Was the Quran written by the Prophet Muhammad?

A draft report by H. Sayoud
http://sayoud.net

Was the Quran written by the Prophet Muhammad? To respond to this question, we made several investigations of stylometry, artificial intelligence and pattern recognition.
More details can be found at http://sayoud.net or www.sayoud.net
Here are some simplified examples and results that illustrate our different experiments and findings.
Our apologies for the poor organization of this report: it is only a quick draft and not an article!

That is, in the following sections, 15 different experiments are reported with the corresponding results.

1. Word length frequency based analysis

The first experiment is an investigation on the word length frequency. Herein, we must define some technical terms employed in our paper:
-The word length is the number of letters composing that word.
-The word length frequency F(n) for a specific length 'n', represents the number (in percent) of words composed of n letters each, present in the text.

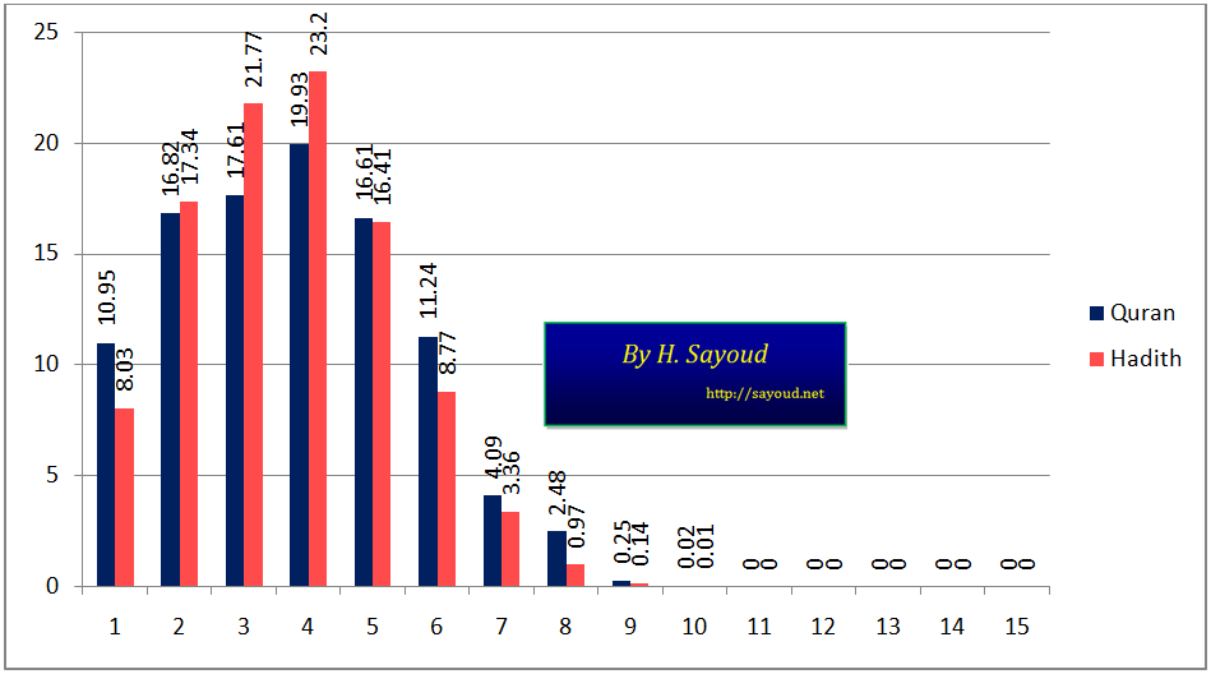

Figure 1.1: Word length frequency in histograms representation.

In figure 1.1, the two spectra are represented simultaneously, which gives an interesting way to compare the two books. So, let us assume that $F_{Quran}(j)$ is the frequency of the words with "j" letters in the Quran and $F_{Hadith}(j)$ is the frequency of the words with "j" letters in the Hadith subset. Then, the observations related to every word length are given here below:

- Length 1: $F_{Quran}(1)=10.95\%$, whereas $F_{Hadith}(1)=8.03\%$; which shows that the words composed of a single letter are much more frequently used in the Quran than in the Hadith subset. For this frequency we notice a great difference between the two books. The Pearson chi-square (uncorrected for continuity) regarding this result is 167.54, involving a probability of consistency $p < 0.0001$, consequently results related to 1-word frequency appear to be significant.

