018815	4.66	111000	101111
ل	0.094731	2.0775	1111	1100
م	0.042866	3.0875	11101	11011
ن	0.039430	3.22	11010	110100
هـ	0.025359	4.0125	01011	110101
و	0.044829	3.020	0100	1110
ى	0.072480	2.375	1001	1111

to simplify the description . The new keypoard has the following characteristics:

1. Both Arabic and English alpha - numerals are put together on the keys . The English characters have the same position as for a traditional machine, and are written in the upper half of the key. The arabic alpha-numerals are written in the lower half of the keys.

2. Some of the keys have shift positions, which is devoted for characters written on the right-hand side of keys. The shift position for Arabic mode of operation can be different than that for English mode.

3. The keys to be pressed by the left hand have Arabic characters that have probabilities summing up to 0.516 which is so close to 0.5 which is the condition necessary for the two hands to share the work equally likely. Note that the statistical

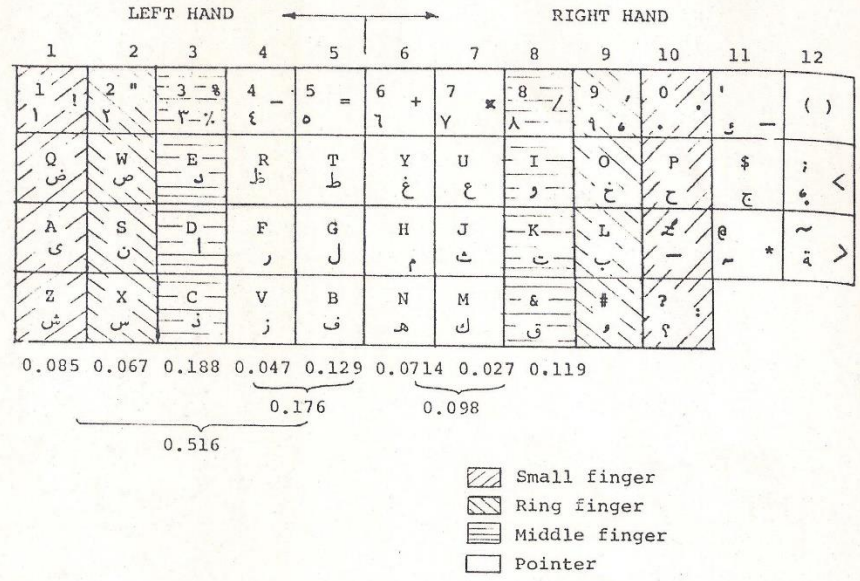


Fig.1. New bilingual keyboard

distribution given in this paper did not include neither the special signs nor the numerals (due to the very large amount of data required to calculate their probabilities). Thus the actual sum of the probabilities of the keys pressed by the left hand may approach 0.5 in a better way.

4. The pointer and the middle fingers are more skill to press the keys than both the ring and small fingers; so each of the first two takes care of characters whose probabilities sum up to 1/6 (one third of a half) while the small and ring fingers together have characters with probabilities summing up to 1/6.

5. For most of the fingers, the character with the higher probability is put in the middle raw of the alpha-keys, such as to minimize up-down displacements.

6. There are 32 basic Arabic characters (the Arabic alphabet excluding لا which is composed by the machine intelligence, plus ة , ء , ى , م (Madda). The compound characters ؤ , ئ, أ, لأ, لـ, آ, لآ, and ﻟ are composed by the machine intelligence.

7. Letters of the same group like (ر,ز) and (د, ذ) are assigned to the same finger for simplicity, or they may be assigned to two or three horizontally adjacent keys such as (ج, ح, خ), (ص, ض) (ش, س) (غ, ع) (ب, ت, ث) and (ط, ظ).

8. There are three kinds of dashes: the dash used both for English writting and as a "minus" sign in Arabic

and English modes (key(1,4)) which is typed above the line, an underlining dash (key(1,11)) which is typed under the line and a third dash used for arabic letters extention which is optional to be used when variable size characters is adopted (key(3,10)) and is typed at the line level.

9. Some of the characters kept their traditional positions whenever it is possible such as (ش, ص, (غ , ع), (خ , ح), ل or near to it such as (س,ش) , (ى).

CONCLUSION

The present work is essential for various applications, the probability disribution of arabic letters is measured here with a known small amount of error. The code proposed is very near to the minimum redundancy code which increases the transmission effeciency for teletype machines. Also, the suggested keyboard seems to be an optimum one for the number of keys and their positions on the board.

REFERENCES

Bendat,J.S., and A.G. Piersol (1971). Random Data Analysis and Measurement Procedures. John Wiley,New York. pp. 177-179.

El-Ramly, S.H. and P. Duhamel (1978). Prévision de l'erreur statistique dans la mesure de desités de probabilités. L'onde électrique, 58, pp. 375-382.

El-Ramly, S.H. and M.A. El-Hamalawy (1979). Low cost bilingual Microcomputer System. Proc. of 5th Conference on Automatic Control and System Engineering. Cairo, Egypt . Nov.1979.

Gilbert, E.N., and E.F. Moore(1959). Variable length binary encodings. Bell Syst. Tech.J.,38,pp. 933-968.

Reza F.M. (1961). An Introduction to Information Theory. McGraw-Hill, New York. pp. 131-138.

المجلة المصرية للحسابات العلمية

المجلد ٦ العدد ١ يناير ١٩٨٣

تصدرها
الجمعية المصرية للحساب الآلي

APGUST: A Standard for Coding Arabic Character Sets.

Mohamed A. El Hamalaway

Department of Systems and Computers Engineering,
Faculty of Engineering, Al Azhar University,
Cairo, Egypt.

Abstract: An Arabic 7-bit coded character set for information processing interchange is proposed. The proposed standard (APGUST) is based upon the ISO 646 standard specification of 1973. The APGUST handles all Arabic-based character sets: Arabic, Farsi, Gawi, Urdo, Swahili and Turkish character sets.

INTRODUCTION

The problem of an Arabic standard character set for information processing interchange is bleeding for a standardization. New models of computers manufactured by the same company are not compatible in their Arabic character set and in their code. Companies don't follow standard code as there is none.

Hereby we propose a standard 7-bit code for the Arabic character sets(APGUST); which stands for Arabic, Farsi, Gawi, Urdo, Swahili and Turkish languages.

The standard is set forth for public discussions.

PROPOSED STANDARD

In the APGUST standard all control codes and graphic symbols are kept the same as in ISO 646 stanmard, international reference version[1]. That was made for possible use of ready made Lattin packages. One have to note that the endings of words and phrases in the Arabic group of languages as well as in the Lattin group is the same. For example full stop in both Arabic and Lattin has the same code (2/14).

Table (1) shows the international APGUST standard for Arabic group of languages. Arabic characters are shown in the table. Positions 5/11, 5/12, 5/13, 5/14, 6/0, 7/11, 7/12, 7/13 & 7/14. could be used for characters coming from languages other than the Arabic one; to allow for the national characters to be included in the standard.e.g. ڤ in Farsi and Urdo.

Table (1) : International APGUST Standard Chart.

b.	b.	b.	b.	row / column	0	1	2	3	4	5	6	7
				b.	0	0	0	0	1	1	1	1
				b.	0	0	1	1	0	0	1	1
				b.	0	1	0	1	0	1	0	1
0	0	0	0	0	NUL	TC. (TLE)	SP	٠§	ة	ج	٠°	ـجـ
0	0	0	1	1	TC. (SOH)	DC.	!	١§	ش	ض	ـشـ	ـضـ
0	0	1	0	2	TC. (STX)	DC.	"	٢§	ق	ف	ـر	ـفـ
0	0	1	1	3	TC. (ETX)	DC.	#	٣§	ذ	س	ظ	ـسـ
0	1	0	0	4	TC. (EOT)	DC.	¤	٤§	ب	خ	ـبـ	ـغـ
0	1	0	1	5	TC. (ENQ)	TC. (NAK)	%	٥§	ق	ه	ـقـ	ـهـ
0	1	1	0	6	TC. (ACK)	TC. (SYN)	&	٦§	ل	د	ـلـ	ـطـ
0	1	1	1	7	BEL	TC. (ETB)	'	٧§	ء	ص	ا	ـصـ
1	0	0	0	8	FE. (BS)	CAN	)	٨§	ت	ث	ـز	ـثـ
1	0	0	1	9	FE. (HT)	EM	(	٩§	خ	ح	ـخـ	ـهـ
1	0	1	0	10	FE. (LF)	SUB	*	:	ن	ى	ـنـ	ـيـ
1	0	1	1	11	FE. (VT)	ESC	+	؛	ر	]°	ـمـ	}°
1	1	0	0	12	FE. (FF)	IS. (IS4)	،	>	ك	\°	ـكـ	!°
1	1	0	1	13	FE. (CR)	IS. (GS)	-	=	لا	[°	ـزـ	{°
1	1	1	0	14	SO	IS. (RS)	.	<	ـة	^°	ـوـ	¯°
1	1	1	1	15	SI	IS. (US)	/	؟	ح	_°	ـئـ	DEL

° National use positions.
§ Shape could change.

It should be noted that no sacrifice was made as for the number or the shape of any of the Arabic letters, as well as for other languages covered by this standard.

The numerals shapes must be allowed to change amongst the languages covered by this standard, but the context must remain the same. For example the number 5 (code 3/5) looks like ٥ in Arabic while it looks like ۵ in Farsi and Urdo. This is true for positions 3/0 through 3/9 in table (1).

Codes used for specific notations e.g. (have been changed only in shape to match the right to left direction of writing for all Arabic group of languages.

Each code indicates one character except the ﻻ (code 4/13) which indicates two characters. This represents no problem in sorting texts as ﻻ (code 4/13) comes after ﺝ (code 4/6) and before ﺟ (code 6/6).

The same information is processed sometimes in more than one shape e.g. ﺑ (code 4/4) and ـﺒـ (code 6/4). This represents no problem in sorting. The character ﺐ (code 4/8) comes before ﺒ (code 6/8).

The main drawback of this proposed standard is that Arabic characters are not in their sorting order.

The proposed standard has payed some attention to the statistics of the occurance of Arabic characters in Arabic texts[2]. It should be noted that the proposed standard adopted to a large extent the existing keyboard configuration of most Arabic typewriters.

The proposed standard has been applied and found satisfactory[3]. It should be stressed that such proposed standard have to float for some time to allow enough feedback for a better standard hopefully based upon this ARGUST.

CONCLUSION

An Arabic 7-bit standard code for information exchange is proposed. Such standard covers Arabic, Farsi, Gawi, Urdo, Swahili and Turkish languages.

REFERENCES

1- 7-Bit coded character set for information processing interchange ISO 646 Standard; First edition 1973; ISO.

2- S. H. El Ramly and Mohamed A. El Hamalawy; Statistical distribution of Arabic letters aids to the design of a new keyboard; IFAC Conference, November 24-27th, 1980; pp 515-519.

3- Patramo Computers Catalog; Patramo systems co.; New York, U.S.A., 1982.

المؤتمر الدولى العاشر

للاحصاء والحسابات العلمية والبحوث الاجتماعية والسكانية

٣٠ مارس ــ ١٠ ابريل ١٩٨٥

Conformance of Current Codes for Arabic Information Interchange to Present Software

by

Ali Ali Fahmy
Military Technical College,
Computer Department,
Cairo, Egypt.

and

Mohamed A. El Hamalaway
Department of Systems and Computers Engineering,
Faculty of Engineering,
Al Azhar University,
Cairo, Egypt.

