

# A Qur'anic Code for Representing the Holly Qur'an (Rasm Al-'Uthmani)

Eng. Khaled M.S. Foda, Prof. Ahmed Fahmy, Prof. Khaled Shehata, Dr. Hanady Saleh  
 Computer Engineering Department, College of Engineering and Technology Cairo branch,  
 Arab Academy for Science, Technology & Maritime Transport, Cairo 2033, Egypt  
 { kfoda@ieec.org , afahmy1610@yahoo.com, khaledshehata58@gmail.com , hanady.issa@cairo.aast.edu }

**Abstract-** Holly Qur'an must be written correctly and precisely without any modification, despite the fact that some characters used in Qur'an does not have a corresponding Unicode representation. The aim of this paper is building a system that translates the Holly Qur'an to Qur'anic code, this translation will be done on three levels; character level, word level, and phrase level. Character level will be translated by extracting all the Arabic characters in the Holly Qur'an with all diacritics available in the Holly Qur'an regardless the linguistic combinations, and adding new character that has a symbol in the Holly Qur'an (Rasm Al-'Uthmani) [1] and has no Unicode representation for it. Word level will be translated by extracting all the duplicated words, generate a special code for these words. Phrase level will be created by extracting all similar patterns of variable lengths using enhanced (Lempel, Ziv, and Welch) LZW data Compression technique. We will use LZW technique for pattern extraction not for data compression.

**Keywords** – Qur'an diacritics, Rasm Al-'Uthmani, Text Size Reduction

## I. INTRODUCTION

Today, computers, Internet, mobiles, and tablets are widely used by millions of people all over the world. The processing power, storage, display capabilities, and various connectivity abilities of these devices made people depend on it in their life style, and reduce many numbers of every day belongs; for example in one device you can find radio, mobile, compass, GPS, and electronic library as well. The most important book for Muslims is the Holly Qur'an. It's originally saved by heart by listening, and reciting. It is a miracle itself [2]. You can find it in every mobile, computer, thousands of websites [3], and tablet applications. They do not want to read the Holly Qur'an only, but they also need to search, translate, and have its explication.

Muslims do not have any problem in having many explications, or translations [4], but the most important thing is that the Holly Qur'an must be written correctly and precisely, which means that any difference in any character or even diacritic is not acceptable at all. So any software developer has to compromise between precision and software features, either to present the Holly Qur'an as images without searching ability or presenting it as a text with taking into consideration a precision risk.

The Holly Qur'an is written in Arabic; which consists of letters, diacritics (التشكيل), punctuation (علامات الوقف) such

as { ٴ ٴ ٴ ٴ }, control conventions (اصطلاحات)

(الضبط) such as { ٴ ٴ ٴ ٴ } you can find more than 30 control conventions at the end of the Holly Quran-, and

marginal information like sura (chapter) name, goz'(part) number, page number, hizb, and quarter number.

## II. CHAPTER, AYA, AND PAGE NUMBERS

In this paper, we will use this representation to represent chapter name, aya number, and page number.

{Chapter number: Aya number: Page Number}.

## III. HARD AND SOFT COPIES OF QUR'AN

Regarding the hard copy of the Holly Qur'an we used Mushaf authorized from Al-Azhar, Islamic Research academy, and General Department for Research writing & Translation, Authorization date the 8<sup>th</sup> of September 1999, and 28/5/1420 Hijri to Dar Al-Maarifa. This Mushaf is 15 lines per page, where each page ends with end of an Aya.

Regarding the soft copy of the Holly Qur'an we downloaded it from www.tanzil.net (Uthmani). This copy was verified to match Medina Mushaf [5].

## IV. PROBLEM DEFINITIONS

In the current electronic Holly Qur'an there are six problems:

### A. Character has no Unicode

There is a character that has no Unicode to represent it in any electronically representation [4]. This character is Medium size Alf -faddara'tom-found in Souret Al-Baqara, Aya number 72, and page 11 shown in figure 1.