- Lengths 2, 3 and 4: For these cases, the Hadith subset contains many more words than the Quran. We conclude that the Hadith subset uses much more short words than the Quran. The number of short words in the Hadith subset is 62.31%, whereas, in the Quran, it is only 53.76%: namely a difference of 8.55%. The Pearson chi-square (uncorrected for continuity) regarding this result is 468.37, involving a probability of consistency $p < 0.00001$, consequently results related to short-word frequency appear to be significant.
- Lengths 5, 6, 7 and 8: For these cases, the Quran uses much more words than the Hadith subset. The number of long words in the Quran is 34.42%, whereas, in the Hadith subset, it is only 29.51%: namely a difference of 4.91%. The Pearson chi-square (uncorrected for continuity) regarding this result is 198.3, involving a probability of consistency $p < 0.0001$, consequently results related to long-word frequency appear to be significant.
- Lengths 9 and 10: the Quran contains approximately a double number of words with 9 and 10 letters than the Hadith. This fact shows that the Quran vocabulary contains more very-long words (*very-long stands for more than 8 letters*) than the Hadith. The Pearson chi-square (uncorrected for continuity) regarding this result is 10.78, involving a probability of consistency $p < 0.001$. Even though the consistency probability is lower in this case, results related to very-long-word frequency appear to be significant enough.

So, according to all these observations we conclude that the two authors have different styles.

2. Discriminative words

In the second experiment, we look for the words that are present in one book and absent in the other.

Definition of *"word"*: In our investigation, a word represents a sequence of characters linked to form a noun, verb, complement, preposition, or a fusion of a preposition and another word (noun/verb) if they are linked without space.

In this experiment, we analyze all the words present in the Hadith, and try to see if there is any occurrence in the Quran. Similarly, on the other hand, we analyze all the words present in the Quran, and try to see if there is any occurrence in the Hadith. If a word is present in only one book, it will be retained; otherwise it will not be taken into consideration. The word can be a name, verb, complement or a simple expression.

We recall that the part of the Bukhari Hadith contains 23068 tokens and 6225 different words. The Quran contains 87339 tokens and 13473 different words.

Results of this experiment show that 62% of the Bukhari Hadith words are untraceable in the Quran and 83% of the Quran words are untraceable in the Bukhari Hadith (see figures 2.1 and 2.2). Such tokens are called *Discriminant Words* (*we chose this appellation due to the proposed application of discrimination*).