Abstract: Few of present codes for Arabic information interchange have been screened as for their agreement with present computer systems software. It was shown that non of these codes would accomodate present system software. Sorting and searshing cause serious problems. This calls for fundamental changes in current system software/codes. Non-conformance with ISO-646 standard as for special signs and numerics causes serious problems in some codes. Other codes 'with minor modifications' could avoid such non-conformance.

INTRODUCTION

This paper traces down the following codes which; to the best of ou knowledge; claim to be standard:

1. AFGUST, 1983(1)
2. ARCII, 1982(2)
3. ASMO, 1982(3)
4. ECMA, 1982(4)
5. SABA-MURSI, 1983(5)
6. SASO, 1983(6)

Our main objective in studying these codes is to see their conformanc with present software. Such software; which forms the conformanc criteria; includes standard languages, widely known operatin systems, common functions essential in many computer systems (e.g sort, edit,...) as well as widely known application packages.

We have directed ourselves to the study of the internal representatio of these codes inside any computer system as for its effect on runnin software on machines employing these codes. Mostly we had paye attention to the linguistic functional differences of these codes Thourough linguistic differences will be shortly published in anothe paper by us.

COMPARISON CHART

The following chart summs up mostly the non-linguistic differences between these codes. The comparison points in the chart have been chosen because of their direct effect on using many system/application software. This is due to the special meaning assigned to the symbols covered by each comparison point.

S.N.	CRITERIA	AFGUST	ARCII	ASMO	ECMA	SABA	SASO
1	BILINGUAL MODE	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
2	DIRECT SORT	NO	NO	NO	NO	NO	NO
3	PREPROCESSING/ DIRECT SORT	NO	NO	NO	NO	NO	NO
4	DIRECT SEARCH	NO	NO	NO	NO	NO	NO
5	"'"IN-STRING SEMANTIC	F.C.	NO	YES	NO	YES	YES
6	"," LINGUISTIC SEMANTIC	YES	YES	YES	YES	YES	YES
7	TATWEEL	F.C.	F.C.	F.C.	NO	F.C.	F.C.
8	REAL NUMBERS ARITHMATIC	F.C.	NO	F.C.	F.C.	F.C.	F.C.
9	NUMERALS	YES	YES	F.C.	YES	YES	F.C.
10	"(",")","{","}" CORRECT SEMANTIC	YES	NO	YES	NO	YES	YES
11	"<",">","[","]" CORRECT SEMANTIC	YES	NO	YES	NO	NO	YES
12	"^" ARITHMATIC SEMANTIC	YES	NO	YES	NO	NO	YES
13	"X","÷"ARITHMA-TIC SEMANTIC	F.C.	NO	F.C.	NO	NO	F.C.
14	"+","-"ARITHMA-TIC SEMANTIC	YES	YES	YES	YES	YES	YES
15	"#"ARITHMATIC SEMANTIC	YES	NO	YES	NO	YES	YES
16	"`","&"COMMAND LANG. SEMANTIC	YES	NO	NO	NO	YES	NO

N.D. means not defined.
YES means criteria and shape are both satisfied by the code.
F.C. means functionally correct but having a different shape.
NO means criteria is not satisfied by the code.

Conformance of Current Codes for Arabic Information Interchange

The following remarks are needed for better clarity of the selected criteria.

1. DIRECT SORT: means using software sort utilities without any modification.
2. PREPROCESSING/DIRECT SORT: means a preprocessing of the input text have taken place before submitting the text to a direct sort utility. A loss of information may occur in this case, and this means a NO on the chart.
3. DIRECT SEARCH: a NO in the chart means a failure in searching for a vowalized / nonvowalized word, or a failure in searching for a syllable in a word.
4. REAL NUMBER ARITHMATIC: relates to the decimal point position in the code chart.
5. CORRECT SEMANTIC: relates to the correct direction of writing Arabic text in order to keep functional compatability.
6. ARITHMATIC SEMANTIC: a NO in the chart means arithmatic signs have different position in the code chart than that of ISO-646(7), or that function does not exist in the code.
7. IN-STRING SEMANTIC and COMMAND LANGUAGE SEMANTIC: has special meaning for some languages. NO means that its plase in ISO-646 is occupied by an arithmatic sign or another code allocated to it.
8. NUMERALS: YES means correct shape and position.

CONCLUSION

Non of the existing Arabic codes for information interchange screened in this paper could run English operating systems and application software without problems. The main area of troubles is a linguistic one. Points 1, 2, 3, 4, 5, 6, 7, 8 & 9 in the comparison chart cover such functional linguistic diferences globally. This calls for fundamental changes in current system software/codes. Non-conformance with ISO-646 standard as for special signs and numerics causes serious problems in some codes. Other codes 'with minor modifications' could avoid such non-conformance. Points 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 & 16 in the comparison chart cover such diferences.

REFERENCES

1. Mohamed A. El Hamalaway; AFGUST: A Standard for Coding Arabic Character Sets; 8th International Congeress for Statistics, Computer Science, Social and Demographic Research; Cairo, Egypt; March 26-31st, 1983; published in Egyptian Computer Science Journal; Vol. 6, No. 1; January 1982, pp. 55-57.

2. ARCII: Arabic Reduced Code for Information Interchange; ALIS RD404.1; Alis Co., Canada; Sept 1982.

3. Draft Proposal for An Arabic Standard; Number 449/1982; ASMO Organization; 1982.

4. Standard ECMA 7-Bit Coded Basic Character Set for the Arabic Alphabet, Final Draft ECMA/TC1/82/7; ECMA; Switzerland; Jan 1982.

5. Mona Mursi and Mohamed Saba; The Standard Arabic Code: Revisited; 8th International Congeress for Statistics, Computer Science, Social and Demographic Research; Cairo, Egypt; March 26-31st, 1983; published in Egyptian Computer Science Journal; Vol. 6, No. 1; January 1983, pp. 32-44.

6. Saudi Standard Draft; Data Processing 7-Bit Coded Arabic Character Set for Information Interchange; Saudi Arabia; 1983.

7. ISO 7-Bit Coded Character Set for Information Interchange; ISO 646 Standard; Second Edition; ISO Organization; 1983.

وقائع المؤتمر

CONFERENCE PROCEEDINGS

Edited By
Dr. Ahmad Bishara

تحرير
د. احمد بشاره

FIRST KUWAIT COMPUTER CONFERENCE

MARCH 27 – 29, 1989

مؤتمر الكويت الأول للحاسوب

27 – 29 مارس 1989

KUWAIT COMPUTER SOCIETY — جمعية الحاسب الآلي الكويتية

KUWAIT FOUNDATION FOR THE ADVANCEMENT OF SCIENCES — مؤسسة الكويت للتقدم العلمي

KUWAIT INSTITUTE FOR SCIENTIFIC RESEARCH — معهد الكويت للابحاث العلمية

KUWAIT UNIVERSITY — جامعة الكويت

A New Font For Arabic Characters Simplifies Recognition Procedure

Salwa H. El Ramly
Electronics and Computers
Engineering Department
Faculty of Engineering
Ain Shams University
Cairo, Egypt

Mohamed A. El-Hamalaway
Department of Systems and
Computers Engineering
Faculty of Engineering
Al Azhar University
Cairo, Egypt

ABSTRACT:

A new and clear shapes of Arabic letters as well as numerals and special characters are proposed. The aim seeked here is to simplify the recognition procedure, speed that procedure up and to reduce memory size. Many applications can be found for this proposal, namely in big financial establishments and in large public service organizations.

1. INTRODUCTION:

In some applications fast automatic optical character recognition is required, such as automatic document handling in banks, insurance companies and other public offices. Many problems are encountered in automatic Arabic character recognition. Cursiveness, overriding, non equal size and many types or styles of writing. These problems are complicated further if they are combined; in other words they must be taken into account all together at the same time. In Arabic, letters may have more than one shape. Actually the mumber of different shapes of a certain character ranges from only one shape (eg. ط ظ و ر ز) up to five shapes(eg. ـه ـھ ھ ـھـ ه). This leads to the increase of the mumber of distinct characters that have to be recognised. This problem could be mostly solved at the writing stage by using an inteligent keyboard having only one shape for each letter, other shapes are deduced through software [1]. At the recognition stage a great effort must be done to discriminate between these large numbers of distinct characters. In litterature few trials are found for automatic Arabic character recognition [2,3,4,5]. The present work is a trial to overcome some of the problems mentioned above by modeling Arabic letters by new shapes. The purpose of this is to facilitate the recognition procedure by reducing the code for each of them and reducing the recognition time through proper choice of character clusters (groups).

2. NEW CHARACTER SHAPES:

In Arabic there are several styles of writing amoung which are

Koufi, Rekaa, Naskh, Diwany and others. The one letter has different shapes in each style. The differnce may be very small or large. This may lead to differnt automatic recognition procedure for each style. For time and effort saving new character shapes are proposed which have the following properties:

a. For each letter the number of different shapes kept are just those required for human eyes recognition and not for the beauty of writing (the number of different shapes is limited to a maximum of three).
b. No overriding is allowed.
c. New shapes must be as simple as possible and in the same time the essential features of habitual shapes must be kept.
d. Some modifications in shapes are necessary to obtain differeut codes for them.

The total number of characters obtaibed is 56 letter shapes in addition to 21 numerals and special signs which sum up to 77 characters. The set of 77 characters is devided into 8 clusters (groups) of different sizes:

a. Concerning horizontal spread: There are two sizes
   -size S is 4 pixels measured from right to left
   -size L is 7 pixels measured from right to left.
b. Concerning vertical spread: There are four sizes:
   -size A is 6 pixels extending from the horizontal base line up to the highest pixel.
   -size B is 8 pixels having 6 pixels above the base line and 2 below it.
   -size C is 10 pixels having 6 pixels above the base line and 4 below it.
   -size D is 12 pixels having 6 pixels above the base line and 6 pixels below it.

The letter shapes are chosen to approch the Koufi font which is the simplest font and is the oldest one; from which other fonts seem to have originated.
The different subsets are described with their horizontal and vertical widths as shown in table (1).
The difference between clusters (NL) and (AL) is that all the righthand pixels are zero.
The letter shapes in cluster (XL); while X stands for either A, B, C or D; are chosen such that the right hand side four pixels do not produce codes of letter shapes constituting cluster (XS), this is to prevent confusion at decoding stage.

3. CODING PROCEDURE:

The codes specific to each character consists of the number of horizontal crossings (comprising the vertical histogram)/the number of vertical crossings (making the horizontal histogram). In the code we count only the variations. For example the letter

(غ) from group (DL) has twelve horizontal crossings (h1, h2,.....h12) which are (0101111111111) from which the parameter H is (0101). Also the vertical crossings (v1, v2,....v7) are (2334322) from which the paramerer V is (23432). This method reduces the length of the code and discards redundant information. The proposed shapes are shown in figure (1), while their codes are given in table (2).

4.RECOGNITION PROCEDURE:

Recognition is achieved through two consecutive decoding processes. The first is to find the cluster in which the character is a member. The second is to detect the character itself. Details are described below:

a.Arabic text used for automatic character recognition is supposed to be written on papers having alignment for both horizontal and vertical base lines at the right handside of the page to indicate the begining of writing. Starting from the righthand side vertical base line pixels of horizontal spread (S); i.e. 4 columns; and vertical spread (D); i.e. 12 raws; are investigated. The horizontal (h1,h2,...h12) and vertical (v1,v2,v3,v4) histograms are obtained from which (H) and (V) are calculated; and the code (H/V) is thus obtained.

By testing the horizontal histogram, the unknown letter is classified in clusters AS, BS, CS or DS (which is an empty cluster) and the code is compared to the codes in the proper cluster (XS).

