Fine-Grained Quran Dataset

Mohamed Osman Hegazi
Department of Computer Science
College of Computer
Engineering and Sciences
Prince Sattam Bin Abdulaziz University
Al-Kharj, Kingdom of Saudi Arabia

Anwer Hilal
Department of Computer Science
Preparatory Year Deanship
Prince Sattam Bin Abdulaziz University
Al-Kharj, Kingdom of Saudi Arabia

Mohammad Alhawarat
Department of Computer Science
College of Computer
Engineering and Sciences
Prince Sattam Bin Abdulaziz University
Al-Kharj, Kingdom of Saudi Arabia

***Abstract*—Extracting knowledge from text documents has become one of the main hot topics in the field of Natural Language Processing (NLP) in the era of information explosion. Arabic NLP is considered immature due to several reasons including the low available resources. On the other hand, automatically extracting reliable knowledge from specialized data sources as holy books is considered ultimately a challenging task but of great benefit to all humans. In this context, this paper provides a comprehensive Quranic Dataset as a first part (foundation) of an ongoing research that attempts to lay grounds for approaches and applications to explore the holy Quran. The paper presents the algorithms and approaches that have been designed to extract an aggregative data from massive Arabic text sources including the holy Quran and tightly associated books. Holy Quran text is transferred into structured multi-dimensional data records starting from the chapter level, the word level and then the character level. All these are linked with interpretations and meanings, parsing, translations, intonation roots and stems of words, all from authentic and reliable sources. The final dataset is represented in excel sheets and database records format. Also, the paper presents models of the dataset at all levels. The Quranic dataset presented in this paper was designed to be appropriate for: database, data mining, text mining and Artificial Intelligence applications; it is also designed to serve as a comprehensive encyclopedia of holy Quran and the Quranic Science books.**

Keywords*—Arabic Language; Holy Quran; Quranic Dataset; Text Mining; NLP

I. Introduction

In recent years, large amount of language datasets and corpora have been developed, these are increased with the spread of cloud computing applications and data linking. Different forms of datasets are available now in the web. However, datasets in Arabic language did not receive attention compared to datasets on other languages like English.

A. Objective of the Study

This study aims to build a group of datasets for the holy Quran, its interpretations, its meanings and related scientific books. Such group of books is considered one of the largest groups of books in Arabic literature. The holy Quran is composed of 114 chapters and about 6,236 verses with 77,477 words [1]. This group of books addresses every verse and word by interpretation, parsing, clarification; supporting these with the reasons of the revelations and the sayings (Hadith) of prophet Mohammad Peace Be Upon Him (PBUH). This results in a massive amount of text which could make it hard to process separately in the form of unstructured text. Also, compiling them to one group requires significant efforts and special processing. Therefore, this study focus in developing a model (algorithms and methodologies) to build a homogeneous dataset that fit all of these contents to produce a set of comprehensive structured data for the Holy Quran and its Scientific books, to be used as an encyclopedia for the Holy Quran and to serve as infrastructure for the technical applications that seek to produce results and carry out research on this vast amount of data.

B. Problem

The Major challenge of this study is how to convert this amount of unstructured text documents into groups of datasets? Provided that they are comprehensive, accurate and they preserve what in the sanctity, characteristics and relationships of this data. The other question is: can these groups of datasets be suitable for all types of applications that require different characteristics and techniques?

C. Previous Studies

Little number of research studies exist that are concerned with the production of datasets related to the Holy Quran and they are not comprehensive [2], [3], [4], [5]. For example [6] built semantic Quranic dataset, which include the Quran in several languages. They have established a set of data through the integration of data from different sources: Arabic, Amharic and Amazigh, Their goal was that the dataset would be appropriate to use in Natural Language Processing (NLP) applications, whereby their study have focused on the development of ontology to organize data. In [7], the author has compiled a dataset for nearly 8,000 Quranic verses related to each other in the form of pairs. They included explanation of the Quran and anaphora information for the purpose of conducting a number of text mining tasks. In a study presented by [8], the authors built a standard dataset for interpretation of the holy Quran.

The rest of the paper organized as follows: Section II is dedicated to present the model of the study, section III illustrates building of the dataset, section IV shows the details of testing the dataset, and finally section V includes conclusion and future directions.

II. The Proposed Model

The proposed model works according to the following algorithm (see figure 1):

- Prepare sources and ensure its credibility and authenticity.
- Convert sources into text files (the first stage of data processing (first phase of data processing)).
- Design databases.
- Process the data records.
- Segment Text and save in data tables (second phase of data processing).