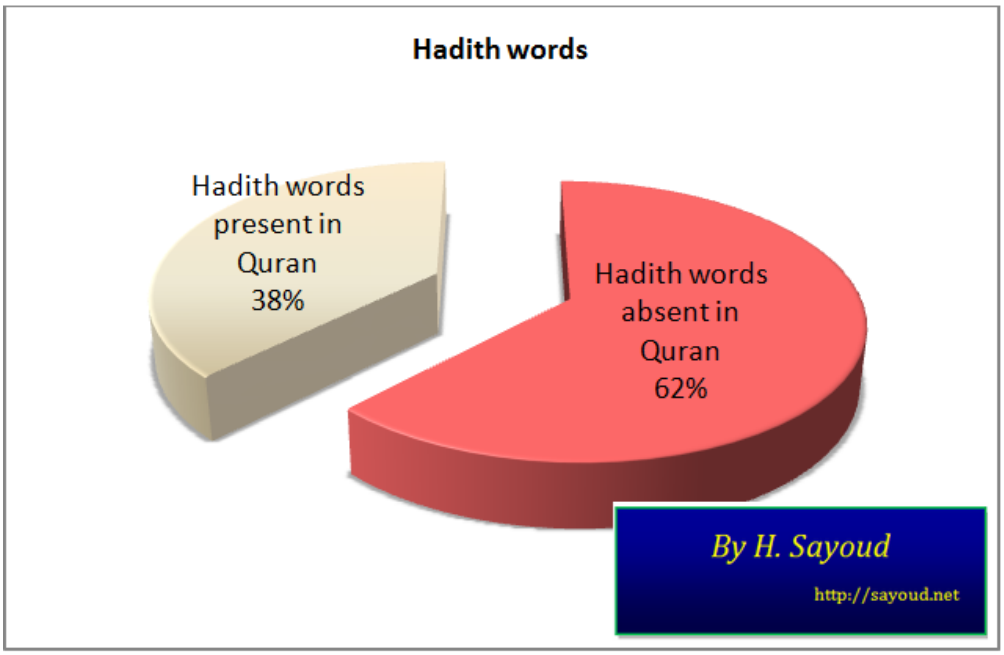

Figure 2.1: Hadith words never used in Quran : 3885 different words (over 6225 total different words contained in Bukhari Hadith) : 3885/6225 =62.41% of words absent in Quran.

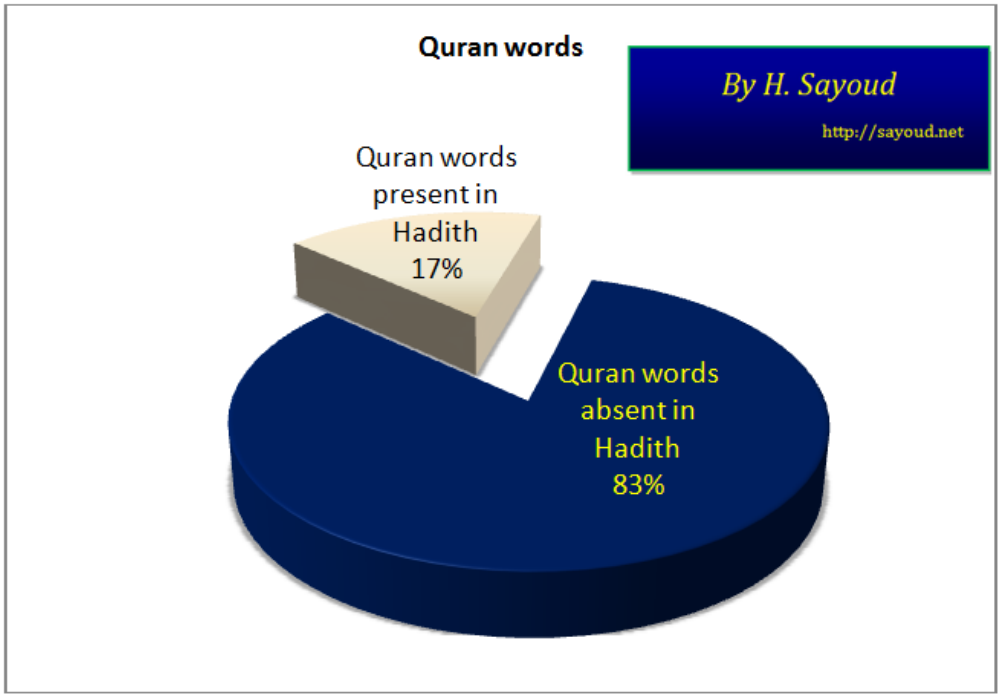

Figure 2.2: Quran words never used in Bukhari Hadith: 11133 different words (over 13473 total different words contained in Quran) : 11133/13473=82.63% of words absent in Hadith.

Observation and Discussion

Practically, it is impossible for a same author to write two books (*related to a similar topic*) with a so great difference in the vocabulary. Therefore, we can deduce that the two books should come from two authors who are characterized by two different vocabularies.

3. Animal citation based analysis

The third experiment investigates the citation of animals in the text.