If a decision is reached the next 4 pixels are investigated to find out the next character. If a decision is not reached then the next 3 pixels are considered with the previous 4 pixels from which (h1, h2, ...h12) and hence H is calculated. Also (v1, v2 ...v7) and hence V is calculated to obtain the code (H/V). The code is compared to cluster (XL) instead of (XS). When deciding whether the unknown character belongs to cluster (AL) or to cluster (NL), v1 is investigated; if it is zero then the cluster is (NL) and not (AL).

b.The characters within a cluster are devided into subgroups each having the same H (but differ in V). The subgroups are arranged from left to right in descending order with respect to their probability of occurrence [1]. Also the characters within a subgroup are arranged from left to right taking into consideration their probability distribution. The measured value of (H) for the unknown character is compared to that of the first subgroup (the most probable) then to the second subgroup,...etc. When a subgroup is identified, the value of the parameter (V) of that unknown character is used to identify it by comparing its (V) value with the (V) values of the characters comprising the subgroup from left to right until a decision is reached. By doing so the processing time is greatly reduced.

The decision tree for cluster (AS) is shown in figure (2) as an example. The recognition process starts from left to compare (H) measured with that of letters (لا) which are the most probable within the cluster and at the same time having the same (H) (H=1). If the two codes are not identical; (H) measured is compared to 0201; which is that of the letter (ـــ) and so on. If the subgroup (لا) is recognised then the measured parameter (V) is used to diffrenciate between the different characters.

The number of subgroups ranges from 2 to 13. The number of characters per subgroup ranges from only one (leading to direct recognition) to four characters. For subgroups having only one character only H is needed. This also gives further reduction of memory size.

Actually for 42 characters out of the 77 information about the horizontal parameter H is only needed to be known. The other 35 characters need both H and V.

5. CONCLUSION:

The new proposed shapes for Arabic characters are found to be useful for office automation in places handling large amount of data e.g. in banks, insurance companies, etc... This proposal for new Arabic characters shapes will help develop fast recognition procedures for special Arabic texts and for different writing fonts which is planned for future work.

6. REFERENCES:

1. S. A. El Ramly and Mohamed A. El Hamalaway; Statistical Distribution of Arabic letters Aids to the Design of a New Keyboard; IFAC Conference on System Approach for Development; Rabat, Morocco; 24-27 Nov., 1980; pp. 515-519.
2. Mohamed A. El Hamalaway and Salwa H. El Ramly; A Novel Arabic Numerals Shapes for Optical Character Recognition; 13th International Conference on Statistics and Computer Science; Cairo, Egypt; 26-31 March, 1988; pp. 127-130.
3. M. F. Tolba and Hani M. K. Mahdi; On the Transformation of Arabic Characters into Unified Mesh Size; 13th International Conference on Statistics and Computer Science; Cairo, Egypt; 26-31 March, 1988.
4. M. A. Sharkawy, M. F. Tolba and E. Shaddad; Fourier Descriptors for Printed Arabic Character Recognition; 13th International Conference on Statistics and Computer Science; Cairo, Egypt; 26-31 March, 1988.
5. H. Y. Abdelazim and M. A. Hashish; Arabic Reading Machine; 10th. National Computer Conference; Riyad, Saudi Arabia; 1988; pp. 733-744.

Table (1) Size Definition of Clusters

Cluster Label	Horizontal Spread	Vertical Spread
AS	S	A
AL	L	A
BS	S	B
BL	L	B
CS	S	C
CL	L	C
DL	L	D
NL	L	B

Table (2) Proposed Shapes Codes

(a)Cluster(AS)

Character	Code(H/V)
ا	1/01
ب ت ث	0201/1212
ج ح خ	10201/12
د ذ	01/12
ر ز	101/123
س ش	101/12
ص ض	201/12
ط ظ	1/1
ع غ	0101/121
ف	0121/121
ق	20121/231
ك	101/1231

(b)Cluster(AL)

Character	Code(H/V)
ل	02021/12121
م	0131/121
ن	012021/121
ه	01/121
و	0101/1321
ي	01/0121
ء	101/021
آ	031/1
أ	012031/121212
ؤ	010131/1231
إ	121/121
ئ	12121/12321
ى	01/12121
ة	101/121
لا	10121/131
لأ	1/1232
لإ	121/1231
لآ	21/0121

(c)Cluster(BL)

Character	Code(H/V)
ب	02101/121
ج	0101/12321
ر	01/01
ز	0101/0121
س ش	012121/121
ص ض	012121/1231
ك	20121/132
ف	121/1
ق	0121/121
ه	12310/12121
ى	012121/131
ي	012121/12321
لا	210/021

(d)Cluster(BS)

Character	Code(H/V)
ت	0101/121
ه	01/1
ة	0102/1212
ـ	010/012

(e)Cluster(CS)

Character	Code(H/V)
و	01/121
م	0121/121

(f)Cluster(CL)

Character	Code(H/V)
ن	0421/1
ن	0120421/121

(g)Cluster(NL)

Character	Code(H/V)
١	1/0210
٢	21/1210
٣	31/1210
٤	1/2320
٥	01210/1210
٦	121/10
٧	21/10
٨	12/10
٩	121/012310
٠	010/010
،	01/01210
؛	1312/10
؟	313/10
!	12121/10
+	2121/120
-	01/10
×	0212/2120
=	01010/20
٪	2120/2120
؟	101/1230
—	010/10

(h)Clust(DL)

Character	Code(H/V)
ع	01/12431
غ	101/12431
ح	0121/232
ح	01/232
خ	0101/23432

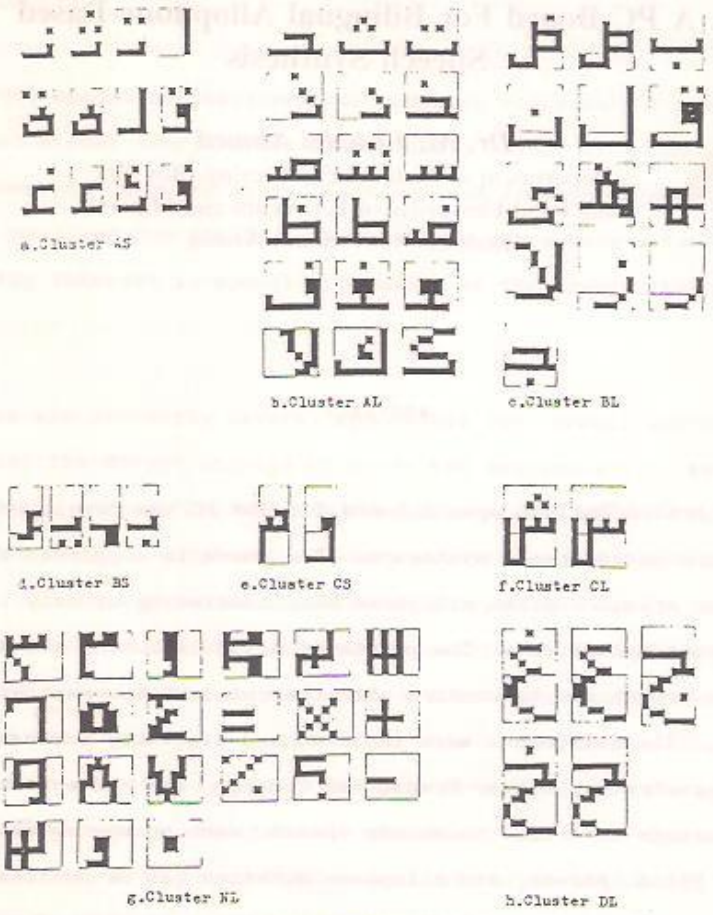


Fig.(1) Characters Shapes

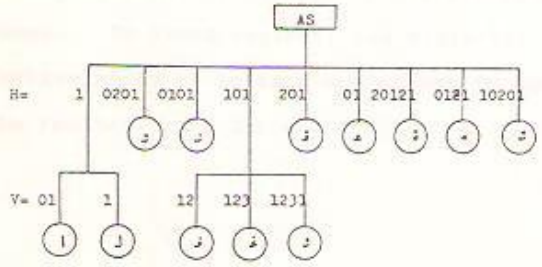


Fig.(2) Recognition Procedure for Cluster AS

PROCEEDINGS OF THE

14th INTERNATIONAL CONFERENCE FOR STATISTICS , COMPUTER SCIENCE , SOCIAL AND DEMOGRAPHIC RESEARCH

CAIRO , EGYPT
25 - 30 MARCH 1989

VOLUME 4

COMPUTER ENGINEERING AND COMPUTER SCIENCE

SCIENTIFIC COMPUTING CENTER
AIN SHAMS UNIVERSITY
ABBASSIA , CAIRO , EGYPT

A New Font For Arabic Characters Simplifies Recognition Procedure

Salwa H. El Ramly
Electronics and Computers
Engineering Department
Faculty of Engineering
Ain Shams University
Cairo, Egypt

Mohamed A. El Hamalaway
Department of Systems and
Computers Engineering
Faculty of Engineering
Al Azhar University
Cairo, Egypt

ABSTRACT:

A new and clear shapes of Arabic letters as well as numerals and special characters are proposed. The aim seeked here is to simplify the recognition procedure, speed that procedure up and to reduce memory size. Many applications can be found for this proposal, namely in big financial establishments and in large public service organizations.

1. INTRODUCTION:

In some applications fast automatic optical character recognition is required, such as automatic document handling in banks, insurance companies and other public offices. Many problems are encountered in automatic Arabic character recognition. Cursiveness, overriding, non equal size and many types or styles of writing. These problems are complicated further if they are combined; in other words they must be taken into account all together at the same time. In Arabic, letters may have more than one shape. Actually the number of different shapes of a certain character ranges from only one shape (eg. ط ظ و ر ز) up to five shapes (eg. ـفـ ـفـ ـپـ ـپ ـف). This leads to the increase of the number of distinct characters that have to be recognised. This problem could be mostly solved at the writing stage by using an inteligent keyboard having only one shape for each letter, other shapes are deduced through software [1]. At the recognition stage a great effort must be done to discriminate between these large numbers of distinct characters. In litterature few trials are found for automatic Arabic character recognition [2,3,4,5]. The present work is a trial to overcome some of the problems mentioned above by modeling Arabic letters by new shapes. The purpose of this is to facilitate the recognition procedure by reducing the code for each of them and reducing the recognition time through proper choice of character clusters (groups).

2. NEW CHARACTER SHAPES:

In Arabic there are several styles of writing amoung which are

Koufi, Rekaa, Naskh, Diwany and others. The one letter has different shapes in each style. The differnce may be very small or large. This may lead to differnt automatic recognition procedure for each style. For time and effort saving new character shapes are proposed which have the following properties:

a. For each letter the number of different shapes kept are just those required for human eyes recognition and not for the beauty of writing (the number of different shapes is limited to a maximum of three).
b. No overriding is allowed.
c. New shapes must be as simple as possible and in the same time the essential features of habitual shapes must be kept.
d. Some modifications in shapes are necessary to obtain differeut codes for them.

The total number of characters obtaibed is 56 letter shapes in addition to 21 numerals and special signs which sum up to 77 characters. The set of 77 characters is devided into 8 clusters (groups) of different sizes:

a. Concerning horizontal spread: There are two sizes
   -size S is 4 pixels measured from right to left
   -size L is 7 pixels measured from right to left.
b. Concerning vertical spread: There are four sizes:
   -size A is 6 pixels extending from the horizontal base line up to the highest pixel.
   -size B is 8 pixels having 6 pixels above the base line and 2 below it.
   -size C is 10 pixels having 6 pixels above the base line and 4 below it.
   -size D is 12 pixels having 6 pixels above the base line and 6 pixels below it.