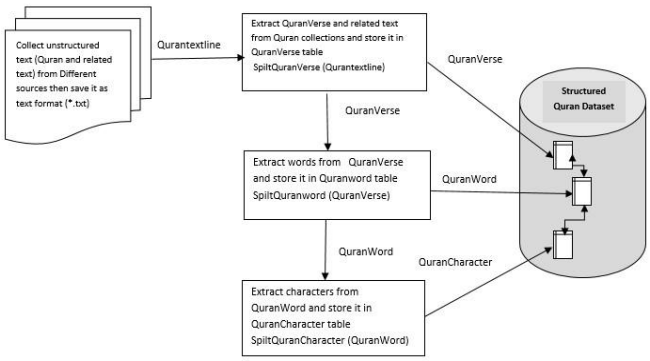

Fig. 1. The proposed model for the construction of the Quranic dataset

A. Sources processing

The dataset designed to carry all possible Quranic texts. At this stage we are going to provide the most available and needed sources, with capability to accept coming and updated sources. The sources that used at this stage are:

1) The holy Quran text with narration of Hafs/Nafi and Ottoman drawing http://tanzil.net/download/.

2) Four books in Quranic interpretation:

- Al-Jalalain interpretation. By Jalal ad-Din alMahalli and Jalal ad-Din as-Suyuti [9]
- Al-Muysar interpretation (King Fahad Quran Complex) [10].
- Al-Tabari interpretation [11].
- Aysar Altfasir interpretation by Abubaker Aljazairy [12].

3) Quran words meaning: Quran words, by Gazi Droopy [13].

4) French translation of Holly Quran text [14]

5) Three books in English translation of Holly Quran text:

- Yusuf Ali [15].
- Mohammad Habib Shakir English translation of Holly Quran text [16].
- Mohammed Marmaduke William Pickthall
- English translation of Holly Quran text [18]

6) One book in Quran words meaning in English.

7) The Transcription

8) One book in Quran syntax or declension, Tafsir Albahr Almuheet, by Alandalusi [19]

9) Quran words root and stemming.

B. Convert sources format to text files (the first stage of data processing)

At this stage, the Quran and its science books are processed and prepared in the form of text files to fit the software that will be prepared later to process and convert them into data records. One text file has been extracted which represents Quran with the narration of Hafs/Nafi according Ottoman drawing download from [i] . Figure 2 illustrates a sample of the text file of the holy Quran. Also all interpretations and others books has been modified and preprocessing manually to be in appropriated text format, figures 3- 4 show examples of a sample of the Al-Jalalain interpretation and Almuheet book respectively.

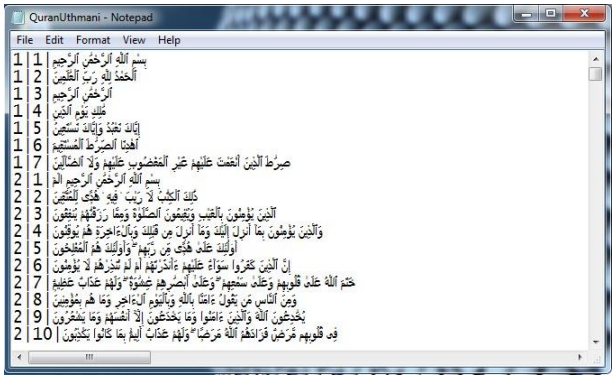

Fig. 2. Quran according to the Othmani drawing in a text file (Hafs from Nafi)

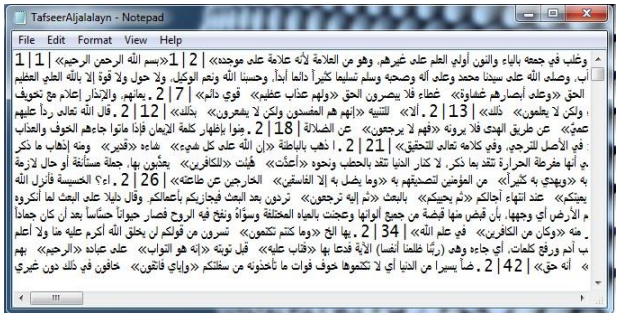

Fig. 3. A sample for some books of interpretations used in the dataset (Al-Jalalain interpretation)

C. Design data records

Datasets will be in two formats: spreadsheets and database records, these will be designed to be comprehensive for all properties in the form of fields in a multi-dimensional approach, based on holy Quran structures (which are character and word):

- Records on the character-level: It contains an index of the character shows chapter and the order of the word in the verse and the order of the character in the word and then the characteristics of the character in intonation and its diacritic (Superimposition character, whispered characteristic), pronunciation rules etc.