The animal citation frequency (*freq*) is defined as follows:

freq in % =100.(frequency of occurrence / total number of animal citations)

First observation:

The following table 3.1 shows that for the seven following animals, the difference in citation between the two books is relatively great:

- The name النعم / انعام (*General name of kamels, cows, sheeps*)) is cited 33 times in the Quran, whereas in the Bukhari Hadith it is cited only 2 times;
- The name كلب (*Dog*) is cited only 5 times in the Quran, whereas in the Bukhari Hadith it is cited 13 times;
- The name شاة (*Sheep*) is completely absent in the Quran, whereas in the Bukhari Hadith it is cited 10 times;
- The name دابة (*Animal*) is cited 17 times in the Quran, whereas in the Bukhari Hadith it is cited only 3 times;
- The name الإبل (*Camel*) is cited only 2 times in the Quran, whereas in the Bukhari Hadith it is cited 7 times;
- The name عجل (*Calf*) is cited 10 times in the Quran, whereas in the Bukhari Hadith it is completely absent;
- The name حوت (*Fish*) is cited only 4 times in the Quran, whereas in the Bukhari Hadith it is cited 8 times;

Table 3.1: Citation frequency of some animals appearing more frequently in one book than in the other.

Animal	Translation	Citation in Quran	Citation in Hadith	Frequency in Quran (%)	Frequency in Hadith (%)
انعام / النعم	General name (*kamels, cows, sheeps*)	33	2	21.3	2.13
كلب	Dog	5	13	3.2	13.83
شاة	Sheep	0	10	0.0	10.64
دابة	Animal	17	3	11.0	3.19
الإبل	Camel	2	7	1.3	7.45
عجل	Calf	10	0	6.5	0
حوت	Fish	4	8	2.6	8.51

Second observation:

In table 3.2, we quote the animals that are quoted in the Quran but completely absent in the Bukhari Hadith. There are 29 such animal names.

We remark that several animal names are not cited in the Bukhari Hadith and particularly the name عجل (calf), which is cited 10 times in the Quran and which is completely absent in the Bukhari Hadith.

Table 3.2: Citation frequency of animals that are quoted in the Quran but completely absent in the Bukhari Hadith.

Citations of a frequency of 1 or 2 are not statistically significant.

Animal	Translation	Citation in Quran	Citation in Hadith
عجل	Calf	10	Absent
نملة	Ant	3	Absent
قرد	Monkey	3	Absent
نعجة	Female sheep	3	Absent
ثعبان	Snake	2	Absent
ذباب	Fly	2	Absent
عنكبوت	Spider	2	Absent
جراد	Grasshopper	2	Absent
غراب	Crow	2	Absent
جآن	Fast snake	2	Absent
السبع	Lion	1	Absent
هدهد	hoopoe	1	Absent
حية	snake	1	Absent
طائر	Bird	1	Absent
صافنات جياد	Type of horse	1	Absent
بعوضة	Mosquito	1	Absent
نحل	Bee	1	Absent
ضأن	Lamb	1	Absent
معز	Goat	1	Absent
قمل	Lice	1	Absent
ضفادع	Frog	1	Absent
الهيم	Thirsty camel	1	Absent
البدن	General name (kamels, cows, sheeps)	1	Absent
الأبابيل	Maybe: type of birds	1	Absent
القسورة	Lions	1	Absent
دابة الأرض (الدودة)	Earthworm	1	Absent
العشار	Pregnant camel (+/-)	1	Absent
الوحوش	Wild animals	1	Absent
حمر مستنفرة	Type of wild monkeys or maybe zebras	1	Absent

Third observation:
In table 3.3, we quote the animals that are quoted in the Bukhari Hadith but completely absent in the Quran. There are 11 such animal names.
A particular observation can be done about the name شاة (*sheep*), which is cited 10 times in the Bukhari Hadith and which is completely absent in the Quran.

Table 3.3: Citation frequency of animals that are quoted in the Bukhari Hadith but completely absent in the Quran. *Citations of a frequency of 1 or 2 are not statistically significant.*

Animal	Translation	Citation in Quran	Citation in Hadith
شاة	Sheep	Absent	10
ثور	Bull	Absent	3
هرة	Cat	Absent	2
عصفور	Bird	Absent	2
ضب	Lizard	Absent	2
حمر النعم	Type of red camels	Absent	1
فرس	Horse	Absent	1
كبش	Sheep	Absent	1
ديك	Rooster	Absent	1
دجاجة	Hen	Absent	1
البراق	Miraculous type of horse (Buraq)	Absent	1

Discussion: Results show that there are different animal name citations in the two books. That is, two cases are possible:

- the two books could be related to two topics that are contextually different, citing a contextual type of animal consequently;
- or the two authors should have different stylistic preferences for animal appellations and citations.

