Proposing an Arabic letters Coding system for different Programming languages

Abdul Monem S.Rahma, Ph.D (Prof.)[1]
110003@uotechnology.edu.iq

Maha A.Hmmood, Ph.D (Asst. Prof.)[2]
100139@uotechnology.edu.iq

Nada Ahmed Juma'a[3]
nadaalzubaidy@yahoo.com

Abstract This paper deals with the problem of using Arabic tex ts by researchers even in programming languages or internet search. The difficulties of Arabic language in programming languages and search the Internet have created a need to propose a system to code Arabic alphabets in all its forms and to be used in various applications and languages software that do not support the Arabic language in their content. Thus, it is possible to deal with the Arabic language and conducts many operations easily to download them through this system by using any Arabic text. The proposed system depends on three algorithms: the first is to code Arabic letters in any text, the second (code-function) is to code letters in any position shape in the text (first letters, middle letters, final letters and separated letters) and the third algorithm is to experiment the success rate of the proposed coding system. The proposed coding system has been examined for different texts representing Arabic poetries. Visual Basic.net 2015 language has been used to apply the proposed system. A confirmation function has been proposed which achieves 100% success rate of the coding system.

Keyword: Coding alphabet letters, Letters shapes, Arabic text.

[1] University of Technology, Department of Computer Science, Iraq, Baghdad
[2] University of technology, Department of Production and metallurgy, Iraq, Baghdad
[3] Ministry of Planning. Central Statistical Organization. Iraq, Baghdad

1. Introduction

The Arabic language faces many challenges, due to the lack of researchers to support Arabic language, and the narrow field of Internet search using Arabic language as compared with English language and others. On the other hand there has been no standard Arabic reference to solve Arabic letters problems [1]. The aim of this work is to develop a program to support Arabic language to be integrated into other computer systems such as automated speech and automatic recognition of Arabic speech, search engines and others through the insertion of Arabic texts and coding them to symbols that can be handled in other software languages.

This proposal helps most of the software systems to deal with the Arabic language letters by coding those using English letters and serial numbers. Coding of the various forms of Arabic letters in any text, poetic poem or article, is important to assist the researcher in using Arabic letters in several programming languages and adapting them in several operations, including statistical and mathematical operations.

Arabic letters coding has received less attention from the researchers. Most of them focus on coding other languages such English, Hebrew and Persian. Other researchers focus on using image processing techniques like Optical Character Recognition devise (OCR) or statistical techniques such as Markov hidden models [2].

In 2013, [3]M. Z. Khader presented research to the University of Jordan where the author reported one of the recommendations stated that minimum number of characters Arabic required for Arabic letters processing should include the following :

- The Arabic letters in different forms of writing, the first word, the middle, the last, and the separate.
- The letters (ى ,ة).
- The letters (ـئ ,ـأ ,ؤ ,آ ,إ ,أ ,ـآ ,ـإ).

In 2013, [4] Ali Abd Almisreb et. al. describe an effective and usable Arabic letters corpus uttered by Malay speakers. This corpus can be used to study the properties and the differences of pronunciations for non

native. The designed corpus consists of 1400 samples recorded by 50 Malay individuals (25 males and 25 females).

In 2008, [5] Ali Salamat suggested a classification method for Arabic letters based on the repetitions of each letters which depended on searching algorithm used for identification fuzzy adaptive resonance theory (ART), by using the neural network.

In 2008, [6] Ben Salm presented a method for classifying Arabic and Roman characters in multilingual documents. The method depended first on morphological transform of line text images, and then analyzing the properties of the text by extracting the text attribute.

2. Letters and Texts Coding

Letter coding is one of the most important fields of the computer programs. Many studies have dealt with this field in different systems to develop efficient secure systems [7].