The letter shapes are chosen to approch the Koufi font which is the simplest font and is the oldest one; from which other fonts seem to have originated.
The different subsets are described with their horizontal and vertical widths as shown in table (1).
The difference between clusters (NL) and (AL) is that all the righthand pixels are zero.
The letter shapes in cluster (XL); while X stands for either A, B, C or D; are chosen such that the right hand side four pixels do not produce codes of letter shapes constituting cluster (XS), this is to prevent confusion at decoding stage.

3. CODING PROCEDURE:

The codes specific to each character consists of the number of horizontal crossings (comprising the vertical histogram)/the number of vertical crossings (making the horizontal histogram). In the code we count only the variations. For example the letter

(ج) from group (DL) has twelve horizontal crossings (h1, h2,.....h12) which are (010111111111) from which the parameter H is (0101). Also the vertical crossings (v1, v2,....v7) are (2334322) from which the paramerer V is (23432). This method reduces the length of the code and discards redundant information. The proposed shapes are shown in figure (1), while their codes are given in table (2).

4.RECOGNITION PROCEDURE:

Recognition is achieved through two consecutive decoding processes. The first is to find the cluster in which the character is a member. The second is to detect the character itself. Details are described below:

1.Arabic text used for automatic character recognition is supposed to be written on papers having alignment for both horizontal and vertical base lines at the right handside of the page to indicate the begining of writing. Starting from the righthand side vertical base line pixels of horizontal spread (S); i.e. 4 columns; and vertical spread (D); i.e. 12 raws; are investigated. The horizontal (h1,h2,...h12) and vertical (v1,v2,v3,v4) histograms are obtained from which (H) and (V) are calculated; and the code (H/V) is thus obtained.

By testing the horizontal histogram, the unknown letter is classified in clusters AS, BS, CS or DS (which is an empty cluster) and the code is compared to the codes in the proper cluster (XS).

If a decision is reached the next 4 pixels are investigated to find out the next character. If a decision is not reached then the next 3 pixels are considered with the previous 4 pixels from which (h1, h2, ...h12) and hence H is calculated. Also (v1, v2 ...v7) and hence V is calculated to obtain the code (H/V). The code is compared to cluster (XL) instead of (XS). When deciding whether the unknown character belongs to cluster (AL) or to cluster (NL), v1 is investigated; if it is zero then the cluster is (NL) and not (AL).

6.The characters within a cluster are devided into subgroups each having the same H (but differ in V). The subgroups are arranged from left to right in descending order with respect to their probability of occurrence [1]. Also the characters within a subgroup are arranged from left to right taking into consideration their probability distribution. The measured value of (H) for the unknown character is compared to that of the first subgroup (the most probable) then to the second subgroup,...etc. When a subgroup is identified, the value of the parameter (V) of that unknown character is used to identify it by comparing its (V) value with the (V) values of the characters comprising the subgroup from left to right until a decision is reached. By doing so the processing time is greatly reduced.

The decision tree for cluster (AS) is shown in figure (2) as an example. The recognition process starts from left to compare (H) measured with that of letters (ـجـ ، ـا) which are the most probable within the cluster and at the same time having the same (H) (H=1). If the two codes are not identical; (H) measured is compared to 0201; which is that of the letter (ـتـ) and so on. If the subgroup (ـجـ ، ـا) is recognised then the measured parameter (V) is used to diffrenciate between the different characters.

The number of subgroups ranges from 2 to 13. The number of characters per subgroup ranges from only one (leading to direct recognition) to four characters. For subgroups having only one character only H is needed. This also gives further reduction of memory size.

Actually for 42 characters out of the 77 information about the horizontal parameter H is only needed to be known. The other 35 characters need both H and V.

5. CONCLUSION:

The new proposed shapes for Arabic characters are found to be useful for office automation in places handling large amount of data e.g. in banks, insurance companies, etc... This proposal for new Arabic characters shapes will help develop fast recognition procedures for special Arabic texts and for different writing fonts which is planned for future work.

6. REFERENCES:

1. S. A. El Ramly and Mohamed A. El Hamalaway; Statistical Distribution of Arabic letters Aids to the Design of a New Keyboard; IFAC Conference on System Approach for Development; Rabat, Morocco; 24-27 Nov., 1980; pp. 515-519.
2. Mohamed A. El Hamalaway and Salwa H. El Ramly; A Novel Arabic Numerals Shapes for Optical Character Recognition; 13th International Conference on Statistics and Computer Science; Cairo, Egypt; 26-31 March, 1988; pp. 127-130.
3. M. F. Tolba and Hani M. K. Mahdi; On the Transformation of Arabic Characters into Unified Mesh Size; 13th International Conference on Statistics and Computer Science; Cairo, Egypt; 26-31 March, 1988.
4. M. A. Sharkawy, M. F. Tolba and E. Shaddad; Fourier Descriptors for Printed Arabic Character Recognition; 13th International Conference on Statistics and Computer Science; Cairo, Egypt; 26-31 March, 1988.
5. H. Y. Abdelazim and M. A. Hashish; Arabic Reading Machine; 10th. National Computer Conference; Riyad, Saudi Arabia; 1988; pp. 733-744.

Table (1) Size Definition of Clusters

Cluster Label	Horizontal Spread	Vertical Spread
AS	S	A
AL	L	A
BS	S	B
BL	L	B
CS	S	C
CL	L	C
DL	L	D
NL	L	B

Table (2) Proposed Shapes Codes

Cluster(AS)
racter Code(H/V)

Character Code(H/V)
1/01
0201/1212
10201/12
01/12
101/123
101/12
201/12
1/1
0101/121
0121/121
20121/231
101/1231

luster(AL)
acter Code(H/V)

Character Code(H/V)
02021/12121
0131/121
012021/121
01/121
0101/1321
01/0121
101/021
031/1
012031/121212
010131/1231
121/121
12121/12321
01/12121
101/121
10121/131
1/1232
121/1231
21/0121

(c)Cluster(BL)
Character Code(H/V)

	Character Code(H/V)
ب	02101/121
ج	0101/12321
ر	01/01
ز	0101/0121
ف	012121/121
ق	012121/1231
ك	20121/132
ل	121/1
ن	0121/121
هـ	12310/12121
و	012121/131
ى	012121/12321
	210/021

(d)Cluster(BS)
Character Code(H/V)

	Character Code(H/V)
ء	0101/121
ة	01/1
ئ	0102/1212
ه	010/012

(e)Cluster(CS)
Character Code(H/V)

	Character Code(H/V)
و	01/121
ر	0121/121

(f)Cluster(CL)
Character Code(H/V)

	Character Code(H/V)
ش	0421/1
ش	0120421/121

(g)Cluster(NL)
Character Code(H/V)

	Character Code(H/V)
١	1/0210
٢	21/1210
٣	31/1210
٤	1/2320
٥	01210/1210
٦	121/10
٧	21/10
٨	12/10
٩	121/012310
٠	010/010
،	01/01210
؛	1312/10
؟	313/10
ـ	12121/10
+	2121/120
×	01/10
=	0212/2120
٪	01010/20
؟	2120/2120
—	101/1230
	010/10

(h)Clust(DL)
Character Code(H/V)

	Character Code(H/V)
ع	01/12431
غ	101/12431
ج	0121/232
ح	01/232
خ	0101/23432

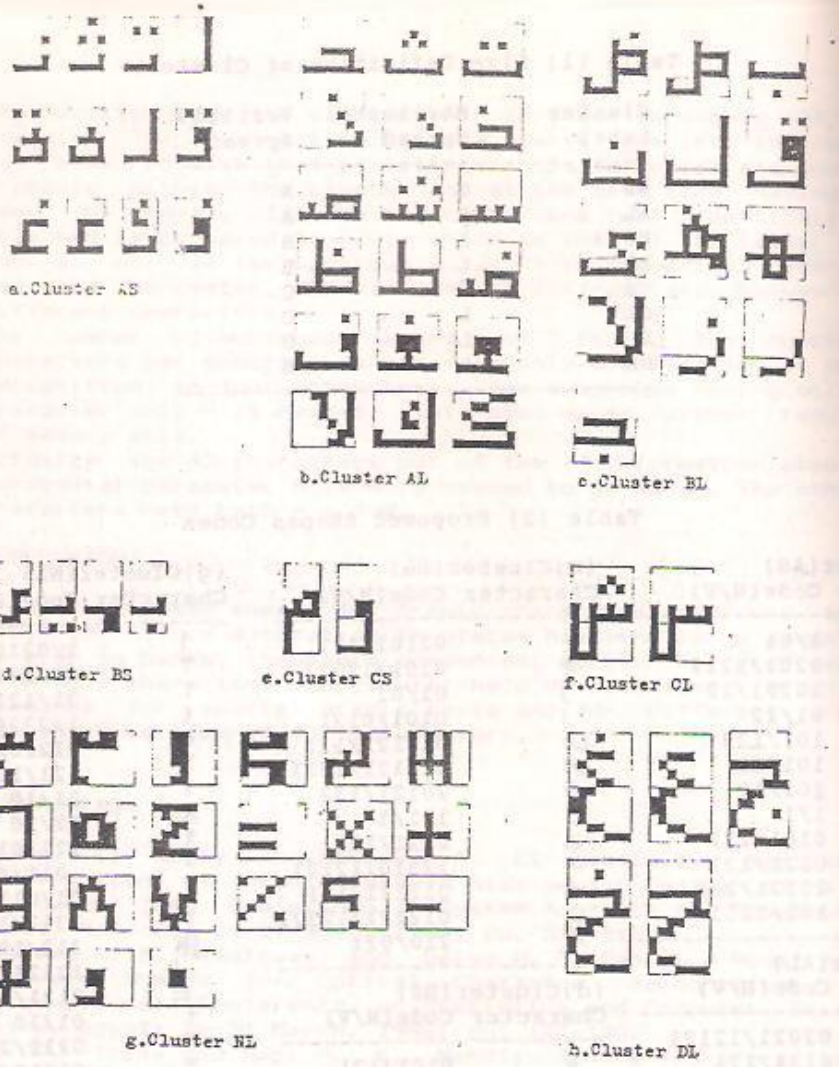


Fig.(1) Characters Shapes

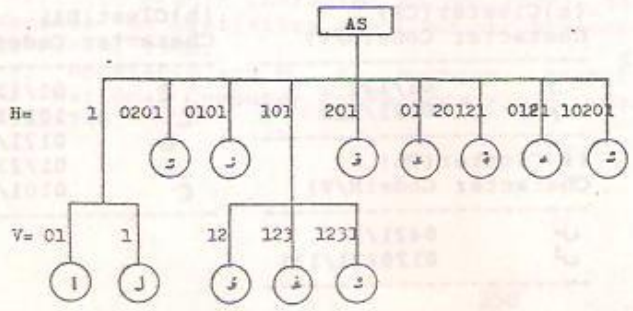


Fig.(2) Recognition Procedure for Cluster AS

شكل جديد للحروف العربيه يسهل طريقة التعرف على اشكال حروفها

سلوى ع. الرملى
قسم هندسة الالكترونيات والحاسبات
كلية الهندسة - جامعة عين شمس
القاهرة - مصر.

محمد يونس ع. الحلاوى
قسم هندسة النظم والحاسبات
كلية الهندسة - جامعة الازهر
القاهرة - مصر.

الملخص

تم اقتراح اشكال جديده وواضحه للحروف العربيه وكذلك للارقام والرموز الخاصه.

والغرض من ذلك هو تسهيل عملية التعرف عليها اليا والاسراع فى عملية التعرف تلك

وتصغير حجم الذاكره المطلوبه لذلك. هذا الاقتراح يمكن ان يكون له استخدامات عديده

فى المؤسسات الماليه الكبيره وهيئات الخدمه العامه وغيرها.