However, when we read the two books, we notice that the topics are mainly the same. This fact proposes that the second case is the most probable in this investigation.

4. Special Ending bigrams

This special investigation is made on six ending bigrams, which are often used in Arabic. The bigram consists of a succession of two successive characters in the text.
For example, in the sentence “The cat is here”, the following syllables “Th”, “he“ and “ca” represent bigrams. Also, in the same sentence, the following syllables “he-”, “at-” and “is-” represent ending bigrams, where the “-“ symbol represents a space or a line-feed.
The different bigrams that have been chosen in this investigation are as follows:

"ين-"
"ون-"
"يم-"
"وم-"
"هم-"
"كم-"

Usually, these bigrams (except the 3rd and 4th ones) are often related to the plural form in Arabic.

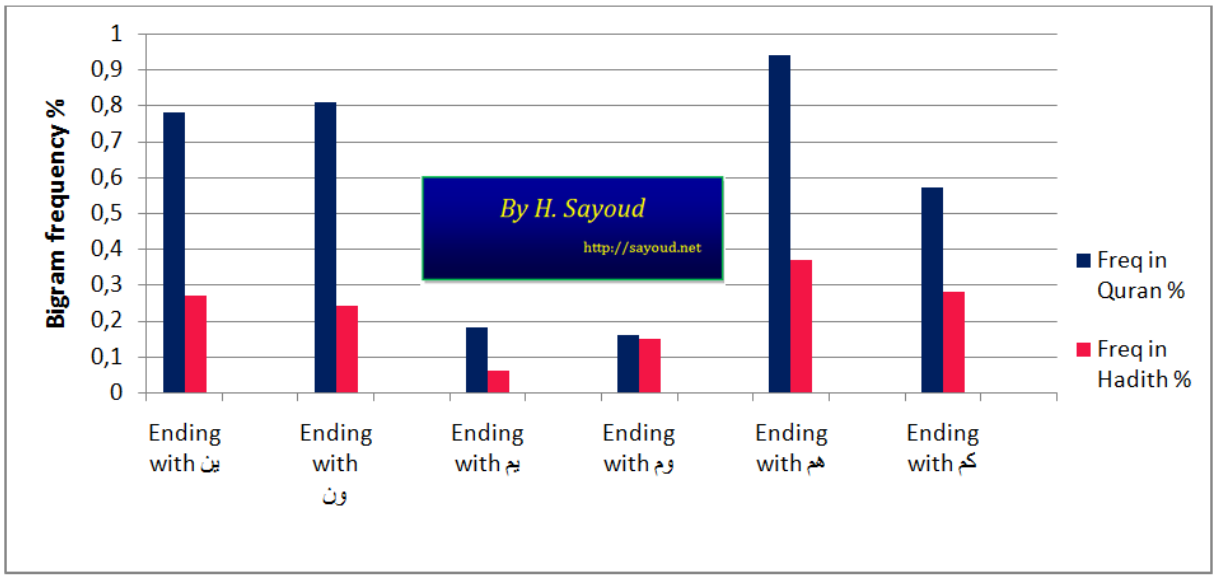

Figure 4.1: Frequency of some ending bigrams.

We notice, in figure 4.1, that there is a great difference in the use of these ending bigrams between the Quran (*where the frequency is relatively high*) and the Bukhari Hadith (*where the frequency is relatively low*), especially for the two first bigrams and the two last bigrams.
This phenomenon can be justified by the fact that the Quran uses much more frequently the plural form in its sentences.
So the authors of the two books appear to have different styles of writing: in the Quran, the plural form is more employed than in Hadith.

5. Segmental analysis (2nd series of experiments)

The fifth experiment analyses the two books in a segmental form: four different segments of texts are extracted from every book and the different texts are analysed and compared.
In such tasks of authorship attribution or discrimination, several linguistic features have been proposed by different researchers. We can quote four main types:
-Vocabulary based Features;
-Syntax based Features;
-Orthographic based features;
-Characters based features.

