Journal of Theoretical and Applied Information Technology
20th April 2014. Vol. 62 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ONLINE DATABASE OF QURANIC HANDWRITTEN WORDS

[1]MUSTAFA ALI ABUZARAIDA, [2]AKRAM M. ZEKI, [3]AHMED M. ZEKI
[1,2]Kulliyyah of Information and Communication Technology
International Islamic University Malaysia Kuala Lumpur, Malaysia
[3]Department of Information Systems, College of Information Technology,
University of Bahrain, Sakhir, Kingdom of Bahrain
E-mail: [1]abuzaraida@yahoo.com , [2]akramzeki@iium.edu.my

ABSTRACT

In this paper, an online Arabic handwritten words database is presented to be the first online Quranic handwritten words dataset. The dataset was collected naturally using Acer computer tablet 1.5 GHz core i3 by writing the words on a smooth touch screen using a special pen stylus. Here, a platform interface was designed to collect the handwritten words using Matlab environment. Handwritten words were chosen as the most common words repeated in the holly Quran. The initial version of (Quranic Handwritten Words) QHW database includes 120 handwritten words and divided equally into two sets to be written by 200 writers in total. The QHW database contains 12000 sample including more than 42,800 characters and 23,300 sub words. Handwritten words were mainly written by variety of people aged between 6 and 50 years from several countries. The main aim of creating this database is to be used in an online Arabic recognition system and make it available to other researchers while there is no public database can be use nowadays. Also, Some preprocessing steps those applied to standardize the all words of this database are presented in this paper.

Keywords: *Quran, Arabic Language, Handwriting Recognition, Online Words Database.*

1. INTRODUCTION

Many studies have been published in the field of handwriting recognition in the past three decades. Generally, there are two types of recognition systems based on the input data, which are Online and Offline type. Offline recognition system is about recognizing scanned printed or handwritten texts. On the other hand, online recognition can be useful for recognizing handwritten texts only. The idea is to recognize the sequence of the coordinates (x,y) of the stylus movements actions on the device surface which representing the handwritten text [1] [2]. In this case, the writing actions are known and representing to the system usually in a text file.

Until the beginning of the nineties, online handwriting recognition approaches were mainly an academic research. The most results were just reported in the open literature due to the limitation of resources and due to the availability of touch devices in that era. However, this situation has completely changed in the past few years with the rapid growth of the pen computing industry [3].

With the growing popularity of touch screen devices such as PDA (personal digital assistant), smart phones, and tablet PCs, online handwritten text recognition field is finding wide adoption as an input method and in applications including electronic dictionaries, games and educational software. Nowadays, these devices are commonly available in the market worldwide dealing with many languages spoken by billions of people around the world such as Latin, Chinese, Japanese, Indian, Korean, Arabic, and many others from textual or speech manner [4].

Online recognition is practical choice when it is difficult to type or there is no enough space to provide keyboard. So in this situation, it is comfortable to write handwritten texts and want the system to recognize them. However, there are some available systems which can recognize online handwritten continuous English words. Some similar works for Japanese and Chinese texts also exist . Although, few researches have been involved in online recognition of handwritten Arabic texts especially for recognizing words [5], and it seems to be necessary to do more research to enhance the output of the current systems [4].

The rest of this paper is organized as follows. In section 2, an overview of the architecture of online handwritten text recognition system is illustrates. Section 3 highlights the characteristic of Arabic

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

language and then followed by presenting the current Arabic Database in section 4. QWH database in its structures is explained in details in section 5. Section 6 presents the output of applying the preprocessing phase steps into QWH database. Finally, section 7 highlights the paper summary provides with the future work.

2. PHASES OF ONLINE HANDWRITTEN TEXT RECOGNITION SYSTEM

Most of the online text recognition systems follow the typical structure of pattern recognition systems which basically consists of four major stages [2] [6] which are text acquisition, preprocessing, feature extraction, recognition phase as illustrated in Figure 1. However, some studies provide extra phases or ignore some phases of the systems depending on the nature of the text or due to classifier structure [1]. Each phase of Online recognition system is explained in the following subsections.