Some ancient Arabic linguists arranged alphabetical letters in alphabetical terms. There is no doubt that the alphabetical order (أ ب ت) and the Arabic alphabet (أ ب ج د هـ و ز), which have already been mentioned based on the alphabetical order (أ ب ت). However, there are many differences with regard to symbols or non-Arabic characters, and the additional letters have gaps between them such as (ء , ى , ة ,). All are not assembled on their number or on their positions in the crews. Also, the configuration which computers consider independent letters is also the subject of a different sequence between the crew and another [8].

3. The Proposed Arabic Coding Letters System

Arabic language needed to be involved in computer and programming world therefore it must be available to be used in all programming languages without any limitations. To achieve this objective, a coding system is proposed in this work to facilitate involving Arabic text in programming and internet search. This must be conducted through some steps and the procedure depends on three algorithms.

The proposed system logic as following: firstly text will be input then reading and partitioning to words then partitioning each word to letters algorithm (1). From here started the main step which is checking

conditions as figure (1) on each letter and coded letters as the condition will achieved algorithm (2) then saving coded letter in array.

To make checking coding validity:

- Depended on encoding array and the original array.
- For each item in array (algorithm1) comparing between encoded and decoded text if they is similar then counter of matching will be increased otherwise the counter of not matching will be increased as shown in(algorithm(1)step 5).

Algorithm (1) illustrates the steps of coding letters in the inserted Arabic input texts, and the accuracy of results using confirmation function.

Algorithm (1): Coding Arabic letters system

Inputs: i=0 , .text

Output: Coded Text (CT)

```
Begin
Step 1 : for each Text in Dataset
                For each word in Text
                   For each letter in word
                              Letter-code = code function (letter)// algorithm (2)
                              Code-array(i)=letter code
                                  I=i+1
                          Next
                 Next
       Next
Step 2: For i = 0 To letter_code_shape.lenghth
          letter_code_break(i) = code_break(letter_code_shape(i))// algorithm(3)
   Next
Step 3: for      i=0      to  code-array.lenghth
            Decode-array=Decode-function(code-array(i))
          Next
Step 4 : for each word in T
                 For each letter in word
                    Original-array(i)=letter
                    i=i+1
                Next
           Next
Step 5 :      similarity count SC=0 , different count DC=0
                  for i=1 to original-array.length
```

```
                If original-array(i)= Decode-array(i) then
                 SC=SC+1
                    Else
                    DC=DC+1
                    End if
          Next
End
```

Texts will be read once they inserted to the proposed system, and then partitioning each word to letters using algorithm (1). The letter is processed by code function and applies any condition (C1,C2,C3,…..,C18) and checking which condition is true or false as figure (1). And then classified as these conditions to letter shape in word(first letter, middle letter, final letter, separate letter) , after getting the letter shape print letter code as classified in table(1). Besides all letters in the words of dataset in the last step of flowchart define two counters. One of them is to similarity letter code. Another counter is different letter code algorithm (1)step5.