PROCEEDINGS OF THE

14th INTERNATIONAL CONFERENCE FOR STATISTICS , COMPUTER SCIENCE , SOCIAL AND DEMOGRAPHIC RESEARCH

CAIRO , EGYPT
25 - 30 MARCH 1989

VOLUME 4

COMPUTER ENGINEERING AND COMPUTER SCIENCE

SCIENTIFIC COMPUTING CENTER
AIN SHAMS UNIVERSITY
ABBASSIA , CAIRO , EGYPT

A Language Dependant Arabic Character Recognition Approach

Salwa H. El Ramly
Electronics and Computers
Engineering Department
Faculty of Engineering
Ain Shams University
Cairo, Egypt

Mohamed A. El Hamalaway
Department of Systems and
Computers Engineering
Faculty of Engineering
Al Azhar University
Cairo, Egypt

ABSTRACT:

Arabic typewritten text automatic character recognition has been considered in this paper for its relative simplicity. The analysis of human inherent rules of character recognition has led to successive clustering procedures. the approach is language dependant based upon the feachures of Arabic characters; and has taken into account the following parameters: the place of the unknown character, the presence or absence of diacritics, the place of diacritics with respect to the unknown character and their number, the place of the character with respect to the horizontal base line & its length and the horizontal and vertical crossings characteristics. The method seems to be promising for application on other styles of Arabic writings.

1. INTRODUCTION:

Arabic writing is different from English writing in many aspects among which we state: cursiveness, overrriding and varios shapes for a given single Arabic letter depending upon its place in the word. Different styles of writing (e.g. Koufi, Rekaa, Naskh,...) give different shapes for Arabic characters. For automatic Arabic character recognition one has thus to specify the style of writing he is going to treat. Most of the published work are for typewritten text and discusses elaborate procedures for recognition [1,2]. It seems however that some kind of simplification of the recognition process can be achieved if one can imitate the actual (human) laws of recognition where no complex mathematical calculations is performed. The human system relies upon the recognition of some features specific to each character and not shared with any other character. The group of subsets describing features for different characters are defined for a certain writing style and may be different for the groups of other styles. This statement can be attested when it is noticed that the human system can read many and many different hand written texts when the hand writting may be highly deformed from the correct writing. This is because the writter is actually

trying to keep most of the main features when writing dropp the complementary features that are for more clarity or beauty the written text. The present work is trying to imitate the hum system way of recognition. The subject is being practica difficult and necessitates long studies. We have devided the wo to be covered into some topics. The first topic was a study the most common apparant features of different characters different styles of writing. Out of these features; as a firs application; a new proposed shapes for Arabic letters, numer characters and special signs were proposed in two successiv papers [3,4]. The ultimate aim of that proposal is to define system using a simple process where only counting horizontal vertical crossings of the characters give mutually exclusive se to be used for recognition. Doing so; time and memory s savings are gained for office automation in places having to de with large amount of data.

In the present paper a further step in this research work exposed. The study presented is confined to typewritten te which is apparantly simpler. Other types of writing are under investigation.

2.LAWS OF THE HUMAN SYSTEM FOR ARABIC CHARACTER RECOGNITION:

The human system for reading Arabic text which has a topologic nature is supposed to follow the following steps:

1. The full text is segmented nto distinct horizontal lines. Th horizontal base line is defined as the place where most of th letters are attached together for cursiveness. Concatinating an Arabic letter to another (for letters which can be connected fro the left side) one has to pass by the horizontal base line. Al most of the Arabic letters are drawn above and including th horizontal base line, while some have parts below the line.
2. The written line is then segmented into distinct words. Th segmentation problem is being dealt with in other work [1,2].
3. The words are further segmented into distinct pieces of Arab words PAW [1]. A PAW can contain only one character or more.
4. The PAWs are finally segmented into distinct letters.
5. Recognition of different letters is achieved by searchi their specific features and comparing them to the memorize features for recognition.

3.PROPOSED CHARACTER RECOGNITION SYSTEM:

Based on the study of human law for recognizing Arab characters; the following set of system design principals has to be followed:

a. Adopting structural analysis recognition method and selecting stable features.
b. Using the fuzzy pattern recognition theory to extract characters features based upon prior knowledge of Arabic characters structure.
c. Adopting layer structure multistage recognition.
The block diagram of the recognition system is shown in figure (1).

3.1.Text preprocessing for automatic optical reading:

The first step is to make alignment to determine the horizontal base line. In second step segmentation is achieved through projection of pixels of a whole written line on a vertical line. Thresholds must be defined during the learning step where the thickness of the writing line Δ is determined. Actually in this step both word and PAW segmentations are achieved too. Spaces between PAWs of value S1 and spaces between words of value S2 are also specified. Actually S2 > S1 and they are both random variables. Statistical measurement of S1 and S2 indicate that they are Gaussian with mean values $\overline{S1}$ & $\overline{S2}$ and variances $\sigma S1$ & $\sigma S2$. Figure 2 shows the defined quantities Δ, S1, S2.

3.2. Clustering and Identification:

Recognition is achieved by successive clustering procedures. Such procedures are:

a. Clustering according to the place of the unknown character C' in the word. The number of these clusters are five as described in table (1). It is to be noted that the five sets of characters are not disjoint.
b. Clustering according to the presence or absence of diacritics.
c. Clustering according to the place of diacritics (above or below the letter).
d. Clustering according to the number of diacritics and their kind (one, two, three dots or hamza).
e. Clustering according to whether the whole character is above the horizontal line or having a part of it below the line.
f. Clustering according to the length of the part of the character above the base line (full length or half length). Also letters partly below the line are devided into two clusters: long and short.
g. Clustering according to the horizontal and vertical crossings.

Two notes have to be taken into account:

1. Clustering procedurs from step (b) to step (g) are not done in the same sequence for all the five clusters defined in step (a).
2. Step (g) which leads to the final recognition of the character is quite dependant upon the group of characters sharing the same

labels in clustering steps (a) to (g). In fact the extracted features are optimally shifted for the reduction of redundant data.

The designed classifier is of a multi-cross layer tree structure [5]. In case of failure of fuzzy recognition procedure in one cluster step, one has to return to probabilistic correlation technique; in which the unknown pattern is matched with the reference patterns till a maximum correlation is attained.

An example of the proposed system is shown in fugure (3).

4.CONCLUSION:

In the present paper a system imitating the laws of human recognition procedures are described for typewritten texts. The merit of this method is that it can be readapted for new sitiuations where more characters are involved (e.g. numeric and special signs). The same algorithm can be followed for other types of writing which is our next research activity.

5.REFERENCES:

1. H. Y. Abdelazim and M. A. Hashish; Arabic Reading Machine; 10th. National Computer Conference;Riyad, Saudi Arabia; 1988; pp. 733-744.
2. M. A. Sharkawy, M. F. Tolba and E. Shaddad; Fourier Descriptors for Printed Arabic Character Recognition; 13th International Conference on Statistics and Computer Science; Cairo, Egypt; 26-31 March, 1988.
3. Mohamed A. El Hamalaway and Salwa H. El Ramly; A Novel Arabic Numerals Shapes for Optical Character Recognition; 13th International Conference on Statistics and Computer Science; Cairo, Egypt; 26-31 March, 1988; pp. 127-130.
4. Salwa H. El Ramly and Mohamed A. El Hamalaway; A New Font for Arabic Characters Simplifies Recognition Procedure; to be published.
5. Duda & Hart; Pattern Classification and Scene Analysis; John Wiley & Sons; New York; 1973.

Table (1)

luster abel	Place of Unknown Character	Description
1	S2 C' S2	Isolated character
2	S2 C' C	End character of a word
3	S1 C' C	End character of a PAW
4	C C' S1 or C C' S2	Start character of a word or PAW
5	C C' C	Middle character

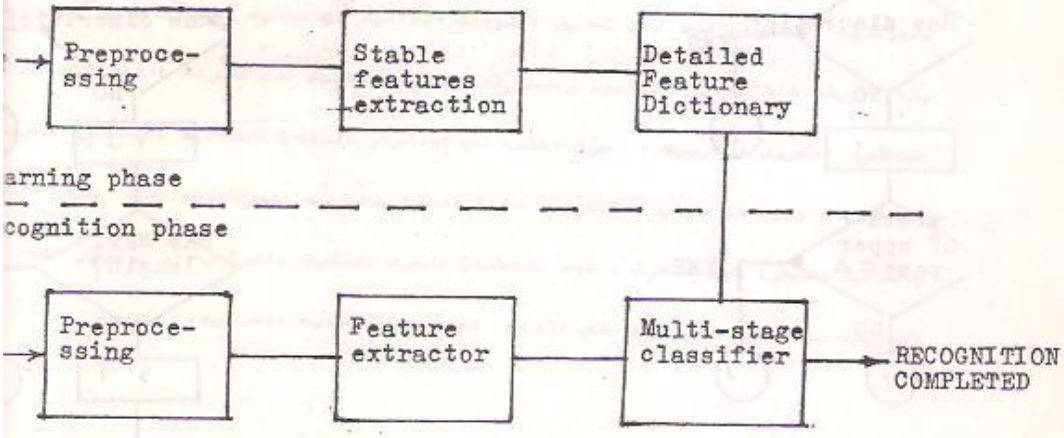


Fig.(1) Block diagram of the recognition system.