Fig. 4. A sample for some of Quranic scientific books (Albahr Almuheet, Quran syntax in a text file)

- Records on the word-level: the word, number of chapter, number of verse, word order in the verse, word meaning, parsing, number of characters in the word, intonation rules, words root, words stem etc.
- Records on the verse-level: verse text, number of chapter, number of verse, number of the chapter, reason for the revelation of verse, parsing, number of the words, Interpretation, meaning etc.

The records appear at each level in multi-dimensional manner as it is shown in figure 5. Because the dataset based on the Quran text, and Quran text is structured by verse, word, and character, then our data schema is based on these components (verse, word, character). The other books components are designed to be properties for the verse, word and/or character. Figure 6 shows the entity relationship diagram (ER) for the dataset schema.

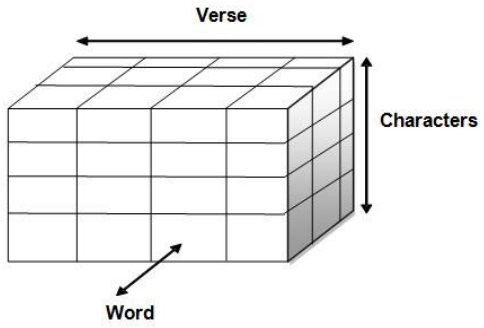

Fig. 5. Multi-dimensional records for the text of the holy Quran

D. *Segment Data into data tables records (the second stage of data processing)*

In this phase the Text documents are segmented and saved in the form of records according to the designed database tables. This phase is composed of three stages:

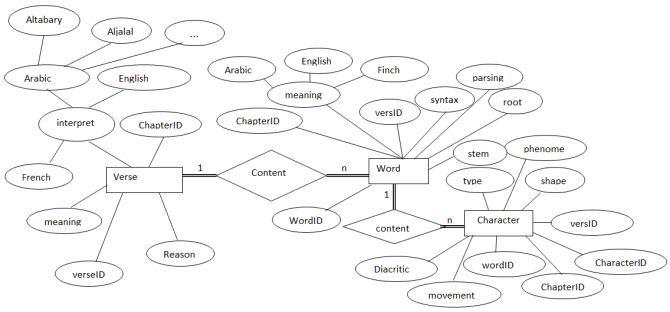

Fig. 6. Entity relation diagram ER of the proposed dataset schema

- Stage one: Identification of templates that will be used to divide the text on the basis of the verse template and the corresponding sentences interpretation and translation, the word template and the corresponding meanings and parsing and the rule of intonation, and character template and the corresponding intonation base statement and diacritic. These templates are considered the basis on which the text of the holy Quran and its sciences will be divided.
- Stage two: Building of algorithms that will read the text and then break it down into:
  - Verses, depending on the punctuation inside the Quranic texts (the Quran and its Sciences) like stop marks that help in the text segmentation process and is not considered as original part of the Quranic text, such as: spaces, comma, semicolon and so. Then, the related text from the Quran science books is inserted accordingly. Figure 7 shows a sample of the algorithms used in this stage.
  - Words and the corresponding rule of intonation and the meanings and translation. Figure 8 illustrates the segmentation algorithm used in segmentation of the words of the Quran.
  - Characters, and its corresponding TAJWEED (type, shape, diacritic, weight and phoneme level).
- Storing the text fragments in the appropriate table according to the designed relational database. So by end of this stage we can reach the acquired a clean datasets.

III. The Application (Building the Dataset)

The proposed model has been applied to a range of texts that have been selected from 14 electronic books as it mentioned in subsection II-A. Then these is text organized and implemented in three structured database tables and electronic sheets, the verse data, word data, and character data. The following subsections show how these data are implemented.