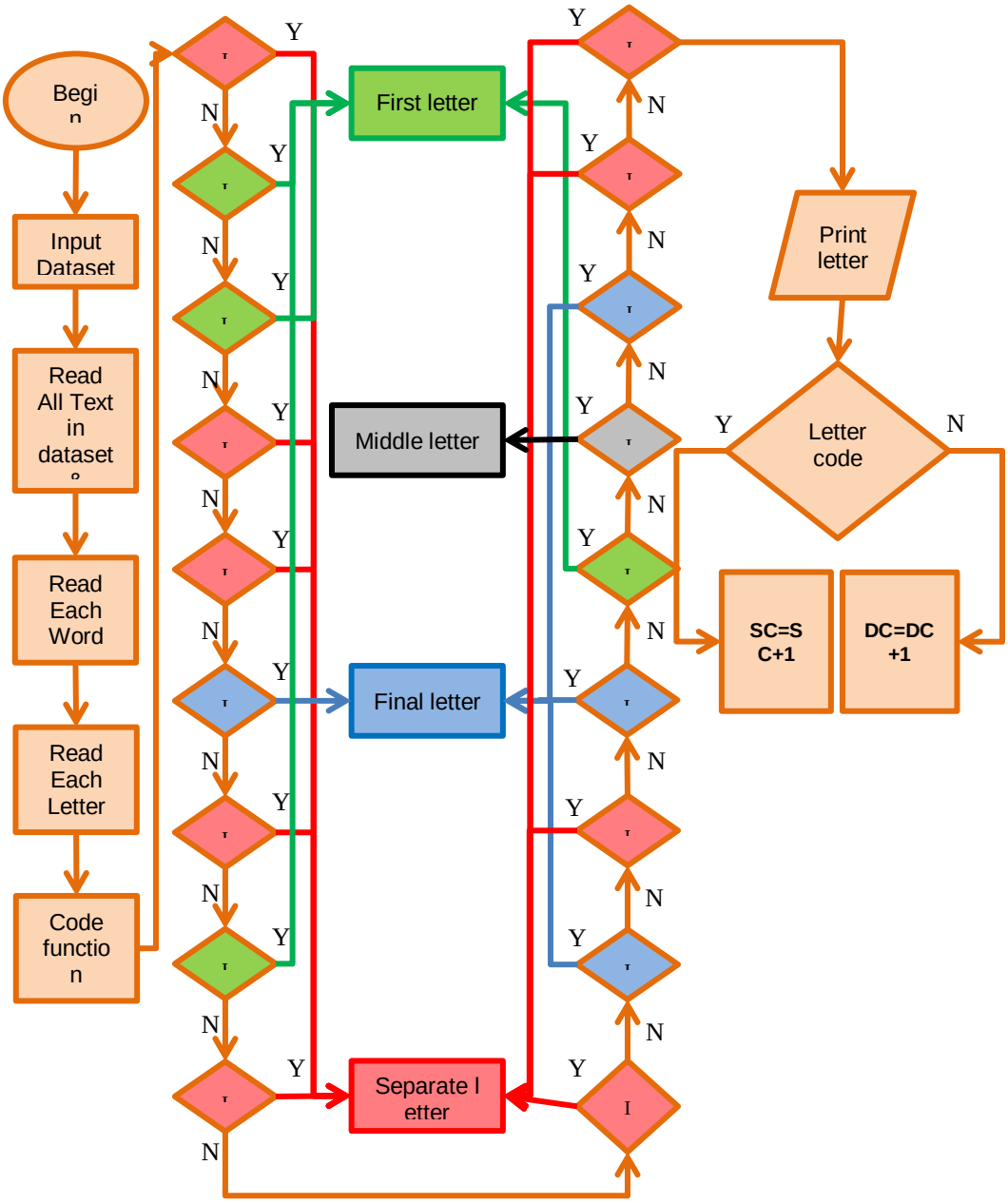


Figure 1. The proposed system diagram

Table 1. Describes all shapes of coded letters in Arabic text that could be inserted to the proposed system that used in algorithm (3) to return x in decoded form.