That is why the Holly Qur'an is still hand written not typed till now; the number of writers known for writing calligraphy script for Mushaf is very limited. Mr. Osman Taha [6] is the

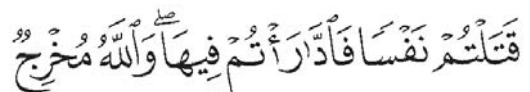


Figure 1. Shows the medium size Alf after the letter reh in faddara'tom {2:72:11}.





Figure 2. Khatat Othman Taha writing calligraphy script Mushaf Medina Quran.



## VII. DEFINITIONS

Line in the calligraphy script means line, but in the electronic representation for the Qur'an means aya, this means that the Qur'an contains 6,236 lines.

Character unit is the combination of the letter with the diacritics, or a letter without diacritics such as {  }, or combination of two letters such as {  }.

Word is a set of character units or a single character that is surrounded by spaces or space and new line. We consider "وَيَوْمَ" as one word because there is no space between "و" and "يَوْمَ".

Number of letters in Arabic language is 28 letter but in the electronic representation is much more.

## VIII. QUR'ANIC CODE GENERATION

This Qur'anic Code will be done on three levels; character level, word level, and phrase level. Character level will be translated by extracting all the Arabic characters in the Holly Qur'an with all diacritics available in the Holly Qur'an regardless the linguistic combinations, and adding new character that has a symbol in the Holly Qur'an (Rasm Al-Uthmani) [1] and has no Unicode representation for it. Word level will be translated by extracting all the duplicated words, generate a special code for these words. Phrase level will be created by extracting all similar patterns of variable lengths using enhanced (Lempel, Ziv, and Welch) LZW data Compression technique.

### A. Character Level

After checking all the letters with all the possible combinations the result was 1924 combinations (37 diacritics \* 52 letters) but in the Holly Qur'an we found only 575 combinations. This is because many combinations are not linguistically allowed, and other combinations are linguistic allowed but not found in the Holly Qur'an. Figure 6 shows the bits mapping of the 16 bits of the Qura'nic Code.

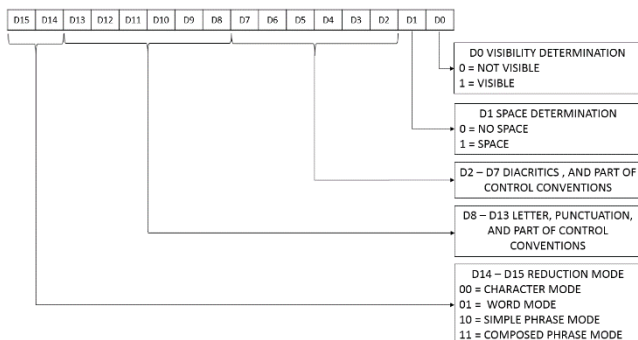


Figure 6. Bit mapping of the Qur'anic Code which is represented by 16 bits

D<sub>0</sub> is a bit used to define the visibility of that character or not. This is valid only for the quarter sign character "◌". Table 1 shows that there are 41 quarter sign that are not found inside the text, but found in the margin. Table 2 shows samples of 199 quarter sign that are already found in the text.

TABLE 1

SAMPLES OF THE quarter SIGNS THAT ARE NOT MENTIONED IN THE HOLLY QUR'AN

Page number	Goz' number	Hizb number	quarter number	Printed in the page margin	Soura number	Aya number
77	4	8	31	نصف الحزب	4	1
106	6	11	43	نصف الحزب	5	1
151	8	16	61	الحزب	7	1
187	10	19	75	نصف الحزب	9	1

TABLE 2

SAMPLES OF THE quarter signs THAT ARE MENTIONED IN THE HOLLY QUR'AN

Page number	Goz' number	Hizb number	Quarter number	Printed in the page margin	Soura number	Aya number
5	1	1	2	رابع الحزب	2	26
79	4	8	32	ثلاثة ارباع الحزب	4	12
112	6	12	45	الحزب	5	27
156	8	16	63	نصف الحزب	7	47

D<sub>1</sub> is a bit defines the last character in the word and has a space after it or not. We will discuss it later in the word level in more details. Table 3 shows all the possible letters and characters used in the Holly Qur'an. Table 4 shows all the possible diacritics used.