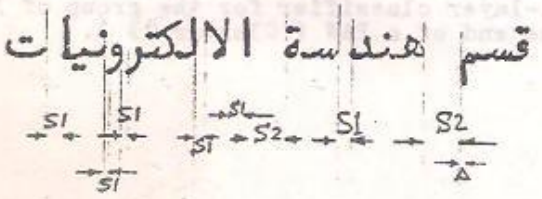


Fig.(2) Definition of S1, S2, Δ

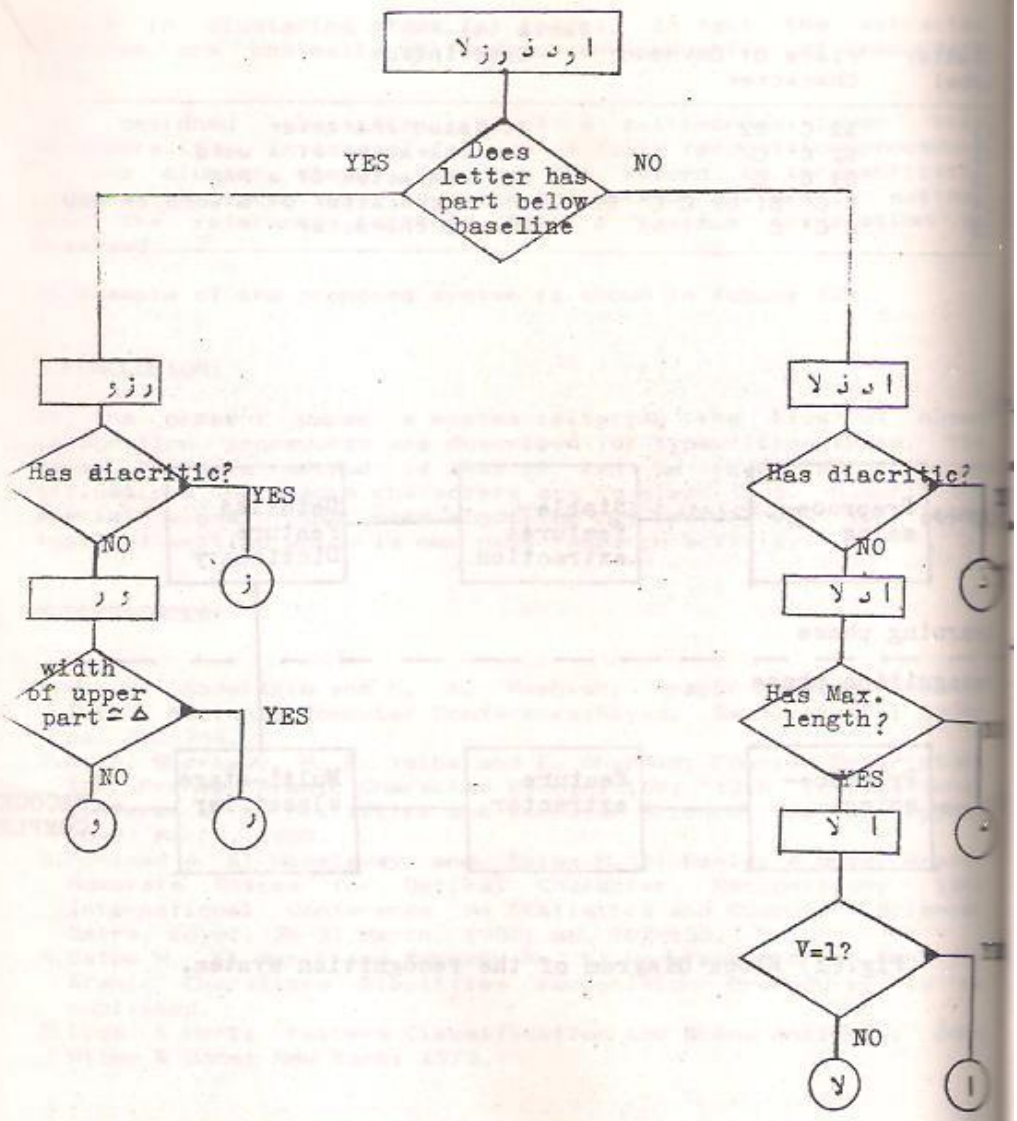


Fig.(3) Multi-layer classifier for the group of letters at the end of a PAW (Cluster C3).

التعرف على الحروف العربيه بطريقه مستمده من اللغه ذاتها

سلوى ع. الرملى
قسم هندسة الالكترونيات والحاسبات
كلية الهندسة - جامعة عين شمس
القاهرة - مصر.

محمد يونس ع. الحلاوى
قسم هندسة النظم والحاسبات
كلية الهندسة - جامعة الازهر
القاهرة - مصر.

الملخص

يختص هذا البحث بالتعرف الالى على النصوص العربيه المكتوبه على الاله الكاتبه نظ
لسهولتها النسبيه. لقد ادى تحليل القواعد الذاتيه الانسانيه للتعرف على الحروف الـ
بناء طريقه ذات تجميع متتابع. والطريقه المتبعه تعتمد على طبيعة اللغه وتستند الـ
خصائص الحروف العربيه ، حيث اخذنا فى الاعتبار النقاط التاليه : مكان الحرف
المجهول ، وجود او غياب التنقيط ، مكان النقط بالنسبه للحرف وعددها، مكان الحر
بالنسبه للسطر، طول الحرف ، عدد تقاطعات الحرف الافقيه والرأسيه . وتبشر هذه
الطريقه بامكانية تطبيقها على انماط اخرى من الكتابه العربيه.

المؤتمر الأول لهندسة اللغة

تنظمه
الجمعية المصرية لهندسة اللغة

تحت رعاية

الأستاذ الدكتور / حسن احمد غلاب

رئيس جامعة عين شمس

١٤ - ١٥ مارس ١٩٩٨

القاهرة - جمهورية مصر العربية

الأبحاث المقبولة

المؤتمر الأول
لهندسة اللغة
١٤-١٥ مارس ١٩٩٨م

منهج لتعريب لغات البرمجة: لغة اللوجو كمثال

محمد يسرى النحاس محمد يونس الحملاوى

نبذة:

فى منظومة مكونات الحاسبات المتقدمة نجد أن تعقيد تلك النظم يتناسب عكسياً مع سهولة استعمال الحاسوب. وينطبق هذا على اللغات الطبيعية وكذلك على مستوى التطبيقات على الحاسوب. فالأصل فى ميكنة أية منظومة هو المستخدم والتسهيل عليه حتى يستطيع أن يبدع من خلال تلك الإمكانيات. وما يحدث من عدم اكتمال هذا الهدف يرجع إلى مستوى التقنية المستخدمة وكلفتها. ويجيئ استعمال اللغة العربية كلغة برمجة متسقاً مع هذا الهدف. ويعرض البحث نموذجاً لبناء لغة لوجو عربية من خلال استعمال البرمجة المرئية لمحاكاة أوامر تلك اللغة بألفاظ ورموز عربية تمكن المستخدم من التعامل مع الحاسوب باللغة العربية بصورة كاملة. ولقد اختيرت تلك اللغة لانتشارها فى المدارس وتتميز بتنمية القدرات الرياضياتية والمنطقية عند الطلبة. وتدلل تلك الخطوة على سهولة هذا الاتجاه على أمل أن تتبعها خطوات أخرى للغات أخرى.

قسم هندسة النظم والحاسبات، كلية الهندسة، جامعة الأزهر.

منهج لتعريب لغات البرمجة: لغة اللوجو كمثال

محمد يسرى النحاس محمد يونس الحملاوى
قسم هندسة النظم والحاسبات، كلية الهندسة، جامعة الأزهر

١ – مقدمة:

تنتشر على الكثير من الحاسبات الموجودة فى المدارس لغة اللوجو وتعتمد وزارة التربية والتعليم فى مصر تلك اللغة كأداة مشوقة لتقدم للطلبة آليات استعمال الحاسوب.[١] ولكن الواقع يشير إلى أن أغلب الطلبة يعزفون عن تعلم تلك اللغة لعدة أسباب أهمها عدم الشعور بالألفة مع رموز تلك اللغة. وهذا الأمر يرتبط أيما ارتباط بقضية تعريب الرموز العربية التى استشرت فى وقتنا الراهن.[٢] ما يهمنا فى هذا الأمر أن ذلك الحائل الذى يفصل بين الطلبة العرب لغة يؤدى فى الأغلب الأعم إلى أن تصبح الألعاب هى كل ما يرتبط بالحاسب، وبالتالى نهدر إمكانية ونضيع فرص تعليم الطلبة خاصة فى السن الصغير لغة برمجة محببة إلى أغلب الأطفال.

ولا يخفى علينا أن تحول الحاسوب إلى ألعاب فقط لهو مضيعة للطاقات وليس اضافة للمتعلم. ولسنا فى مجال التدليل على أن ممارسة لغة للحاسب سوف يضفى على فكر النشأ ليس فقط الصغير بل والكبير كذلك التسلسل المنطقى فى تعامله مع الحاسوب. وهذا الأمر سوف ينطبع على المتعلم فى العديد من ممارسته بل وسلوكياته. وهذا هو عين ما استفادته البلاد المتقدمة من خلال نشر المعرفة بالحاسوب وبأساليب برمجته لتضفى على النشأ عندها اعتياد التفكير المنطقى وتقبل الخطوات الاجرائية فى العديد من السلوكيات بل والتفكير.

كما أن لتعريب الحاسوب وبرمجياته انعكاساته على تشجيع المتعلم على اعتياد الحاسوب ليسهل استعماله لحل الكثير من الصعوبات التى تعترضنا عامة. وواضح كذلك أن فى استعمال اللغة الأم الكثير من التيسير على المتعلم وهو عين الهدف من استعمال اللغات عالية المستوى مثل اللوجو والفورتران بدلاً من اللغات قليلة المستوى مثل لغات التجميع. كما أن تعليم غير المتخصص لغات الحاسب المتقدمة لابد من أن يصاحبه التيسير عله بمخاطبته بلغته. ليس المقصود من تعليمنا لغة للحاسوب أن نثير فى ذهن المتلقى سواء أكان طفلاً أم غيره الرهبة لجهاز الحاسوب بل إن العكس لهو المطلوب. فتيسير التعلم لهو

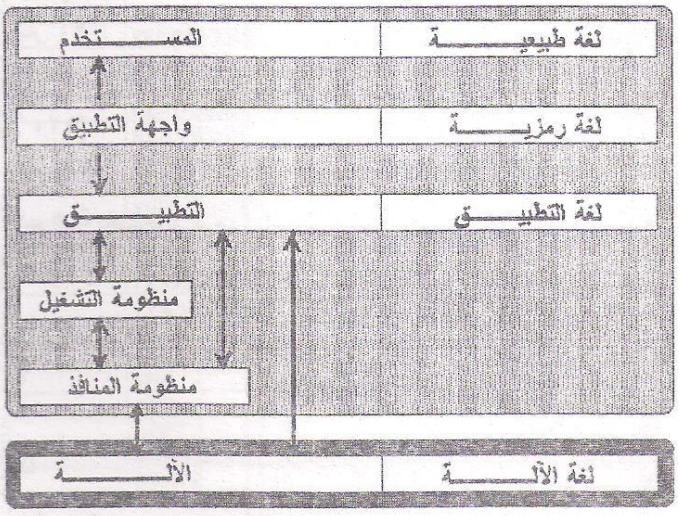
الهدف. ولا يخفى علينا مقدار الاستفادة من اعتياد نمط التعامل مع الحاسوب فى تقييس نمط تفكيرنا وتيسير تعاملنا مع مختلف المعارف.

كما أن لغة اللوجو تمكن المتعلم من محاطبة الآلة وتعليمها الكلمات والجمل مما يؤدى إلى تفهمه لبراعة ودقة تراكيب لغته الأم. ولغة اللوجو كذلك أداة طيعة لتعلم مفاهيم الرياضيات الحاسوبية. فهذه اللغة مدخل طبيعى لادراك المفاهيم الآساسية للرسم بالحاسوب وعالم بناء الصور بتكرار العناصر المرئية الصغيرة (الفراكتال). كما أنها بحكم نشأتها أكثر ملاءمة لتصوير المبادئ العلمية المستخدمة فى التحكم فى الانسان الآلى. وبهذه الأدوات مجتمعة تعتبر لغة اللوجو أفضل أدوات علم النفس لدراسة وتطوير ملكة الابداع عند الأطفال.

٢- منهج تعريب لغة البرمجة:

توجد العديد من الأساليب كى نعرب استعمالنا مع الحاسوب. أول تلك الأساليب هو تعريب واجهات حزم البرامج مع الابقاء على البرنامج كوظائف وكتصميم بدون تغيير. وثانى تلك الأساليب هو تعريب كلمات اللغة التى من خلالها نتعامل مع البرنامج. وثالث تلك الأساليب هو تعريب العبارات والتى تعنى محاكاة وظائف اللغة الأصلية باستخدام عبارات عربية مماثلة ومقيدة ودالة على تلك الوظائف. ورابع تلك الأساليب هو وضع لغة برمجة عربية جديدة تنافس لغة أخرى أو مجموعة من اللغات الأخرى. ونلاحظ أن سهولة بناء الواجهة يتناسب عكسياً مع عمقها. ويتناسب العمق مع كفاءة العمل كذلك.

ويمكن توضيح علاقة هذه المناهج المختلفة ببيئة الحاسوب بالشكل رقم ١ التالى:



شكل رقم ١ البيئة الداخلية لتطبيقات الحاسوب

ونلاحظ أننا كلما اقتربنا من الآلة فى الشكل السابق كلما تعمقنا فى مستوى تعاملنا مع الآلة. وتندرج التطبيقات العربية لحزم البرامج تحت تلك المستويات الأربع وإن زاد عددها كلما بعدنا عن الآلة.