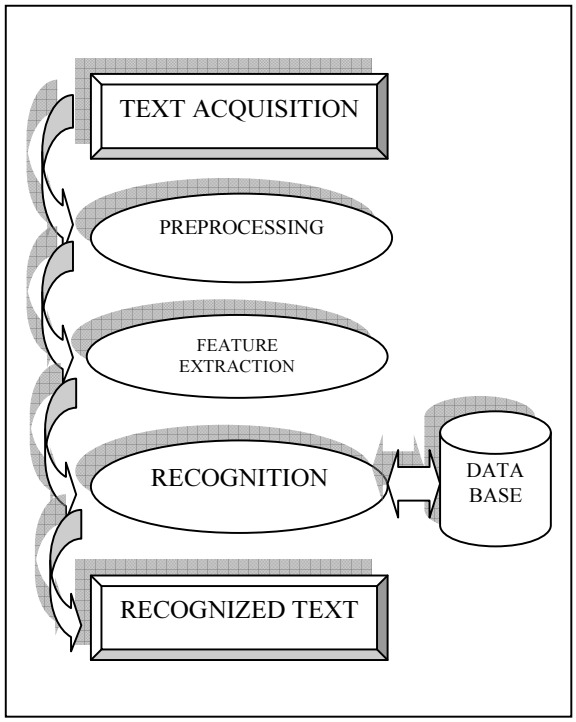

Figure 1: Phases of Online Handwriting Recognition System

2.1 Data acquisition

Online handwriting recognition requires a transducer that captures the writing as it is written. The most common of these devices is the electronic tablet or digitizer. These devices use a pen that is digital in nature. Writing on the acquiring device is the first step in any online handwriting recognition system that collects the sequence of coordinate points (x,y) of the pen movements. The pen movements consist of three actions, namely, Pen-Down, Pen-Move and Pen-Up action. When one press, moves, lifts the pen up consecutively, and more than one point collected, the stroke number is incremented. Pen-Move action stores movements of pen on writing pad. The points of the list are denoted by (x_n,y_n); where n is total number of points in the list [7].

2.2 Preprocessing phase

Preprocessing phase in online handwriting recognition is performed to minimize the noise which may accrue in the handwritten text as mentioned earlier. Preprocessing phase includes several multiple steps and each step does a specific function in this phase. Preprocessing steps could improve the overall recognition rate and considers as one of the essential phases of online handwriting recognition and most of the researchers have discussed its challenges for various scripts from time to time [1]. In section 6, the steps of preprocessing phase those conducted into QWH database will be explained in details.

2.3 Features extraction

Text features are measurements applied to the word or a character and listed together to produce a measurement vector "feature vector". Any character, word, digit, or stroke has its own specific feature vector to identify it from any other characters, words or strokes [6]. For pattern recognition in general and text recognition approaches in particular, extracting an appropriate set of features and an efficient extraction method have been stated as the most important factors in achieving high recognition performance [6].

2.4 Recognition phase

Recognition (classification) phase considers as the heart of the text recognition system and for many pattern recognition approaches such as face, voice, signatures, and image processing recognition systems. Recognition phase is a phase of making the system decision. Based on the extracted features, the Recognition algorithm of the system attempts to identify the pattern that represents the input features. On the other hand, the success rate of any recognition system depends not only on the recognition techniques but it depends on several reasons such as the features extraction technique, the preprocessing stage, or the segmentation step.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

However, until now there is no such thing as the "best recognition method" [2].

3. CHARACTERISTIC OF ARABIC LANGUAGE

Arabic language is spoken by more than 280 million people as a first language in all Arabic countries and by 250 million as a second language worldwide. In the recent years , Arabic language comes as the fifth rank of most commonly used languages in the world and as one of the six official languages of the United Nations [8]. In addition, there are some other languages related to Arabic language. These languages have some similarities with Arabic language whither from the letters shapes or from the pronunciation. These languages are *Jawi, Persian, Urdu, Pashto, Bengali*, and others. These languages are spoken by millions of people in many Islamic countries such as Iran, Afghanistan, Pakistan, parts of India, Bangladesh, Sri Lanka, Malaysia, Indonesia, and other countries [2].

Several reasons make Arabic script different to other languages scripts from the shape and the writing style [1]. Here are some of these reasons:

• First, Arabic 28 characters come in different shapes depending on their position in the word. These positions are beginning, middle, terminate, and isolated shape. Although, the characters (ا, د, ذ, و, ر, ز) come in two shapes which are single and terminate shape only.

• Second, several Arabic characters have the same main body and it can be distinguished from each other by special extra strokes that may be dots or Hamza (ء).

• Third, any Arabic word must be written cursively and the characters must be connected unless one of the characters (ا, د, ذ, و, ر, ز) exist in the middle of the word to make sub-words.