No	Letter X	First Letter	R1	Middle Letter	R2	Final Letter	R3	Separated Letter	R4
1	ء	-	-	-	-	-	-	ء	D11
2	ا	-	-	-	-	ـا	C12	ا	D12
3	ب	بـ	A13	بـ	B13	ـب	C13	ب	D13
4	ت	تـ	A14	تـ	B14	ـت	C14	ت	D14
5	ث	ثـ	A15	ثـ	B15	ـث	C15	ث	D15
6	ج	جـ	A16	جـ	B16	ـج	C16	ج	D16
7	ح	حـ	A17	حـ	B17	ـح	C17	ح	D17
8	خ	خـ	A18	خـ	B18	ـخ	C18	خ	D18
9	د	-	-	-	-	ـد	C19	د	D19
10	ذ	-	-	-	-	ـذ	C20	ذ	D20
11	ر	-	-	-	-	ـر	C21	ر	D21
12	ز	-	-	-	-	ـز	C22	ز	D22
13	س	سـ	A23	سـ	B23	ـس	C23	س	D23
14	ش	شـ	A24	شـ	B24	ـش	C24	ش	D24
15	ص	صـ	A25	صـ	B25	ـص	C25	ص	D25
16	ض	ضـ	A26	ضـ	B26	ـض	C26	ض	D26
17	ط	طـ	A27	طـ	B27	ـط	C27	ط	D27
18	ظ	ظـ	A28	ظـ	B28	ـظ	C28	ظ	D28
19	ع	عـ	A29	عـ	B29	ـع	C29	ع	D29
20	غ	غـ	A30	غـ	B30	ـغ	C30	غ	D30
21	ف	فـ	A31	فـ	B31	ـف	C31	ف	D31
22	ق	قـ	A32	قـ	B32	ـق	C32	ق	D32
23	ك	كـ	A33	كـ	B33	ـك	C33	ك	D33
24	ل	لـ	A34	لـ	B34	ـل	C34	ل	D34
25	م	مـ	A35	مـ	B35	ـم	C35	م	D35
26	ن	نـ	A36	نـ	B36	ـن	C36	ن	D36
27	ه	هـ	A37	هـ	B37	ـه	C37	ه	D37
28	و	-	-	-	-	ـو	C38	و	D38
29	ي	يـ	A39	يـ	B39	ـي	C39	ي	D39
30	ة	-	-	-	-	ـة	C40	ة	D40
31	ؤ	-	-	-	-	ـؤ	C41	ؤ	D41
32	ئ	ئـ	A42	ئـ	B42	ـئ	C42	ئ	D42

33	أ	-	-	-	-	ـأ	C43	أ	D43
34	إ	-	-	-	-	ـإ	C44	إ	D44
35	ى	-	-	-	-	ـى	C45	ى	D45
36	آ	-	-	-	-	ـآ	C46	آ	D46

Algorithm (2) experiments the conditions algorithm (code-function) letters in all shape (first, middle, final, separated):

Algorithm (2):code-function (conditions)

Input: Text ,Groups of letters (table 2):A, B, C , D , E ,F, G

Output : letter-code

```
Begin
Step 1: reading each text in dataset

Step2: for each letter in word
      If  letters =”ء” then letter-code=”D11”
            Else If A(letters) And j = 1 Then  letter_code = G(letters)
            Else If Not equal (A(letters)) And j = 1 And letter not equal  "ء" Then
              letter_code = D(letters)
            Else If Not (A(letters)) And j = 1 And letter equal "ء" Then
              letter_code = C(letters)
             Else If letters not equal"ء" And j equal word Len And A(letter) Then
               letter_code=C(letters)
            Else If letters not equal "ء" And j equal word Len And Not (A(letter))
Then
              letter_code =    E(letters)
            Else If A(letters) And letter equal " " Then
              letter_code = C(letters)
            Else If Not (A(letters)) And letter equal " " And letter2 not equal"ء"
Then
               letter_code = D(letters)
            Else If Not (A(letters)) And letter equal " " And letter2 equal "ء" Then
              letter_code = C(letters)
            Else If letters not equal "ء" And A(letter) And letter2 equal " " Then
              letter_code = C(letters)
            Else If letters not equal "ء" And Not (C(letter)) And letter2 equal " "
Then
              letter_code = E(letters)
             Else If A(letters) And A(letter) Then
              letter_code = C(letters)
             Else If A(letters) And Not (A(letter)) Then
```

```
                letter_code equal E(letters)
            Else If Not (A(letters)) And A(letter) And letter2 not equal "ء" Then
                letter_code = D(letters)
            Else If Not (A(letters)) And Not (A(letter)) And letter2 not equal "ء"
Then
                letter_code = F(letters)
            Else If Not (A(letters)) And Not (A(letter)) And letter2 equal "ء" Then
                letter_code = E(letters)
            Else If letters not equal "ء" And letter equal " " And letter2 equal " "
Then
                letter_code =  C(letters)
             Else If  A(letter) And A(letter)  Then
                letter_code = C(letters)
            End If
        Next
Step 2: return letter-code
End
```