TABLE 3

ALL THE LETTERS USED IN THE HOLLY QUR'AN

Letter	D8- D13	Letter	D8- D13	Letter	D8- D13
	Count		Count		Count
-	0	آ	18	ج	36
	2		13,819	ح	3,317
ا	1	ا	19	ح	37
	15		25,184	ح	4,364
ا	2	أ	20	خ	38
	199		8,900	خ	2,497
ا	3	إ	21	د	39
	1		5,088	د	5,991
ا	4	ء	22	ذ	40
	12		3,059	ذ	4,932
ا	5	ـ	23	ر	41
	1972		495	ر	12,627
ا	6	ى	24	ز	42
	68		6,605	ز	1,599
ا	7	ي	25	س	43
	22		18,334	س	6,122
ا	8	ئ	26	ش	44
	603		921	ش	2,124
ا	9	ب	27	ص	45
	1,682		11,603	ص	2,074
ا	10	ة	28	ق	46
	5		2,344	ق	7,034
ا	11	ت	29	ك	47
	24,970		10,520	ك	10,497
ا	12	ث	30	ل	48
	706		1,414	ل	38,550
ا	13	ن	31	ه	49
	27,071		27,380	ه	14,962
ا	14	و	32	غ	50
	1,257		9,837	غ	1,221
ا	15	و	33	ع	51
	995		1	ع	9,405
ا	16	ف	34		
	1,686		8,747		
	17		35		

ط	1,273	ظ	853		
---	-------	---	-----	--	--

By looking to the last row in table 4, you can find “No Diacritics” - 11, and “Can’t have Diacritics” - 23. No diacritics means that letter comes without any diacritics, such as “ب, ي, و” or there are other characters that can’t come without diacritics “س, ز, ج”. Table 5 shows the characters that always come with diacritics, and the characters that can come without diacritics with their numbers. “Can’t have Diacritics” comes with characters such as “ٲ, ٳ”.

TABLE 4  
ALL THE DIACRITICS USED IN THE HOLLY QUR’AN

Diacritics	D2- D7	Diacritics	D2- D7	Diacritics	D2- D7
	Count		Count		Count
◌ْ	0	◌ُ◌ُ◌ُ	12	◌ُ◌ُ◌ُ	24
	270		5	◌ُ◌ُ◌ُ	99
◌ُ◌ُ	1	◌ُ◌ُ	13	◌ُ◌ُ	25
	30		1708		129
◌ُ◌ُ	2	◌ُ◌ُ	14	◌ُ◌ُ	26
	548		2,034		35,286
◌ُ◌ُ◌ُ	3	◌ُ◌ُ◌ُ	15	◌ُ◌ُ	27
	72		1		18
◌ُ◌ُ	4	◌ُ◌ُ	16	◌ُ◌ُ	28
	98		1,863		581
◌ُ◌ُ	5	◌ُ◌ُ	17	◌ُ◌ُ◌ُ	29
	4196		42446		6
◌ُ◌ُ◌ُ	6	◌ُ◌ُ	18	◌ُ◌ُ	30
	222		42		2,679
◌ُ◌ُ	7	◌ُ◌ُ	19	◌ُ◌ُ	31
	100		692		16,291
◌ُ◌ُ	8	◌ُ◌ُ	20	◌ُ◌ُ	32
	107,105		37372		1
◌ُ◌ُ	9	◌ُ◌ُ	21	◌ُ◌ُ	33
	3988		5,376		1
◌ُ◌ُ	10	◌ُ◌ُ	22	◌ُ◌ُ	34
	1		1		1
No Diacritics	11	Can't have Diacritics	23	◌ُ◌ُ	35
	36		37		1
◌ُ◌ُ	1	◌ُ◌ُ	1		

TABLE 5  
EACH LETTER AND THE NUMBER OF UN-DIACRITIC ASSOCIATED WITH IT

letter	Number of un-diacritic	letter	Number of un-diacritic
◌ْ	8,815	◌ُ	1,092
◌ُ	828	◌ُ◌ُ	4
◌ُ◌ُ	5,140	◌ُ◌ُ	1,331
◌ُ◌ُ◌ُ	5,729	◌ُ◌ُ	1
◌ُ◌ُ	1	◌ُ◌ُ	1
◌ُ◌ُ	0	◌ُ◌ُ	2
◌ُ◌ُ	0	◌ُ◌ُ	8
◌ُ◌ُ	0	◌ُ◌ُ	3
◌ُ◌ُ	0	◌ُ◌ُ	0
◌ُ◌ُ	0	◌ُ◌ُ	9
◌ُ◌ُ	3	◌ُ◌ُ	46