A. Verse template dataset

As illustrated in figure 9, all verses of the holy Quran become in the form of data records, the record contains an index for each verse with its interpretation and translation in English language... etc., parsing, and ...etc. The results of the table match the original data without errors, inconsistency or missing data compared to the verses of the holy Quran and its science books.

```
Input      : QurantextLine
Output     : QuranChapterNo, QuranVerseNo, QuranVerse
Function   SpiltQuranVerse (QurantextLine)
     Repeat
          For each QurantextLine Input
          QuranChapterNo   : = CUT (QurantextLine,
                    start_num, FIND ('|', QurantextLine,
                    start_num)
          QuranVerseNo: =CUT (QurantextLine, LENGTH
          (QuranChapterNo) +٢, (FIND ("|",     QurantextLine,
               LENGTH (QuranChapterNo) +٢) - LENGTH
          (QuranChapterNo)-٢))

          QVerse     : =RIGHT (QurantextLine, LENGTH
                     (QurantextLine) -       (Length
                     (QuranVerseNo) + LENGTH (QuranChapterNo)
                     +٢))
          Insert tblQuranVerse (ChapterNo, VerseNo, and Verse)
                 Values (QuranChapterNo, QuranVerseNo,
                     QuranVerse)
     Until EOF.
Return (QuranChapterNo, QuranVerseNo, and QuranVerse)
```

Fig. 7. Text reading algorithm and fragmentation to the verses

```
Input      : QurantextLine
Output     : QuranChapterNo,QuranVerseNo,WordOrder,QuranWord
Function   SpiltQuranWord (QurantextLine ())
Var i=١
     Repeat
          For each QurantextLine (i)
          QuranChapterNo   : = CUT (QurantextLine (i), ١,
                    FIND ('|', QurantextLine (i), ١)-١)
          QuranVerseNo: =CUT (QurantextLine, LENGTH
          (QuranChapterNo) +٢, (FIND ("|",     QurantextLine
          (i),  LENGTH (QuranChapterNo) +٢) - LENGTH
          (QuranChapterNo)-٢))

          QuranVerse     : =RIGHT (QurantextLine (i), LENGTH
                    (QurantextLine) -       (Length
                    (QuranVerseNo) + LENGTH (QuranChapterNo)
                    +٢))
          Word=spilt (QuranVerse," ")
          Var J: int, Quranword: String
          QuranWord=Word (٠)
          For J=١ to length (Word)
          Insert tblQuranWord (ChapterNo, VerseNo, and Verse,
                    word (j))
               Values (QuranChapterNo, QuranVerseNo,
                    WordOrder, QuranWord)
               J=j+١
          Next
               i=i+١
     Until EOF.
Return (QuranChapterNo, QuranVerseNo, and WordOrder,
QuranWord)
```

Fig. 8. Text reading algorithm and fragmentation to words

These records of data make it easy to extract information related to the verses and their meanings and make translation to English or French an easy task. For example, one can search for a verse by its significance or its translation or through the content of the word or the subject, besides the possibility of linking between verses and the meanings and reasons of revelation, also one can search based on themes, or similarity. Besides the possibility of carrying out some statistical and relational tasks easily.

SuraNameEng	ChapterNo	VerseNo	VerseTextClean	VerseText	QuranWordRoot	InterpreterAlJalalyn	Int
1-Al-Fatiha	1	2	الحمد لله رب العالمين	ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ	حمد ، الله|اله|ل|وله ، ربب ،	«والحمد لله» جملة خبرية قصد بها الثنا	حمد
1-Al-Fatiha	1	3	الرحمن الرحيم	ٱلرَّحْمَٰنِ ٱلرَّحِيمِ	رحم ، رحم ،	«الرحمن الرحيم» أي ذي الرحمة وهما	فلق
1-Al-Fatiha	1	4	مالك يوم الدين	مَٰلِكِ يَوْمِ ٱلدِّينِ	ملك ، يوم ، دين ،	و يوم القيامة، وخص بالذكر لأن	و يو
1-Al-Fatiha	1	5	اياك نعبد واياك نستعين	إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ	ايك|اي|ك ، عبد ، ايك|اي	«اياك نعبد واياك نستعين» أي نخصك	وحد
1-Al-Fatiha	1	6	اهدنا الصراط المستقيم	ٱهْدِنَا ٱلصِّرَٰطَ ٱلْمُسْتَقِيمَ	هدى ، صرط ، قوم ،	«اهدنا الصراط المستقيم» أي أرشدنا	قوم
1-Al-Fatiha	1	7	صراط الذين انعمت عليهم غير المغضوب عليهم ولا الضالين	صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ ٱلْمَغْضُوبِ عَلَيْهِمْ وَلَا ٱلضَّآلِّينَ	صرط ، الذى ، نعم ، على|هم	«صراط الذين أنعمت عليهم» بالهداية	لصد
2-Al-Baqara	2	1	الم	الٓمٓ	١|ل|م ،	«الم» الله أعلم بمراده بذلك.	طعم
2-Al-Baqara	2	2	ذلك الكتاب لا ريب فيه هدى للمتقين	ذَٰلِكَ ٱلْكِتَٰبُ لَا رَيْبَ فِيهِ هُدًى لِّلْمُتَّقِينَ	ذلك ، كتب ، لا ، ريب ، فى|ه	«ذلك» أي هذا «الكتاب» الذي يقرؤه	ك
2-Al-Baqara	2	3	الذين يؤمنون بالغيب ويقيمون الصلوة ومما رزقناهم ينفقون	ٱلَّذِينَ يُؤْمِنُونَ بِٱلْغَيْبِ وَيُقِيمُونَ ٱلصَّلَوٰةَ وَمِمَّا رَزَقْنَٰهُمْ يُنفِقُونَ	الذى ، امن ، ب|غيب ، قوم|و	«الذين يؤمنون» يصدقون «بالغيب»	ه ح