To perform the proposed coding system, Arabic letters were put in 7 groups. Each group has involved different number of letters depending on some criteria as letters shape located in word. Table (2) contains all letters groups used in algorithm (2) that describes conditions of letter location in Arabic text input each group having letters that:

Table 2. letters Groups

NO.	A	B	C	D	E	F	G
1	ا	ا	ء	ب	ا	ب	ا
2	أ	أ	ا	ت	ب	ت	أ
3	آ	آ	ب	ث	ت	ث	آ
4	إ	إ	ت	ج	ث	ج	إ
5	و	و	ث	ح	ج	ح	و
6	ؤ	ؤ	ج	خ	ح	خ	ؤ
7	د	د	ح	د	خ	د	د
8	ذ	ذ	خ	ذ	د	ذ	ذ
9	ر	ر	د	ر	ذ	ر	ر
10	ز	ز	ذ	ز	ر	ز	ز
11		ة	ر	س	ز	س	
12		ى	ز	ش	س	ش	
13			س	ص	ش	ص	
14			ش	ض	ص	ض	

15			ص	ط	ض	ط	
16			ض	ظ	ط	ظ	
17			ط	ع	ظ	ع	
18			ظ	غ	ع	غ	
19			ع	ف	غ	ف	
20			غ	ق	ف	ق	
21			ف	ك	ق	ك	
22			ق	ل	ك	ل	
23			ك	م	ل	م	
24			ل	ن	م	ن	
25			م	ه	ن	ه	
26			ن	و	ه	و	
27			ه	ي	و	ي	
28			و	ئ	ي	ئ	
29			ي		ة		
30			ة		ؤ		
31			ؤ		ئ		
32			ئ		أ		
33			أ		إ		
34			إ		ى		
35			ى		آ		
36			آ				

Similarity or difference of letters coded depend on (X) value if letter code break returned same value of (X) that matches with table (1). Then counter of matching (similarity) will be increased otherwise if the value of (X) is not matching then the counter of not matching (difference) will be increased. For example if we use the word "عيناك" from input text "عيناك غابتا نخيل ساعة السحر" suppose dealing with the letter "ع" in the first of word matching with (X) value = “A29” (depend on table(1), column(R1)) then counter of similarity will be increasing one time and waiting checking on next letter from word that input in text. In algorithm (3) where Xn= X1,X2,X3,…..X36 and Dn = D12, D13, D14, ... ,D46 and Cn = C12, C13, C14, ... ,C46 and Bn = B13, B14, B15, ... , B42 , and An = A13, A14, A15, … , A42 and Ri = R1, R2, R3 as mentioned in table (1).

Algorithm (3):A proposed validity algorithm (letter code break)
Input : letter Xn
Output : code value

Begin
Step 1: If X = "D11" Then X = "ء"
Else
Step 2 : for I =1 to 4
For Xn 1 to 36
For Dn = 12 to 46
For Cn 12 to 46
For An 13 to 42
For Bn to 42
If Xn = Dn or Else Xn = Cn or else Xn= Bn or else Xn=An then X=Ri
Next
Next
Next
Next
Next
Step 3 : Return (X)
End.

4. The Interface

The proposed system is designed to consist of an interface that is embedded in the popular Visual Basic.net version 2015 (figure 2) containing the interface on the implementation code which contains the complete code of the encoding process. Then the interface consists of the text box to display the result of the coded letters for the input text and then processed with coded function and confirmed the accuracy of the proposed system by checking with similarity counter and different counter. Later the results have been applied in practice in the program MATLAB that does not support the Arabic language in characteristics and found that coded letters are suitable for many statistical and mathematical operations.