◌ْ	0	◌ُ	7
◌ُ	0	◌ُ◌ُ	1
◌ُ◌ُ	17	◌ُ◌ُ	0
◌ُ◌ُ◌ُ	10	◌ُ◌ُ	0
◌ُ◌ُ	7,044	◌ُ◌ُ	5,315
◌ُ◌ُ	0	◌ُ◌ُ	9,380
◌ُ◌ُ	0	◌ُ◌ُ	0
◌ُ◌ُ	0	◌ُ◌ُ	0
◌ُ◌ُ	18,450	◌ُ◌ُ	13,819

One of our gains is the reduction ratio ( $RR$ ),  $S_{old}$  is the size of the old representation, and  $S_{new}$  is the size of the new representation.  $RR$  is given by

$$RR = \frac{S_{old} - S_{new}}{S_{old}} \quad (1)$$

We concluded that the Qur’anic codes for the character level are: 575 unique characters. Table 6 shows sample of the Qur’anic code.

TABLE 6  
SAMPLE OF ONE CHARACTER UNIT FROM THE QUR’ANIC CODE

Qur’anic Code	(00- 010111- 011110-0-1) <sub>2</sub>		
Character Unit	◌ُ	Example	شَيْئًا
Number of Repetitions:	65		
Old Representation			
-	◌ُ	◌ُ	◌ُ
0x0640	0x0654	0x064B	0x06ED
Size of old representation in bytes:	8		
Total size of the old representation in bytes:	520		
Total size of the new representation in bytes:	130		
Reduction ratio:	75%		

The total reduction ratio at the character level (TRR) is given by equation (2) where  $R_i$  is the number of repetitions of each character unit,  $S_{old}$  is the size of the old representation, and  $S_{new}$  is the size of the new representation.

$$TRR = \sum_{i=0}^{574} \frac{S_{old_i} - S_{new_i}}{S_{old_i}} \quad (2)$$

As a result, the Qur’anic code based on character level achieves 46% reduction ratio. We made reduction from 1,277,564 bytes to 689,916 bytes.

### B. Word Level

Total number of words in the Holly Qur’an are 82,456 words, 12,452 unique words-repeated only once- such as {بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ}, 7,028 repeated words such as {اللَّهُمَّ صَلِّ عَلَى مُحَمَّدٍ وَآلِهِ}, 19,480 distinct word-non redundant words-, and. In Table 7 you can see samples of words that are repeated many times, and the reduction ratios for each word.

TABLE 7  
SAMPLE OF QUR'ANIC REDUCTION BASED ON WORD LEVEL

Qur'anic Code			
Word	فِي	اللَّهُ	الَّذِينَ
Number of Repetition	1,098	940	810
Size of old representation in bytes / word	6	14	18
Size of Qur'anic code representation in bytes / word	2	2	2
Total size of the old representation in bytes	6,588	13,160	14,580
Total size of the Qur'anic code representation in bytes	2,202	1,890	1,632
Reduction ratio:	66.57%	85.63%	88.80%

We found that many words are repeated many times but we will not use word reduction level on every repeated word, we will use it if equation (3) is satisfied. Where  $f_i$  is the frequency –number of repetition- of the word, and  $l_i$  is the length of the word as show the condition in (3)

$$f_i \times l_i - f_i - l_i - 1 > 0 \quad (3)$$

Equation (3) is driven form comparing the total old size of a word (number of repetition \* number of character per word \* number of bytes to represent each character) to the total size of Qur'anic code representation ((number of repetition + number of character per word + 1 which is added in the Qur'anic Code reference)\*number of bytes to represent each Qur'anic code which is 2 bytes) as shown in figure 7.