Fig. 9. Sample for a verse template data

B. Word Template Dataset

As illustrated in figures 10, all words of the holy Quran became a valuable form of data records and the record contains a single index for each word, besides the word and its meaning, parsing and meaning in English, and its root and stem …etc. The results were identical to the original text without any error or missing data. These records make accessing words of the holy Quran easy and quick in several ways. One can also use some of the words characteristics to link between the subjects and conclude results from the holy Quran and interpretation books. For example, if we wish to conduct research on usury or prayer or Zakat, can search all words that contain the desired subject through the root or stem, or directly through the word then connect the verses and interpretation, and then we can come up with an integrated research on what is stated in the Quran and the interpretation book about the search. Beside the possibility to make any statistical or relational tasks on the word level easily and quickly.

VerseNo	WordNo	QuranWord	WordRoot1	QuranCleanWord	QuranWordUsmani	WordRoot2	WordPronounce	MeaningEnglish
1	1	بسم	س م و	بسم	بِسْمِ	ب|سمو|وسم	Bismi	In (the) name
1	2	الله	ا ل ه	الله	ٱللَّهِ	الله|اله|ا|وله	Allahi	(of) Allah,
1	3	الرحمن	ر ح م	الرحمن	ٱلرَّحْمَٰنِ	رحم	alrrahmani	the Most Gracious
1	4	الرحيم	ر ح م	الرحيم	ٱلرَّحِيمِ	رحم	alrraheemi	the Most Merciful
2	1	الحمد	ح م د	الحمد	ٱلْحَمْدُ	حمد	Alhamdu	All praises and tha
2	2	لله	ا ل ه	لله	لِلَّهِ	الله|اله|ا|وله	lillahi	(be) to Allah,
2	3	رب	ر ب ب	رب	رَبِّ	ربب	rabbi	the Lord
2	4	العالمين	ع ا ل م	العالمين	ٱلْعَٰلَمِينَ	علم	alAAalameena	of the universe
3	1	الرحمن	ر ح م	الرحمن	ٱلرَّحْمَٰنِ	رحم	alrrahmani	the Most Gracious
3	2	الرحيم	ر ح م	الرحيم	ٱلرَّحِيمِ	رحم	alrraheemi	the Most Merciful
4	1	مالك	م ل ك	مالك	مَٰلِكِ	ملك	Maliki	(The) Master
4	2	يوم	ي و م	يوم	يَوْمِ	يوم	Yawmi	(of the) Day
4	3	الدين	د ي ن	الدين	ٱلدِّينِ	دين	aldeeni	(of the) Judgment.
5	1	اياك	ايك|اي|ك	اياك	إِيَّاكَ	ايك|اي|ك	iyYaka	You Alone

Fig. 10. Sample of Word dataset according to the word root and the pronunciation and meaning in English

C. Character template data set

As can be seen in figure 11, all the characters of the holy Quran become data records, whereby records contain a single index for each character besides: character, diacritic and time. These records of the data make it easy to produce reports at the character level. For example, in [17], the authors presented statistics on the letters only without offering to provide any information to distinguish between like-like characters ((ا،إ،أ) (ى,ي)). Character dataset can be considered more statistically accurate and comprehensive (see figures 14- 15). One of the interesting aspects of this dataset is its comprehension; the methodology didn't ignore even the diacritic of the character

which allows users to handle the character with or without diacritic. Such aspect provides multiple benefits, such as the beneficial aspects of intonation studies. Figure 16 shows statistics on diacritic.