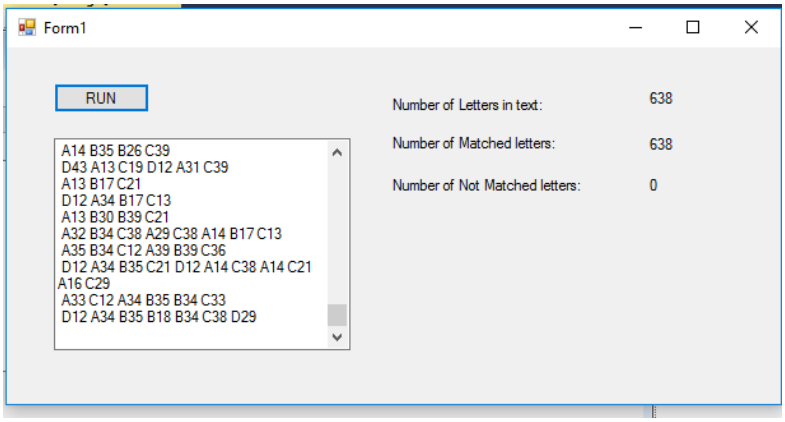
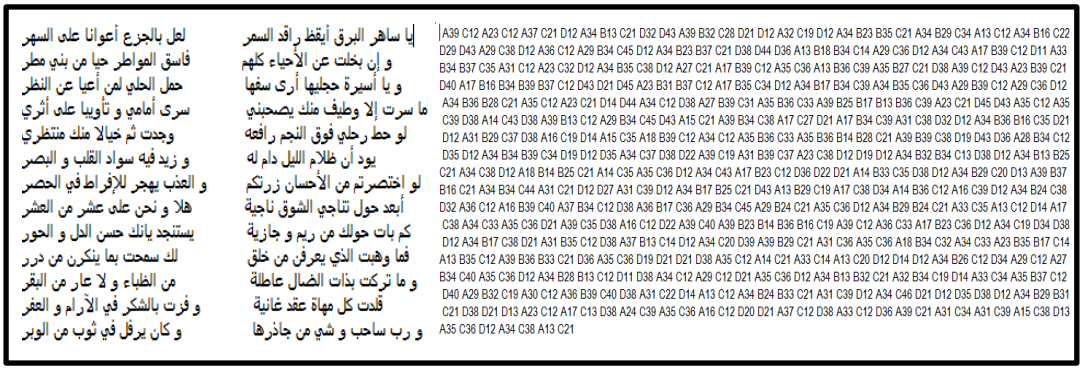


Figure 2. Interface of proposed system

5. Results and Discussion

The size of sample is 30 text file represented the poems file as example is shown in figure (3), all letters in poem will be coded as algorithm (1) depended on special conditions that guaranty from coded letter as their shape (first, middle ,final, separate) in the word. the program will beginning with first word in poem (يا ساهر) the letter shape (يـ) the first letter in word (يا) the code will become (A39) then letter shape (ـا) the final letter in word (يا) the code will become (C12), the word (ساهر) beginning with letter shape first (سـ) that code become (A23) then letter shape final(ـا) coded as (C12) then letter shape first (هـ) coded as (A37) and letter shape final (ـر) coded as (C21).

Let taking another word in the same poem (القلب) the first letter shape separate (ا) that code become (D11) next letter (لـ) that code become (A34) and letter shape middle (قـ) that code become (B32) , letter (لـ) that code become (B34), letter (ـب) that code become (C13).



(a) Letters in text (b) Coded letters

Figure 3. letters in text (a) coded letters (b)

The propose system achieved a success rate 100% by applying accuracy measurement equation(1) depends on number of poems that entered to system (30) poems, and number of poems that become coded (30) poems as following :

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{no.poems\ coded\ correctly}}{\boldsymbol{no.all\ poems}} \quad (1)$$

Where

$$\text{Accuracy} = \frac{30}{30} * 100 = 100\%$$

6. Conclusion

The use of Arabic language has encountered problems and difficulties in the beginnings of information and its multiple fields. The proposed technical developments system has contributed effectively to solving most of these problems. More clear, new and easy idea is suggested to coding Arabic letter, which helped many software to introduce Arabic alphabets in their forms to their systems and processing through statistical and mathematical operations used in solving many problems related to the Arabic language and improving their reality, a proposed system will which achieves 100% success rate when applied on samples contain from 30 poems they be coded all.