٣- تصميم لغة اللوجو العربى:

شهدت لغات البرمجة العديد من المداخل السابقة منذ بدأ استخدام الحواسب فى المنطقة العربية. وبدأت محاولات تعريب تلك اللغات بالعديد من الخطوات تناسبت مع مستوى التقانة المتوفرة وقت التطبيق. وتعددت تلك المحاولات فى فترات بعينها وخبت فى فترات أخرى. وتواكبت تلك المحاولات مع الرغبة فى تحسين مستوى أداء منظومة العمل والتعليم.

وفى التطبيق الحالى للغة اللوجو العربى استخدمنا الاسلوب الثالث من أساليب تعريب استعمالات الحاسوب بمحاكاة وظائف لغة لوجو الأصلية باستخدام عبارات عربية مماثلة ومقيدة ودالة على تلك الوظائف. وقد اخترنا مجموعة جزئية من أوامر لغة اللوجو التقليدية تشتمل على المجموعة الجزئية التى يتم تدريسها فى المدارس المصرية. ويوضح الجدول رقم ١ مجموعة الأوامر التى تم تعريبها. ونلاحظ أننا لم نجر أية اختصار على تراكيب التحكم فى التدفق لتسهيل فهم البرنامج. والمثال التالى يوضح اجرائية بسيطة لرسم مربع بداخله مثلث:

```
إلى مربع
كرر ٤[أمام ٨٠ يمين ٩٠]
يمين ٩٠ أمام ٢٠ يسار ٩٠
رفع_قلم أمام ١٠ يمين ٣٠ ضع_قلم
كرر ٣[أمام ٥٠ يمين ١٢٠]
نهاية
```

ويمكن كتابة نفس الاجرائية باستخدام الرموز المختصرة كما فى الجدول رقم ١ لتصبح كالآتى.

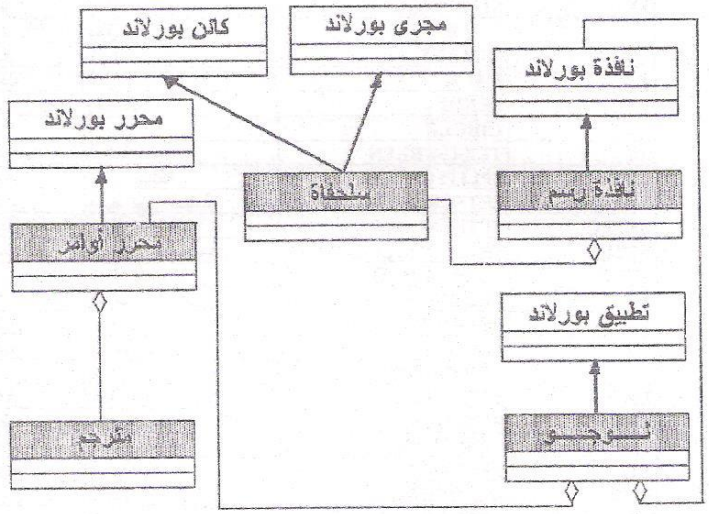
```
إلى مربع
كرر ٤[أم ٨٠ يم ٩٠]
يم ٩٠ أم ٢٠ يس ٩٠
رق أم ١٠ يم ٣٠ ضق
كرر ٣[أم ٥٠ يم ١٢٠]
نهاية
```

قائمة الاختصارات

Abbreviation	Command	الأمر	الاختصار
	LAST	أخير	أخ
<	LESSP	أقل من	
>	GREATERP	أكبر من	
	TO	إلى	
FD	FORWARD	أمام	أم
	OR	أو	
HT	HIDETURTLE	اخفاء سلحفاة	خس
	IF	إذا	
CO	CONTINUE	استمر	سر
ST	SHOWTURTLE	أظهر سلحفاة	ظس
	HOME	بيات	
ED	EDIT	تحرير	حر
	SAVE	تخزين	
PPT	PENPAINTt	تلوين قلم	لتق
+	SUM	جمع	
SE	SENTENCE	جملة	جم
	LOAD	حمل	حم
OP	OUTPUT	خرج	خج
BK	BACK	خلف	خف
PU	PENUP	رفع قلم	رق
CRC	CIRCLE	دائرة	دا
TS	TEXTSCREEN	شاشة نص	شنص
SS	SPLITSCREEN	شطر شاشة	شطر
SETSC	SETSCREENCOLOR	ضبط لون شاشة	ضلق
SETPC	SETPENCOLOR	ضبط لون قلم	ضلق
SETFC	SETFLOODCOLOR	ضبط لون مساحة	ضلم
SETH	SETHEADING	ضبط اتجاه	
*	PRODUCT	ضرب	*
PD	PENDOWN	ضع قلم	ضق
PR	PRINT	طباعة	طبع
-	DIFFERENCE	طرح	-
BL	BUTLAST	عدا آخر	عخ
BF	BUTFIRST	عدا أول	عل
PX	PENREVERSE	عكس قلم	عكس
RL	READLIST	قراءة قائمة	قرقم
RC	READCHAR	قراءة حرف	قرح
RCS	READCHARS	قراءة حروف	قرحو
RW	READWORD	قراءة كلمة	قرك
/	QUOTIENT	قسمة	/
	REPEAT	كرر	
FS	FULLSCREEN	كل شاشة	كلش
	NOT	لا	
ER	ERASE	محو	مح
PE	PENERASE	محو قلم	محق
CS	CLEARSCREAN	مسح شاشة	مسش
ERF	ERASEFILE	محو ملف	محم
CT	CLEARTEXT	مسح نص	مسنص
	END	نهاية	
	AND	واو	
LT	LEFT	يسار	يس
=	EQUALP	يساوى	=
RT	RIGHT	يمين	يم

٤- تصميم مترجم اللوجو العربى:

صمم مترجم لغة اللوجو العربى وبيئة عمله باستخدام منهج تصميم الكائنات ونفذ باستخدام لغة بورلاند سى++ المرئية وذلك لوضوح هيكل التصميم وسرعة تنفيذ المترجم. وكذلك استخدمت فى بداية التجارب لغة البيسيك المرئية لتنفيذ المترجم لسهولة استخدامه فى بناء النموذج الأولى. ويضح الشكل رقم ٢ مخطط أصناف بيئة اللوجو العربى، حيث مربعات الأصناف الرمادية فى الرسم تمثل الأصناف التى صممت ونفذت بواسطة الفريق البحثى أما مربعات الأصناف الأخرى فتمثل الأصناف التى ورثت منها والمتاحة فى بيئة الـ سى++٢,٤. ويتمحور العمل كله حول تصميم المترجم الذى يعمل بطريقة التحليل علوى-سفلى. [٣][٤]



شكل رقم ٢ مخطط أصناف بيئة اللوجو العربى

٥- الملخص:

يتسق هدف البحث مع منظومة تعريب العلوم بهدف رفع كفاءة العملية التعليمية لنصل بالمتلقى لمرحلة الفهم العميق لأساسيات ما يدرس، ولن يتم ذلك بدون اللغة الأم. ويعرض البحث نموذجاً لبناء لغة لوجو عربية من خلال استعمال البرمجة المرئية لمحاكاة أوامر تلك اللغة بألفاظ ورموز عربية تمكن المستخدم من التعامل مع الحاسوب باللغة العربية بصورة

كاملة. كما يعرض البحث المجموعة الجزئية للأوامر التى تم تنفيذها والمنهج الذى اتبع فى بناء مترجم لغة اللوجو العربى.

٦- المراجع:

١- فوقية رشوان الزهيري وآخرون؛ الحاسب الالكترونى للصف الأول الثانوى؛ وزارة التربية والتعليم؛ القاهرة؛ دار النشر هاتييه؛ ١٩٩٤م.

٢- محمد يونس الحملاوى ومحمد يسرى النحاس؛ الرموز وقضية التعريب؛ ندوة الرموز ومكانتها فى قضية التعريب؛ القاهرة؛ الجمعية المصرية لتعريب العلوم؛ ٢٧ فبراير ١٩٩٧م.

3- Alfred V. Aho & Jeffrey D. Ullman; Principles of Compiler Design; Addison-Wesley Publishing Co.; Massachusetts, U.S.A.; 1977.

4- Robert Sedgewick; Algorithms in C; Addison-Wesley Publishing Co.; Massachusetts, U.S.A.; 1990.

LOGO5.DOC

بسم الله الرحمن الرحيم

التحديات التقنية التى تواجه اللغة العربية

أ.د. محمد يونس عبد السميع الحملاوى
أستاذ هندسة الحاسبات، كلية الهندسة، جامعة الأزهر

قد يكون من الأصوب أن يكون العنوان: التحديات التقنية التى تواجه أهل العربية، فاللغة العربية لغة مكتملة الأركان بلا خلل فى تراكيبها صوتًا وكتابةً. ورغم ذلك فلم تحظْ الدراسات المقارنة لأصوات وحروف اللغة العربية واللغات الأخرى باهتمام معقول من أهل العربية. [١-٢] كما لم يحظ إسهام اللغة العربية فى العلم العالمى بالقدر المناسب من الدراسات.[٣] ورغم أهمية هذه الدراسات لإذكاء روح الولاء فى أبناء العربية إلا أن الأمر لم يجد الاهتمام المناسب به على مختلف الأصعدة العربية!

من التحديات التقنية التى تواجه أبناء العربية نقل العلم للعربية مما يستلزم الاهتمام بالترجمة الآلية من لغات الدول المتقدمة إلى لغتنا العربية. ويمكن من رصد الحالة العلمية فى المنطقة أن نستشعر تصحر المخرج العلمى العربى من ناحية الجودة ومن ناحية الهدف فغالبًا ما يصب جهدنا البحثى الجيد فى غير صالح المنطقة وبالتالى فى غير صالح لغتنا العربية. ولم تحظْ برمجيات الترجمة الآلية بالاهتمام المؤسسى ولهذا لم تتكامل الجهود فى هذا الإطار.[٤-٥] لقد لعب التقييس

---

١ محمد يونس الحملاوى؛ دراسة مقارنة بين أشكال الحروف العربية والحروف الإنجليزية؛ المؤتمر الدولى عن العربية وتقنية المعلومات، المجلس الأعلى للغة العربية، الجزائر؛ ٢٨-٢٩ ديسمبر ٢٠٠٢م

٢محمد يونس الحملاوى ومحمد يسرى النحاس؛ تقييم الخواص الشكلية لفئتى الأرقام العربية المشرقية والغبارية الغربية؛ المؤتمر الثالث لهندسة اللغة؛ القاهرة؛ ٢٢ أكتوبر ٢٠٠٢م

٣ Neil deGrasse Tyson, Natural History Magazine, Research Triangle Park, NC, USA; February 2003

٤ محمد يونس الحملاوى؛ نحو رؤية لدور الترجمة فى منظومة النهضة العلمية؛ المؤتمر الدولى للترجمة ودورها فى تفاعل الحضارات؛ القاهرة؛ ٢٣-٢٥ يونيو ١٩٩٨م

٥ محمد يونس الحملاوى وآخرون؛ بحث مشترك مع آخرين؛ إنشاء محلل كلمات ومعجم آلى فى مجال الترجمة الآلية من الإنجليزية إلى العربية؛ المؤتمر السنوى الثالث لتعريب العلوم؛ القاهرة؛ ١٢-١٣ مارس ١٩٩٧م