D. Application interface

The dataset are available and ready for use, its implemented in three format, access table, excel sheet, and SQL-server database. Also an application for searching the content of the dataset is generated and available online at http://www.anwermustafa.com/su/. This application provide searching engine for the all the contents of the dataset for example, searching using the exact Quranic word, the meaning, the English or French word, any part of sentence, the root, the stemming ... etc. Figure 12 shows an example of using this application in searching for the word "نافق".

ChapterNo	VerseNo	WordNo	LetterOrder	WordText	Letter	SuraNameA	IntitialLetter	InitialLetterOrder
2	2	1	1	ذٰلك	ذ	2. البقرة		
2	2	1	2	ذٰلك	ل	2. البقرة	Yes	
2	2	1	3	ذٰلك	ك	2. البقرة		
2	2	2	1	الكتاب	ا	2. البقرة	Yes	
2	2	2	2	الكتاب	ل	2. البقرة	Yes	
2	2	2	3	الكتاب	ك	2. البقرة		
2	2	2	4	الكتاب	ت	2. البقرة		
2	2	2	5	الكتاب	ا	2. البقرة	Yes	1
2	2	2	6	الكتاب	ب	2. البقرة		
2	2	3	1	لا	ل	2. البقرة	Yes	2
2	2	3	2	لا	ا	2. البقرة	Yes	1
2	2	4	1	ريب	ر	2. البقرة		
2	2	4	2	ريب	ي	2. البقرة		
2	2	4	3	ريب	ب	2. البقرة		
2	2	5	1	◌	◌	2. البقرة		
2	2	6	1	فيه	ف	2. البقرة		
2	2	6	2	فيه	ي	2. البقرة		

Fig. 11. Model for a character template data in accordance with the shape, diacritic, and the Word

Fig. 12. A snapshot of the search engine of the Quranic dataset

IV. Testing The Dataset

This section presents the aspects that confirm the algorithms used and the correctness of the data produced. Because we used automatic algorithms for transferring the text to data base table, it will be important to validate these algorithms by comparing the results with the actual data (the original books). As its shows in figures 9- 16 the results are matching the real data, for example:

1) The holy Quran is composed of 114 chapters and 6,236 verses with 77,477 words [1], when we compare this statistics with the statistics that generated from our data set we found the same result. Figures 13- 14 show statistical reports that generating form our dataset.

2) All Quranic verses are included in the data set and linked correctly with its appropriated chapters and its parameters, such as interpretation, translation ...etc.

Figure 9 shows an example for this aspect.

3) The Quranic words that generated by our algorithm are the same as that in the holy Quran, same number of words, same relation to its verse and chapter, and linked correctly with its corresponding meaning, translation, syntax, root, stemming, ...etc. Figure 10 shows an example of this aspect.

4) The statistical analysis of Quran character provide a good result in comparison with the previous work example [17]. Figures 14- 16 show examples of such results.

عدد الأحرف	عدد الكلمات	عدد الآيات	عدد السور
332,837	72,135	6,236	114

Fig. 13. Statistical summary for Quran components

ء	ا	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض
1578	59,616	11,603	2,344	10,520	1,414	3,317	4,364	2,497	5,991	4,932	12,627	1,599	6,124	2,124	2,072	1,686
ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	ئ	Total	
1,273	853	9,405	1,221	8,747	7,034	10,497	38,639	27,071	27,382	14,962	24,813	673	2,592	22,085	1,182	332,837