• Fourth, in such a writing way some Arabic words may have overlap case especially in handwritten way. Here, some characters might come in vertical position above their next character. Overlapping may cause an ambiguity to the system to segment the word and then to recognize the proper word.

4. ONLINE ARABIC DATABASES

Providing a large amounts of handwritten data for training and testing purpose, is a fundamental procedure for designing any Online or Offline handwriting recognition system with high recognition accuracy rate. Furthermore, the comparison of various recognition methods has become a focus of attention recently. Therefore, the acquisition and distribution of standard databases has become an important issue in the handwriting recognition research community [9] [10].

Several of handwriting datasets have been published since 1990s. Off-line datasets had been more publically available before Online datasets for handwriting recognition field. Starting from digit or letter datasets to city name, further to sentence and that application fields have expanded from small lexicon domains, such as bank check reading and address recognition, to large lexicon and general unconstrained domains. Constructing datasets is a costly and time consuming process but the availability of public datasets allows researchers to pay more attention to enhance their systems rather than collecting large datasets to work on .

For Arabic the only publically available dataset is the *IFN/ENIT* offline handwritten dataset. It is a dataset of isolated words representing a limited vocabulary lexicon of 937 Tunisian city names [11]. For online Arabic datasets, the same institution introduced a new dataset called ADAB [12] at the tenth International Conference on Document Analysis and Recognition (ICDAR'2009). It has been presented in a handwriting recognition competition for systems output evaluation, but it is not yet launched for public use. The database in version 1.0 consists two parts: the training part and the test part. The training part is composed of 15158 Arabic words handwritten by more than 130 different writers, most of them selected from the narrower range of the *l'Ecole Nationale d'Ing`enieurs de Sfax (ENIS)*. The text written is from 937 Tunisian town/village names. The ADAB-database is split in 3 sets. Details about the number of files, words, characters, and writers for each set 1 to 3 are shown in Table 1. The test part is composed of 1562 files, 2418 words and 12648 characters written by 24 writers [12].

Table 1: Features of ADAB datasets 1, 2, and 3

Set	Files	Words	Characters	Writers
1	5037	7670	40500	56
2	5090	7851	41515	37
3	5031	7730	40544	39

Journal of Theoretical and Applied Information Technology
20th April 2014. Vol. 62 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. QURANIC HANDWRITTEN WORDS DATABASE

QWH database is consisted of 120 common Quranic words those are the most repeating in the Holly Quran. As a way of facilitating the acquiring the handwritten words, the dataset is divided into two sets equally. Each writer was asked to write the 60 words separately without any repetition. Table2 illustrates the words list of QHW database.

The writers were chosen from several countries those who can write Arabic scripts. The age of writers male and female was between 8 to 50 years old with different level of education. However, the majority of them were right hand persons with a few number of left hand persons.

For collecting QWH database, 1.5 GHz core i3 Accer computer tablet has been used in a nature handwriting way. This computer tablet provided with touch screen can easily be used to acquire the Quranic handwritten words with a simple way of normal writing on the touch screen using a special stylus as illustrated in Figure 2. The way of writing on this computer tablet can minimize the noise and errors while recording on the tablet surface.

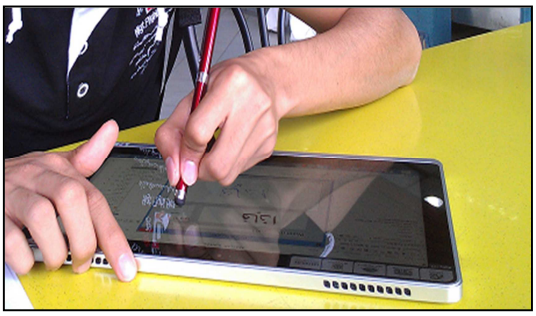

Figure2: Collecting data using Acer tablet

In order to build an Online database for Arabic Handwritten text recognition, a data collection platform is designed using Matlab environment to write each word separately as illustrated in Figure 3. The platform consists of writing area to write the word that appears in the top of the platform in image format. The writer can start writing the words on the area of writing just after writing his/her writer identification code. After writing the word, the writer is asked to press the next button to save the coordinates (x,y) of the handwritten word and then to start writing the next word. If the writer made any mistake or wrote the word wrongly, he/she can reset and clean the writing area by clicking on reset button.