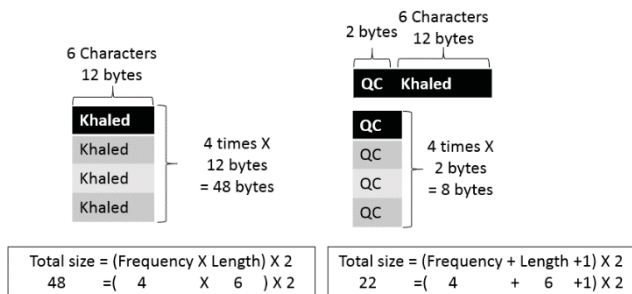


Figure 7. Comparing the size of the old representation with the Qur'anic Code

If we apply (3) to “ك”, its frequency is 1,972 and its length is one character then  $1972 \times 1 - 1972 - 1 - 1 = -2$  then we will not apply this techniques on it.

The 16 bit representation of the word level is shown in table 8, where the values of  $D_{14}$ - $D_{15}$  are 01,  $D_{11}$ - $D_{13}$  are defined by the number of characters to represent this word, but if we found this number is equal to 001 this means we have to look at  $D_9$ - $D_{10}$  to determine the length of the word if its length is greater than 7 as shown in table 8, X means don't care.

TABLE 8  
QUR'ANIC CODE REPRESENTATION OF WORD LEVEL

Length of word	Number of repeated words	Qur'anic code representation	Number of bits required to represent the word
2	145	01-010-XXX-D7---D0	8
3	881	01-011-X-D9---D0	10
4	1708	01-100-D10---D0	11
5	1749	01-101-D10---D0	11
6	1409	01-110-D10---D0	11
7	731	01-111-X-D9---D0	10
8	329	01-001-00-D8---D0	9
9	68	01-001-01-XX-D6---D0	7
10	7	01-001-10-XXXXX-XD2---D0	3
11	1	01-001-11-00000-0000	1

As a result, the Qur'anic code based on word level we reached 62.26% reduction ratio. We made reduction from 1,277,564 byte to 482,126 byte.

$$Num. of Spaces = Num. of Words - Num. of Lines \quad (4)$$

After looking to the Holly Qur'an we found that we have 82,456 word, 6,236 line and 76,220 space. This relation is found in (4) as shown in figure 8. By using  $D_1$  bit in the Qur'anic code in the character level we can reach 64.71% reduction ratio. We made reduction from 1,366,254 bytes to 482,126 bytes.

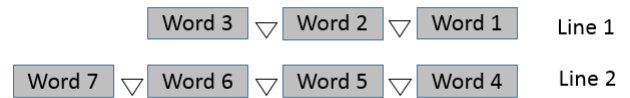


Figure 8. The relationship among spaces, number of words, and lines

### C. Phrase Level

In the Holly Qur'an many phrases are repeated such as

“بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ” {1:1:1}. This phrase was repeated 114 time, and its size is 38 character. So its original size is 8,664 bytes, but by applying the Phrase Qur'anic Code the new size is reduced to only 268 bytes;  $2 \times (114 + 19 + 1)$ . This leads to 96.90% reduction ratio.

We searched for the 25 sequential words repeated in the Holly Qur'an but we didn't find, but when we searched for 24 we found a phrase that is repeated twice “وَإِنْ كُنْتُمْ مَرْضَىٰ أَوْ عَلَىٰ سَفَرٍ أَوْ جَاءَ أَحَدٌ مِنْكُمْ مِنَ الْغَائِطِ أَوْ لَمَسْتُمُ النِّسَاءَ فَلَمْ تَجِدُوا مَاءً

فَتَيَمَّمُوا صَعِيدًا طَيِّبًا فَامْسَحُوا بِوُجُوهِكُمْ” {4:43:85}, and {5:6:108}. But of course not all the repeated phrases have the same number of words, table 9 shows the relationship between the number of words and number of repeated phrases found.

We used N-GRAM and fast pattern extraction algorithm [8] which was driven from LZW data compression technique [9] to extract table 9. There is a problem in that technique; the shortest pattern may be repeated in the longer pattern. That is why we will not provide any results related to the reduction

ratio of the phrase level. We think this work will be completed in another research.