دورًا محوريًا فى تقدم العلوم فى العصر الحالى ولم تحظ اللغة العربية بالجهد الذى تستحقه فى هذا الصدد، فما تم هو جهود فردية وإن كُتِبَ لها النجاح إلا أنها لم تتكامل مع غيرها من الجهود.[٦-٧] ورغم غزارة المصطلحات التى تم اعتمادها من مختلف الهيئات اللغوية العربية وغير العربية إلا أن حوسبة تلك المصطلحات لتصب فى آلية عمل عربية لم تحظْ بالاهتمام الواجب. [٨] كما أن المعاجم الآلية لم تنل جهدًا عربياً كافيًا نتيجة عدم وجود مُنتَج تصب فيه تلك المعاجم.[٩-١٠] ومن المجالات التى لم تنل ما تستحقه من جهد مؤسسى تعريب أسماء مواقع شبكة الإنترنت وعناوين البريد الإلكترونى حيث أُنجِز توصيف قياسى لها ولكن لم تلتفت مختلف المؤسسات لهذا التوصيف وبالتالى بات حبرًا على ورق.[١١-١٢] كما لم تحظْ قضية التعرف على الحروف العربية المطبوعة ما تستحقه من اهتمام وكأننا تركنا المجال لعلماء الغرب.[١٣] كما لم نبذل نبذل أى جهد

٦ محمد يونس الحملاوى وآخرون؛ مسودة مشروع مواصفة الأرقام العربية والعلامات الحسابية الأساسية؛ ندوة الأرقام العربية والمواصفات القياسية؛ الهيئة المصرية العامة للتوحيد القياسى وجودة الإنتاج؛ القاهرة؛ ٢٠ مايو ٢٠٠٣م

٧ محمد يونس الحملاوى؛ التقييس والبرمجيات العربية؛ ندوة صناعة البرمجيات فى مصر؛ جمعية المهندسين المصرية؛ القاهرة؛ ٢١ سبتمبر ٢٠٠٣م

٨ محمد يونس الحملاوى؛ المصطلح العلمى العربى والحوسبة؛ الندوة الخامسة للمسئولين عن تعريب التعليم العالى فى الوطن العربى؛ الخرطوم؛ ٢٨-٣٠ نوفمبر ٢٠٠٤م

٩ محمد يونس الحملاوى وآخرون؛ ملاحظات حول الذخيرة اللغوية العربية وحوسبتها؛ ندوة حوسبة بنوك المعطيات النصية مع التطبيق على الذخيرة اللغوية العربية؛ الجزائر؛ ٣-٤ نوفمبر ٢٠٠١م

١٠ محمد يونس الحملاوى؛ المعاجم المُحَوْسَبَة والتعريب؛ ندوة المسئولين عن تعريب التعليم العالى فى الوطن العربى، مسقط، ٤-٦ نوفمبر ٢٠٠٦م

١١ محمد يونس الحملاوى وآخرون؛ مشروع مواصفة تعريب أسماء مواقع الشبكة العالمية للمعلومات؛ ندوة تعريب أسماء مواقع شبكة المعلومات العالمية؛ الهيئة المصرية العامة للمواصفات والجودة، القاهرة؛ ٤ مايو ٢٠٠٥م

١٢ محمد يونس الحملاوى؛ أسماء مواقع الإنترنت ومجموعة اللغات العربية؛ مؤتمر ائتلاف أسماء مواقع الإنترنت باللغات المختلفة؛ تونس؛ ٢٥-٢٦ أكتوبر ٢٠٠٣م

١٣ محمد يونس الحملاوى وآخرون؛ التعرف على حروف اللغة العربية باستخدام خوارزمات التنحيف والتقطيع؛ المؤتمر الخامس لهندسة اللغة؛ كلية الهندسة، جامعة عين شمس؛ القاهرة؛ ١٤-١٥ سبتمبر ٢٠٠٥م

مؤسسى فى مجال لغات البرمجة.[١٤] وفى نفس الوقت لم نهتم بقضية التشكيل الآلى للنصوص العربية.[١٥]

ولعل مجال استنطاق الآلة والإملاء الآلى والتعرف على الكلام يحظى بقدر من اهتمام المؤسسات العربية فلقد أصبح المترجم الآلى للكلمات المنطوقة والمذيع الآلى وأصبحت آلات الرد الآلى وآلات المحادثة الآلية فى اللغات الأخرى قاب قوسين أو أدنى من الاستخدام التجارى.[١٦] ورغم ما يحدث من اضطراب فى مجال رسم الحرف العربى إلا أننا لم نهتم بالقدر الكافى لهذه القضية فانتشرت الأبناط غير المقيسة للحرف العربى فى الاستخدامات الالكترونية وبالتالي انتقلت إلى الصحف والافتات.[١٧] هذا الأمر سوف يؤئد السليقة اللغوية ليس فقط للحرف المكتوب بل كذلك للحرف المسموع حيث نجد أن التوليد الآلى لبعض كلمات اللغة فى بعض أشكال الحرف العربى يعطى كلمات عربية خاطئة.

ولعل الجانب المهم فى التحديات التى تواجهنا كعرب هو الحفاظ على هويتنا، فعلينا أن ننظر للغة بصورة موسوعية، فبجانب العديد من التعريفات للغة هناك وظيفة أساسية لها هى نقل المعارف بين أفراد المجتمع والحصول عليها من المجتمعات الأخرى عبر آلية الترجمة وكذلك البناء على الكيان المجتمعى الممتد لقرون عدة ومنه التراث العلمى، فالوعاء المجتمعى لأية أمة يتشكل داخل لغتها. اللغة؛ أية لغة؛ تحمل المضامين الثقافية للمجتمع الذى تحيا فيه وبه ويحيا هو بها. ورغم ذلك تتمايز اللغات فيما بينها فمنها ما هو مكتمل البنيان ومنها ما يقصر عن التعبير عن بعض الأزمنة على سبيل المثال فالإنجليزية واليابانية مثلًا لا تحتويان على صيغة للفعل فى الزمن المستقبل بل يلزم إضافة بعض الكلمات للتعبير عن المعنى فى اللغتين. كما أن الفعل فى اليابانية لا يرتبط بالفاعل. هذه الاختلافات هى فى حقيقتها مؤشرات تعبر بجانب عدد الكلمات وعدد جذور اللغة عن درجة ثراء تلك اللغة وعن طريقة تفكير أهل تلك اللغة فى نفس الوقت. كما أن اللغة

---

١٤ بحث مشترك مع محمد يسرى النحاس؛ منهج لتعريب لغات البرمجة: لغة اللوجو كمثال؛ المؤتمر الأول لهندسة اللغة؛ القاهرة؛ ١٤-١٥ مارس ١٩٩٨م

١٥ Rehab Alnefaiea , Aqil M. Azmi; Automatic minimal diacritization of Arabic texts; 3rd International Conference on Arabic Computational Linguistics; Dubai, United Arab Emirates; ACLing 2017, 5–6 November 2017

١٦ Hassan Satori, H Harti & N. Chenfour; Arabic Speech Recognition System Based on CMUSphinx, 3rd International Symposium on Computational Intelligence and Intelligent Informatics – ISCIII 2007; Agadir, Morocco; March 28-30, 2007

١٧ محمد يونس الحملاوى ومارك فان وورمهاوت؛ ملاحظات حول طباعة وإظهار بعض الحروف غير اللاتينية؛ مؤتمر الاتحاد الدولى للتحكم الآلى؛ القاهرة؛ ٢٦-٢٩ نوفمبر ١٩٧٧م

وعاء حضارة المجتمع فمن خلالها يتواصل الفرد مع غيره ويتواصل مع تاريخه ويتواصل مع مستقبله من خلال ما يتركه من آليات للحضارة. فاللغة ليست فقط أداة تواصل بين أفراد المجتمع بل هى بالأساس مضامين حضارية والتى بدورها تشمل مفاهيم ثقافية. فاللغة هى الوعاء الأكبر لأى مجتمع. وكلما كانت اللغة نقية كلما كانت فرصة نهوض مجتمعها أفضل، فهى ليست العامل الوحيد ولكنها عامل مهم فى عملية التنمية. كما أن للغة دور مهم فى مسألة القومية وصهر المجتمع سياسياً وفكرياً فى بوتقة واحدة ويجدر بنا أن نتذكر دور المفكر الألمانى هردر فى القرن الثامن عشر الميلادى فى هذا السياق الذى نظَّر لوحدة الأقاليم الألمانية المتناحرة مما عَبَّد الطريق أمام بسمارك لتوحيد ألمانيا.[١٨] كما يجدر بنا أن نشير إلى أن اللغة وعاء الفكر حيث تؤكد العديد من الدراسات إلى أن للغة حضور أساسى فى عملية التفكير وتكاد تجمع تلك الدراسات على أن الفكر يتشكل داخل وعاء اللغة.[١٩-٢٠-٢١-٢٢] ولن نستطيع أن نحل مشاكل العاميات والتعليم وتهميش اللغة العربية اجتماعيًا إلا من خلال استعمال اللغة فى مختلف مناشط المجتمع وأولها تعريب التعليم العام والجامعى.[٢٣-٢٤-٢٥]

لا تشكل النقاط التى عرجت عليها بأى حال من الأحوال كل القضية أو جلها فاللغة العربية بحر واسع علينا أن نغوص فيه حفاظًا على هويتنا ولكننى أردت أن أعرض بعض الفرص المتاحة للنهوض باللغة العربية ككيان متكامل على أمل رصد الحالة اللغوية بما فيها التطبيقات التقنية

---

١٨ The most natural state: Herder and nationalism; Alan Patten; History of Political Thought. Vol. XXXI. No. 4. Winter 2010; Imprint Academic 2010.

١٩ Through the language glass: why the world looks different in other languages; Guy Deutscher; First Edition 2010; Metropolitan Books; New York, USA.

٢٠ Language and Mind, Noam Chomsky; 3rd Edition; Cambridge University Press, Cambridge. U. K.; 2006.

٢١ How Language Shapes Thought; Lera Boroditsky; February 2011, Scientific American Magazine; pp. 63-65.

٢٢ محمد يونس الحملاوى؛ العربية حضارة؛ المؤتمر السنوى الخامس لمجمع اللغة العربية الفلسطينى؛ غزة؛ ٢٢ أبريل ٢٠١٨م

٢٣ محمد يونس الحملاوى؛ مشروع تعريب التعليم والعلوم والمعارف حل جذرى للحفاظ على اللغة العربية؛ المنتدى العربى الثانى للنهوض باللغة العربية؛ جامعة الدول العربية؛ القاهرة؛ ١٨-١٩ ديسمبر ٢٠١٦م

٢٤ محمد يونس الحملاوى وآخرون؛ استراتيجية النهوض باللغة العربية؛ تقرير داخلى؛ جامعة الدول العربية؛ القاهرة؛ ٢٠ يوليو ٢٠١٧م

٢٥ محمد يونس الحملاوى؛ الحرف العربى فى المجتمع: أداة للنهوض باللغة العربية؛ تقرير داخلى؛ لجنة الخبراء المصغرة لوضع استراتيجية النهوض باللغة العربية بجامعة الدول العربية؛ القاهرة؛ ١٤ يونيو ٢٠١٧م

الخاصة باللغة العربية والتى يلزمنا القانون برصدها.[٢٦-٢٧] ولعلى أشير إلى أن عدم توظيف التقنيات الحديثة فى مجال اللغة العربية لعامل هدم لبنيان اللغة ولبنيان المجتمع ككل.

٢٦ محمد يونس الحملاوى؛ اللغة العربية: الآفاق والتحديات؛ ورشة عمل الحوار المصرى الأمريكى، مركز الدراسات الحضارية وحوار الثقافات، جامعة القاهرة؛ ٤ نوفمبر ٢٠٠٩م

٢٧ محمد يونس الحملاوى؛ اللغة فى سياقها المعرفى التنموى؛ المؤتمر السنوى السابع عشر لتعريب العلوم؛ جامعة أسيوط؛ أسيوط؛ ١١-١٢ مايو ٢٠١٣م