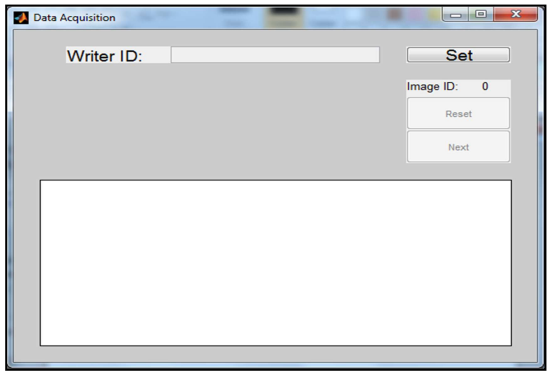

Figure 3: Handwritten words acquisition platform

The stylus movements are distributed at various positions on writing area of the acquisition platform and then joined from the first position (x_1,y_1) to the last (x_n, y_n) to present the appearance of drawn text. The stylus movements are stored in text file starting from the start point (x_1,y_1) until the last point (x_n,y_n) where n is total number of points in the writing movements list. Figure 4 illustrates a sample of stylus movements which have been saved in a text file.

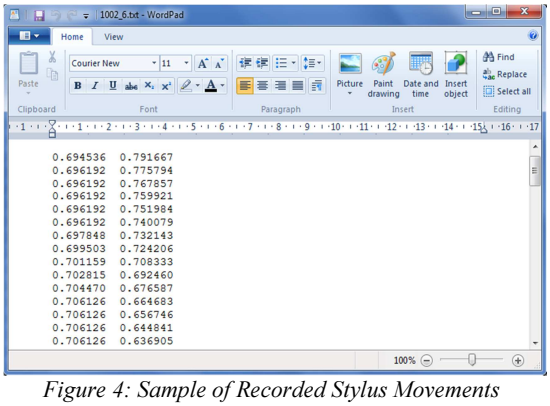

Figure 4: Sample of Recorded Stylus Movements

6. PREPROCESSING PHASE STEPS

Preprocessing phase in online handwriting recognition is performed to minimize the noise which may accrue in the handwritten text as mentioned earlier. Preprocessing phase includes several multiple steps and each step does a specific function in this phase. Preprocessing steps could improve the overall recognition rate and considers as one of the essential phases of online handwriting recognition and most of the researchers have discussed its challenges for various scripts from time to time.

To insure that all the words are standard, clean, and ready to be fed to the next phase, a number of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

preprocessing steps are done for every word of QHW database. The details for each step are explained in the following subsections:

6.1 Word Smoothing

Flickers and zigzag shapes could exist in handwriting lines might accrue due to individual handwriting way or due to the quality of the used acquisition hardware. These noise can be minimized using specific smoothing algorithms in preprocessing phase [13]. In the proposed system, a smoothing technique is used to smooth the handwritten curves called Loess filter. This filter is based on conducting the local regression of the curves points using weighted linear least squares and a second degree polynomial model.

In this technique, each smoothed value is determined locally by neighboring data points defined within the writing curve. The process is weighted and a regression weight function is defined for each data point contained within the writing curve. The local regression smoothing algorithm is presented in following steps for each data point [14].

I. Calculate the regression weights for each data point in the writing curve by the tricube formula:

$$Wi = (1 - |\frac{x - xi}{d(x)}|^3)^3 \quad (1)$$

Where x is the predictor value associated with the response value to be smoothed, x_i is the nearest neighbor of x as defined by the curve, and $d(x)$ is the distance along the *x-axis* from x to the most distant predictor value within the curve. The weights have these characteristics. The data point to be smoothed has the largest weight and the most influence on the fit. Data points outside the curve have zero weight and no influence on the fit.

II. A weighted linear least squares regression is performed. Here for loess method, the regression is based on a second degree polynomial.

III. The smoothed value is given by the weighted regression at the predictor value of interest.

6.2 Word Simplification

Data points simplification is the process of reducing the number of data points acquired by a digital device [15], through removing the redundant points which could inappropriate for pattern classification. This processing directly affects and enhances the recognizer performance [16]. However, *Douglas Peucker's* algorithm [17] was adopted to simplify the acquired handwritten word point sequence.

Basically, Douglas Peucker's algorithm is done by considering an imaginary line between the first and the last point in a set of a curve points. The algorithm then checks which point in between is farthest away from this line. Although, if the point or all other in-between points are closer than a given distance, it removes all these in-between points. However, if this outlier point is farther away from the imaginary line than a specific value "tolerance", the curve is split in two parts. Here, Douglas Peucker's algorithm has been applied with a tolerance of .01 which is determined empirically.