TABLE 9  
THE RELATIONSHIP BETWEEN THE NUMBER OF WORDS AND THE NUMBER OF REPEATED PHRASE FOUND IN THE HOLLY QUR'AN

Number of words	Number of repeated phrase found	Number of words	Number of repeated phrase found
24	1	23	2
22	4	21	6
20	9	19	12
18	15	17	20
16	27	15	36
14	53	13	79
12	113	11	164
10	238	9	368
8	558	7	839
6	1304	5	2073
4	3371	3	5375
2	8023		

### IX. RESULTS AND COMPARISONS WITH TRADITIONAL SYSTEMS

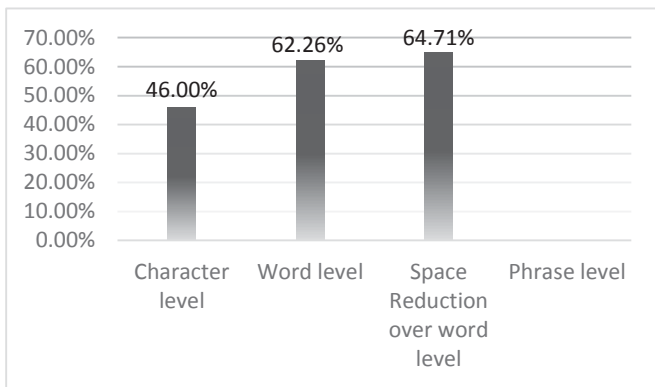


Figure 9. Show reduction ratios for each level

After using the Qur'anic code based on character level we reached 46% reduction ratio, from 1,277,564 byte to 689,916 bytes.

After using the Qur'anic code based on word level we reached 62.26% reduction ratio, from 1,277,564 byte to 482,126 bytes.

If we apply the space reduction convention then reached 64.71% reduction ratio, from 1,366,254 byte to 482,126 byte.

We didn't have the final reduction ratio of the phrase level as we didn't apply it yet as we mentioned before.

### X. CONCLUSIONS

The Unicode representation is not the best way to represent the Holly Qur'an, but by using the Qur'anic code we have solved 5 problems out of 6, we reached more than 65% reduction ratio, more searching capability, and standard way to store and present the Holly Qur'an on any electronic device.

### ACKNOWLEDGMENT

We Would like to thank Mr. Muhammad Mostafa El-Hossary who is graduated from Al-Azhar helped us in understanding all the diacritics, and revising all the result we have found for his effort in this research.

### REFERENCES

- [1] The Qur'anic Manuscripts, islamic-awareness.org, retrieved April 2, 2006
- [2] Zulkurnaini, N.A., Abdul Kadir, R.S.S., Murat, Z.H., Isa, R.M.. "The Comparison between Listening to Al-Quran and Listening to Classical Music on the Brainwave Signal for the Alpha Band," *Third International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*, 2012, pp. 181-186.
- [3] Bakar, A.B.A., "Evaluating the accessibility and visibility of Quran websites," *International Symposium in Information Technology (ITSim)*, 2010, pp. 1-4.
- [4] Tabrizi, A.A., Mahmud, R., "Issues of coherence analysis on English translations of Quran," 1<sup>st</sup> International Conference on Communications, Signal Processing, and their Applications (ICCSA), 2013, pp. 1-6
- [5] [http://http://tanzil.net/wiki/Tanzil\\_Project](http://http://tanzil.net/wiki/Tanzil_Project).
- [6] [http://en.wikipedia.org/wiki/Uthman\\_Taha](http://en.wikipedia.org/wiki/Uthman_Taha)
- [7] [http://en.wikipedia.org/wiki/Arabic\\_script\\_in\\_Unicode](http://en.wikipedia.org/wiki/Arabic_script_in_Unicode)
- [8] Kraty, M., Baca, R. ; Bednar, D., Walder, J., Dvorsky, J., Chovanec, P., "Index-based n-gram extraction from large document collections," *Sixth International Conference on Digital Information Management (ICDIM)*, 2011, pp. 73-78.
- [9] Basavamja, S.V., Sreenivas, T.V., "LZW Based Distance Measures for Spoken Language Identification," *The Speaker and Language Recognition Workshop*, 2006, pp. 1-6.