In this section, the author proposes some types of features and describes five related experiments: an experiment using discriminative words, a word length frequency based analysis, an experiment using the COST parameter, an investigation on discriminative characters and an experiment based on vocabulary similarities.
In these experiments, the different segments are chosen as follows: one segment is extracted from the beginning of the book, another one from the end and the two other segments are extracted from the middle area of the book. A segment size is about 10 standard pages and all the segments are distinct and separated (*without intersection*). These segments are denoted Q1 (*or Quran 1*), Q2 (*or Quran 2*), Q3 (*or Quran 3*), Q4 (*or Quran 4*), H1 (or Hadith 1), H2 (*or Hadith 2*), H3 (*or Hadith 3*) and H4 (*or Hadith 4*). Finally, these eight texts segments are more or less comparable in size.

6. Discriminative words

This sixth experiment investigates the use of some words that are very commonly used in only one of the books. In practice, we remarked that the words: الذين (*in English: THOSE or WHO in a plural form*) and الأرض (*in English: EARTH*) are very commonly used in the four Quran segments; whereas, in the Hadith segments, these words are rarely used, as we can see in the following table.

Table 6.1: Some discriminative words and their frequencies.

Word	Frequency (%) in the Quran segments				Frequency (%) in the Hadith segments			
	Quran 1	Quran 2	Quran 3	Quran 4	Hadith 1	Hadith 2	Hadith 3	Hadith 4
الذين	1.35	1.02	1.12	0.75	0.11	0.03	0.02	0.08
الأرض	0.34	0.63	0.59	0.42	0.23	0.13	0.18	0.15

For الذين the frequency of occurrence is over 0.7% in the Quran segments, but it is between 0.02% and 0.11% in the Hadith segments (*namely almost the 1/10th of the Quran frequency*).
For الأرض the frequency of occurrence is about 0.5% in the Quran segments, but it is between 0.13% and 0.23% in the Hadith segments (*namely about the half*).

These results show that the author of the Quran uses much more frequently these particular words than the Hadith author does.

7. Word length frequency based analysis

The seventh experiment is an investigation on the word length frequency. In the following figure (*figure 7.1*), the different curves (*smoothed curves*), representing the « word length frequency » versus the « word length », show the following two important points:

- The Hadith curves have more or less a gaussian shape that is pretty smooth; whereas the Quran curves seem to be less Gaussian and present some oscillations (*distortions*).
- The Hadith curves are easily distinguishable from the Quran ones, particularly for the lengths 1, 3, 4 and 8: for the lengths 1 and 8, Quran possesses higher frequencies, whereas for the lengths 3 and 4, Hadith possesses higher frequencies.

The statistical consistency of the discrimination between the two groups, using frequency of monograms, trigrams, tetragrams or octograms based words, which is evaluated with Fisher's exact test, corresponds to a probability p of 2.86%.