The following steps are explain how Douglas Peucker's algorithm roughly works:

1. Read the points of the handwritten word.

2. Make a polyline between the two endpoints. This is the initial approximation to the simplified Word shape.

3. Find the distance between the edge and the remaining vertices. If the tested vertex is further away from the defined tolerance, then the vertex is added to the simplification points.

4. Repeat step 3 until the end of the points sequences.

6.3 Word Size Normalization

Size of the acquired handwritten word depends on how the writer moves the stylus on writing area. The handwritten words are generally written in different sizes when the stylus is moved along the border of writing area that may cause some ambiguity in the next phases. Size normalization is a necessary step that should be performed in order to recognize any type of text. This can be achieved by converting the acquired handwritten word with assumed fixed size format. The algorithm for size normalization of the acquired handwritten word is given below:

1. Read the coordinates series $S = (x_1,y_1), (x_2,y_2), \ldots (x_n,y_n)$ of the acquired word.

2. Find the maximum point of the series $S (x_{max}, y_{max})$

3. Convert all the series points by

$$S_{new} = S / S(x_{max},y_{max}).$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6.4 Centering of the Word

After resizing the acquired handwritten word, the current coordinates are needed to be shifted to the centering axis (x_0,y_0) to make sure that all handwritten words points are in the equal formatting and all data are translated to the same spot relative to the origin. This step is done using the following algorithm:

1. Read the coordinates series $S = (x_1,y_1), (x_2,y_2), \ldots (x_n,y_n)$ of the resized word.

2. Find the lower point of the series $S (x_{min}, y_{min})$.

3. Convert all the series points to the origin by

$$S_{new} = S - S(x_{min},y_{min})$$

After passing the four steps of preprocessing phase, the handwritten words points are almost in standard formatting. However, in this proposed system, a simple and less number of steps were performed to eliminate the complexity and to minimize any processing delay that may happen.

7. CONCLUSION AND FUTURE WORK

A database consisting of handwritten Qurqnic words has been described in this paper. The database can serve as a basis for research in online recognition of Arabic handwritten text. However, with few preprocessing procedures and without segmentation step, the system successfully gave good results to recognize the handwritten words dealing with the QHW database. The preprocessing phase includes Word Smoothing, Word Simplification, Word Size normalization, and Centering of the word.

QHW dataset is the first online database for Arabic language which contained of Quranic common words. The data collection way is considered to get real, natural, simple and clear database. QHW database consists of 120 handwritten words which divided equally into two sets to be written by 200 writers. The QHW database contains 12000 sample including more than 42,800 characters and 23,300 sub words. Handwritten words were mainly written by variety of people aged between 6 and 50 years from several countries. The dataset is collected using a layout described in details. It saved in text files and each file contains the data of a stylus movements action for writing a specific word.

The main aim for the future research is to enlarge the database to cover as much Quranic words as possible to build comprehensive Arabic lexica vocabulary in online Arabic text recognition. Then to make this database available online throw internet to be shared with other researchers.

Additionally, classifying the writers according to gender, handedness, age and education to study the influence of variations in these categories, can be another point of a future study related to this topic.

REFRENCES:

[1] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Segmentation Techniques for Online Arabic Handwriting Recognition: A survey," In Proceeding of the 3rd International Conference on Information and Communication Technology for the Moslem World: ICT Connecting Cultures, ICT4M 2010, Jakarta, Indonesia, 2010, pp. D37-D40.

[2] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Recognition Techniques for Online Arabic Handwriting Recognition Systems," In Proceeding of the International Conference on Advanced Computer Science Applications and Technologies (ACSAT2012), Kuala Lumpur, Malaysia, 2012.

[3] R. Elanwar, M. Rashwan and S. Mashali, "OHASD: The First On-line Arabic Sentence Database Handwritten on Tablet PC," In Proceeding of the Proceedings of World Academy of Science, Engineering and Technology (WASET), International conference on Signal and Image Processing ICSIP, 2010, pp. 910-915.

[4] F. Faradji, K. Faez and M. S. Nosrati, "Online Farsi Handwritten Words Recognition using a Combination of 3 Cascaded RBF Neural Networks," In Proceeding of the International Conference on Intelligent and Advanced Systems (ICIAS 2007) 2007, pp. 134-138.