References

[1] كنعان ، أحمد ، (2012)،"اللغة العربية والتحديات المعاصرة وسبل معالجتها" ، جامعة دمشق ، كلية التربية ، ص7.

[2] الغامدي واخرون ، (2006)، "نظام حاسوبي لتشكيل النص العربي",ص8.

[3] علي العناسوة، محمد، (2013)، "اللغة العربية واستخداماتها في المعلوماتية (التوثيق والمكتبات والمعلومات)"، المجلة الاردنية في اللغة العربية وآدابها، المجلد (9) العدد (3).

[4] Almisreb and others, (2013) , *Arabic Letters Corpus Based Malay Speaker-Independent* , IEEE 3rd International Conference on System Engineering and Technology, Shah Alam, Malaysia.

[5] Ali Selamat and Ng Choon Ching, (2008), *Arabic Script Documents Language Identifications Using Fuzzy ART* , Second Asia International Conference on Modelling & Simulation, IEEE.

[6] S. Ben Moussa, A. Zahour, A. Benabdelhafid, A.M. Alimi, (2008) , *FractalBased System for Arabic/Latin, Printed/Handwritten Script Identification* , IEEE Pattern Recognition, , ICPR 2008. 19th International Conference on 8-11 Dec. 2008.

[7] Salomon, David, 2007, *Variable-length Codes for Data Compression* , British Library Cataloguing in Publication Data.

[8] خضر ، محمد ، (2001) ، "الحرف العربي والحوسبة "، الموسم الثقافي لمجمع اللغة العربية " عمان- الاردن ، ص4.

أقتراح نظام لترميز الحروف العربية يناسب مختلف اللغات البرمجية

أ. د. عبد المنعم صالح رحمه [1]
100139@uotechnology.edu.iq

أ. م. د. مها عبد الكريم [2]
110003@uotechnology.edu.iq

الباحث : ندى أحمد جمعه[3]
nadaalzubaidy@yahoo.com

المستخلص يتناول هذا البحث مشكلة استخدام النصوص العربية في مختلف اللغات البرمجية من قبل الباحثين في البرمجة او البحث في الانترنت. نظرا للصعوبات التي واجهت استخدام اللغة العربية في اللغات البرمجية والبحث في الانترنت فلقد اوجدت طريقة لاقتراح نظام ترميز الحروف العربية بجميع اشكالها وادخالها الى نظام الحاسوب واستخدامها في مختلف التطبيقات واللغات البرمجية التي لا تدعم اللغة العربية في محتواها وبذلك اصبح بالامكان التعامل مع اللغة العربية واجراء العديد من العمليات عليها بكل سهولة، يعتمد النظام المقترح على ثلاث خوارزميات: الاولى هي لترميز الحروف العربية لأي نص أما الثانية فهي خوارزمية لدالة ترميز الحروف في اي شكل من موقعها من النص (الحرف الابتدائي، الحرف الوسطي، الحرف النهائي، الحرف المنفصل، والخوارزمية الثالثة هي دالة لاختبار نجاح نظام الترميزالمقترح. م فحص النظام المقترح على عينة من الملفات النصية , وتم تطبيق النظام بشكل عملي على برنامج VB.NET 2015 ومن خلال فحص العينات المدخلة للنظام فقد حقق نسبة نجاح 100% وذلك على إعتبار التأكد من ترميز جميع حروف اللغة العربية من خلال دالة التأكيد التي تم اقتراحها ونجاحها في تلبية متطلبات النظام المقترح.

الكلمات المفتاحية : ترميز الحروف الابجدية ,أشكال الحروف ,نص عربي.

[1]الجامعة التكنلوجية / قسم علوم الحاسوب . بغداد / العراق
[2]الجامعة التكنلوجية / قسم هندسة الانتاج و المعادن . بغداد / العراق
[3]وزارة التخطيط / الجهاز المركزي للإحصاء . بغداد / العراق