Fig. 14. Sample of statistics for Quran characters

Chapter Name	ء	ا	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ
1. الفاتحة		26	4		3			5		4	1	8		3		2	2	2	
2. البقرة	125	4,719	919	216	970	128	200	330	191	458	330	876	107	452	168	155	133	99	62
3. آل عمران	68	2,662	575	100	557	52	93	173	104	252	218	510	67	228	86	88	66	50	36
4. النساء	71	3,011	479	116	561	75	136	198	128	301	181	491	51	307	82	122	101	65	45
5. المائدة	62	2,303	395	87	392	59	93	173	85	231	182	381	57	215	63	74	53	52	18
6. الأنعام	80	2,164	454	70	404	58	117	162	75	199	207	506	66	213	132	69	66	52	48
7. الأعراف	76	2,654	509	105	466	61	165	166	129	246	229	534	41	300	83	97	62	56	43
8. الأنفال	22	898	174	30	162	18	38	54	34	87	95	187	23	71	27	29	27	22	8
9. التوبة	26	1,900	350	87	307	31	102	138	86	207	145	412	55	194	49	63	61	34	29
10. يونس	40	1,359	239	43	253	30	83	100	40	126	115	257	26	130	42	32	43	17	32
11. هود	49	1,424	284	55	236	32	79	88	57	146	96	325	35	123	43	60	32	28	24
12. يوسف	31	1,388	270	33	233	29	80	103	71	141	81	257	34	167	42	61	32	19	14
13. الرعد	28	629	152	26	94	16	35	44	23	77	47	137	21	71	22	20	24	12	6
14. إبراهيم	23	617	135	18	114	17	34	35	31	73	53	160	21	77	25	24	22	11	11
15. الحجر	12	530	101	18	73	4	43	44	24	41	22	96	19	67	16	23	14	11	9
16. النحل	41	1,297	270	67	239	45	71	91	72	123	122	289	45	139	64	33	32	25	22
17. الإسراء	28	1,252	259	51	202	27	85	90	61	143	87	301	38	148	42	31	44	29	16

Fig. 15. An example of some of the Quran letter statistics

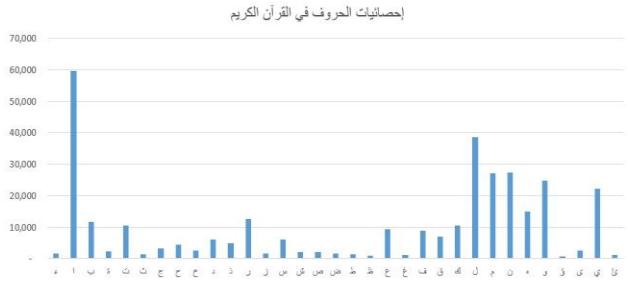

Fig. 16. Statistics for the letters for the holy Quran

ChapterNo	SuraNameAr	◌	◌	◌	◌	◌	◌	◌	◌
2	2. البقرة	4	6	179	4		19	161	117
3	3. آل عمران			61	1		10	63	81
4	4. النساء			154	2		11	60	39
5	5. المائدة	4	7	100	4		8	66	21
6	6. الأنعام		2	96	3		9	101	37
7	7. الأعراف	2	4	85	1	1	9	89	19
8	8. الأنفال		6	42			3	13	14
9	9. التوبة		6	91			8	36	28
10	10. يونس		3	58	1		5	48	10
11	11. هود			66	1		6	53	11
12	12. يوسف			43			5	64	5
13	13. الرعد		1	29		1	3	24	17
14	14. إبراهيم	2		20			2	23	8
15	15. الحجر		1	3			1	3	1
16	16. النحل		10	52		1	5	35	16

Fig. 17. an example of some of the characters diacritic statistics (formation)

رقم الجزء	رقم السورة	التشكيل ً	ٌ	ٍ	َ	ُ	ِ	ّ	ْ	Grand Total
1	1				41	6	31	11	14	103
	2	89	84	51	3,815	1,275	1,419	1,330	685	8,748
2	2	79	118	70	3,878	1,299	1,526	1,356	843	9,169
3	2	32	57	37	1,512	506	566	503	317	3,530
	3	56	51	34	2,274	719	1,025	741	519	5,419
4	3	54	73	52	2,733	1,033	1,173	1,051	594	6,763
	4	62	18	22	908	338	311	317	185	2,161
5	4	264	47	53	3,762	1,207	1,546	1,312	748	8,939
Grand Total		**636**	**448**	**319**	**18,923**	**6,383**	**7,597**	**6,621**	**3,905**	**44,832**

Fig. 18. An example of some of the characters diacritic statistics (formation)

V. CONCLUSION

This paper presented a model that contains algorithms that read Arabic texts and convert them into datasets in the form of relational database records and electronic sheet. This model is applied to one of the most important and largest Arabic texts; the holy Quran and more than 13 books include English and French translations, and sciences books such as interpretations, syntaxes, word meaning... etc. Algorithms used have achieved accurate results when its compare with the original text. This makes this model an appropriate model for the treatment of Arabic texts and building datasets from Arabic texts.

Applications of the model in the construction of the Quranic dataset proved that the automated algorithms provide correct result when converting any text from Quran and its Science books into data tables. The finding is tested and compared with the original sources and shows that the data in the records are same as the original text, this implies that the model presented in this paper is trusted and can be used in accommodate any future data within any of the prepared templates or by creating new templates, all can be done automatically and without new settings.