[5] F. Biadsy, J. El-Sana and N. Habash, "Online Arabic Handwriting Recognition using Hidden Markov Models," In Proceeding of the 10th International Workshop on Frontiers in Handwriting Recognition, La Baule (France), 2006, pp. 85-90.

[6] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Feature Extraction Techniques of Online Handwriting Arabic Text Recognition," In Proceeding of the 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M), 2013, pp. 1-7.

[7] M. A. Abuzaraida, A. M. Zeki and A. M. Zeki, "Problems of Writing on Digital Surfaces in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Online Handwriting Recognition Systems," In Proceeding of the 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M) 2013, pp. 1-5.

[8] K. Versteegh, M. Eid, A. Elgibali and M. Woidich, Encyclopedia of Arabic language and linguistics: Boston, 2006.

[9] U.-V. Marti and H. Bunke, "The IAM-database: an English Sentence Database for Offline Handwriting Recognition," International Journal on Document Analysis and Recognition, vol. 5, pp. 39-46, 2002.

[10] H. Alamri, J. Sadri, C. Y. Suen and N. Nobile, "A novel comprehensive database for Arabic off-line handwriting recognition," In Proceeding of the Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, ICFHR, 2008, pp. 664-669.

[11] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze and H. Amiri, "IFN/ENIT-database of Handwritten Arabic Words," In Proceeding of the CIFED, 2002, pp. 127-136.

[12] H. Boubaker, et al., "Online Arabic Databases and Applications," in Guide to OCR for Arabic Scripts, ed: Springer, pp. 541-557, 2012.

[13] C. Loader, "Smoothing: Local Regression Techniques," in Handbook of Computational Statistics, ed Berlin Heidelberg: Springer, pp. 571-596, 2012.

[14] L. Clive, Local Regression and Likelihood vol. 47: springer New York, 1999.

[15] K. Jumari and M. A Ali, "A survey and Comparative Evaluation of Selected Off-line Arabic Handwritten Character Recognition Systems," Jurnal Teknologi, vol. 36, pp. 1–18, 2012.

[16] M. Al-Ammar, R. Al-Majed and H. Aboalsamh, "Online Handwriting Recognition for the Arabic Letter Set," Recent Researches in Communications and IT, 2011.

[17] D. David and P. Thomas, "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature," Cartographica: The International Journal for Geographic Information and Geovisualization, vol. 10, pp. 112-122, 1973.

Journal of Theoretical and Applied Information Technology
20th April 2014. Vol. 62 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: QWH database words list

No	Word	No	Word	No	Word	No	Word	No	Word	No	Word
1	من	21	كان	41	لم	61	منهم	81	الله	101	لمن
2	الله	22	بما	42	إذا	62	إنه	82	يشاء	102	دون
3	في	23	قل	43	إذ	63	عليه	83	الدنيا	103	خلق
4	ما	24	الأرض	44	شيء	64	حتى	84	إنما	104	فمن
5	إن	25	ذلك	45	هذا	65	يأأيها	85	ولكن	105	إله
6	لا	26	له	46	كفروا	66	بالله	86	ربهم	106	منه
7	الذين	27	الذي	47	عذاب	67	وهم	87	مما	107	سبيل
8	على	28	هو	48	كنتم	68	أولئك	88	ولو	108	فإذا
9	إلا	29	ءامنوا	49	السماوات	69	إني	89	الحق	109	فما
10	ولا	30	هم	50	الناس	70	وإذا	90	السماء	110	كذلك
11	وما	31	وإن	51	وهو	71	رب	91	منكم	111	منها
12	أن	32	قالوا	52	فإن	72	أم	92	ربنا	112	يؤمنون
13	قال	33	كل	53	عليكم	73	ولقد	93	موسى	113	وقال
14	إلى	34	فيها	54	والذين	74	عليم	94	ربكم	114	العذاب
15	لهم	35	والله	55	الكتاب	75	فيه	95	النار	115	وكان
16	ومن	36	يوم	56	والأرض	76	بل	96	فلما	116	لنا
17	ثم	37	كانوا	57	فلا	77	قد	97	ألا	117	تعملون
18	لكم	38	عن	58	إنا	78	عند	98	مبين	118	لو
19	به	39	ربك	59	بعد	79	قوم	99	أنزل	119	يعلمون
20	أو	40	عليهم	60	وإن	80	قبل	100	ربي	120	رحيم