Quranic datasets presented in this paper has put the Quran and its science books text in the form of encyclopedia to serve all areas of research which contain a tremendous amount of knowledge and meanings and relationships. Making it a scientific contribution in this area since most previous studies did not reach such a level of datasets.

The model capabilities which are used in the construction of these groups of datasets makes it flexible to accommodate new set of text titles related to the holy Quran. It also allows issuing reports and search results in multi-dimensional manner being a relational datasets.

Quranic datasets presented in this paper were placed in templates and then converted into relational database records which establish the infrastructure for the implementation of multi-technology applications; such as: database applications, data mining and knowledge discovery applications, Artificial Intelligence applications, machine learning applications, natural language processing NLP and statistical applications. Accordingly, our next work will exploit these generated datasets in exploring and discovering worthy results appropriate to the greatness of the holy Quran and its unlimited knowledge for everything ”And you went book gives an account of everything, and a guidance and a mercy and glad tidings to Muslims” (Nahl: 89)). In addition, we expect to use this dataset in conducting research work in the area of Arabic language processing and/or Arabic text mining because the holy quran is considered as a standard Arabic reference for more than 14 century.

REFERENCES

[1] M. Alrabiah, N. Alhelewh, A. Al-Salman, and E. Atwell, “An empirical study on the holy quran based on a large classical arabic corpus,” *International Journal of Computational Linguistics*, vol. 5.1, pp. 1–13, 2014.

[2] M. Shoaib, M. Nadeem Yasin, U. Hikmat, M. Saeed, and M. Khiyal, “Relational wordnet model for semantic search in holy quran,” in *International Conference on Emerging Technologies, 2009. ICET 2009.*, Oct 2009, pp. 29–34.

[3] M. Yunus, R. Zainuddin, and N. Abdullah, “Semantic query for quran documents results,” in *Open Systems (ICOS), 2010 IEEE Conference on*, Dec 2010, pp. 1–5.

[4] K. Dukes, E. Atwell, and N. Habash, “Supervised collaboration for syntactic annotation of quranic arabic,” *Language Resources and Evaluation*, vol. 47, no. 1, pp. 33–62, 2013.

[5] K. Dukes. (2014, Dec.) Quranic arabic corpus. [Online]. Available: http://corpus.quran.com/

[6] M. A. Sherif and A.-C. Ngonga Ngomo, “Semantic quran: A multilingual resource for natural-language processing,” *Semantic Web Journal*, vol. 1, pp. 1–5, 2013.

[7] A. B. Muhammad, “Annotation of conceptual co-reference and text mining the quran,” Ph.D. dissertation, University of Leeds School of Computing, 2012.

[8] M. Safi, M. A. Latif, and A. Tayeb, “The construction of a standard set of data to interpret the holy quran,” in *Taibah International University conference in employing information technology to serve the Holy Quran and its Sciences (NOORIC’1435 / 2013), Medina, Kingdom Saudi Arabia*, 2013.

[9] Suyuti, Jalaluddin 2010 “Al-Jalalain interpretation”, DKI, Beirut 2010.

[10] Aa’d Alqrny ،Aldktwr(2011) “altfsyr almysr” King Fahad Quran Complex. 2011.

[11] Tabari, Jarir (2009) “Al-Tabari Interpretation” DKI, 2009

[12] Abubaker Aljazairy (2007) “Aysar Altfasir” Maktabat al-'Ulum wa-al-Hikam, KSA, 2007.

[13] Shamela.ws. (2015) Explanations of the meamings of the words of the holy quran.

[14] French translation of Holly Quran text King Fahad Quran Complex. 2007.

[15] Abdullah Yusuf Ali (2000), “ The Holy Qur'an” Wordsworth Edition Limited 2000

[16] Mohammad Habib Shakir (2010)” (Quran Translator) MobileReference 2010.

[17] M. Z. Khadr and M. A. Zaki, “A statistical study of the words of the qur’an,” in *Third Conference of the Arab Language and Literature ”Recent trends in linguistic and literary studies.” International Islamic University, Malaysia*, 2011.

[18] Marmaduke Pickthall (1930)”The Meaning of the Glorious Quran” Global Grey edition 1930

[19] Abi Hian, Ather Aldein Mo. (2010) “ Tfsyr Albhar Almhayt” Dar al kotob al ilmiyah - Beirut – Lebanon 2010

[i] -http://tanzil.net/download/