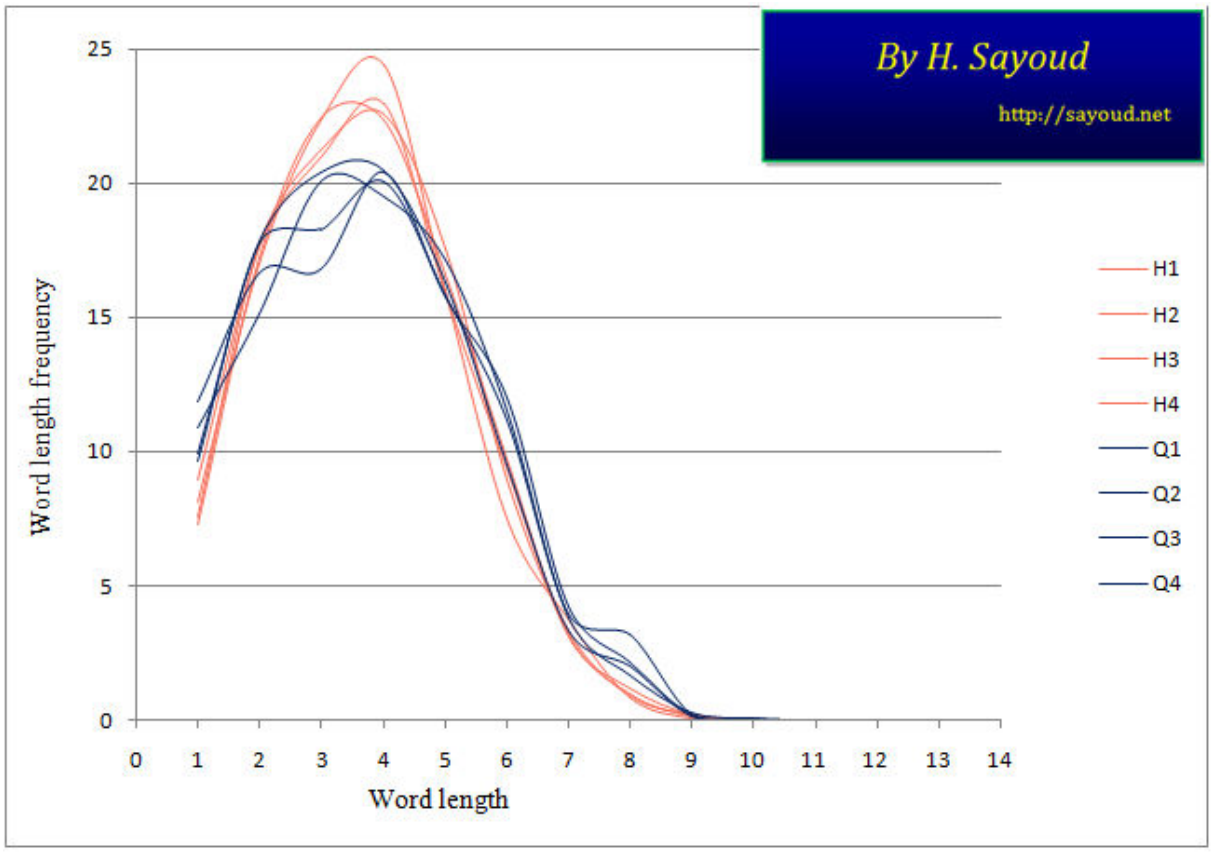

Figure 7.1: Word length frequency (*smoothed lines*).

Although these results cannot be used accurately in authorship discrimination, they can give preliminary information on the sizes of the preferred words by each author. That is, according to these results we should expect that the two text groups correspond to different authors.

8. COST parameter

The eight experiment concerns the new COST parameter which appears non-null only in the Holy Quran, as we can see in table 8.1. The COST parameter is explained in section 4.1.3 of a previous article (see the article I published in LLC journal, in 2012).
In fact, it measures the termination similarity between the neighboring sentences of a text, such as a same final syllable or letter. That is, the COST parameter gives us assessment on the text organization in term of ending structure.
The following table shows the average COST values of the 8 different segments.

Table 8.1: Average COST values for the different segments.

	Quran 1	Quran 2	Quran 3	Quran 4	Hadith 1	Hadith 2	Hadith 3	Hadith 4
$COST_{average}$	2.2	2.6	2.6	2.38	0.46	0.47	0.43	0.47

We notice that the average value of COST is practically constant for all the Quran segments: it is about 2.2 at the beginning of the Quran, 2.4 at the end and it is about 2.6 in the area of the middle.
Similarly, this parameter appears constant for all the Hadith segments: it is about 0.46.
In addition, we notice that the mean values of the COST for Quran and Hadith are very different. This great difference involves distinctive writing styles for the two books (i.e. two different styles concerning the sentence ending).

9. Discriminative characters

The ninth experiment investigates the use of some characters that are very commonly used in only one of the books.
In reality, we limited our investigation to one of the most interesting character, which seems to be very discriminative between the two books: it is the character " و" , which is a consonant and vowel in a same time (*in English, it is equivalent to the consonant W when used as consonant; or the vowel U when used as vowel*).
Furthermore, this character is important because it also represents the preposition AND (*in English*), which is widely used in Arabic.
So, by observing the table below, we notice that this character has a frequency of about 7% in all Quran segments and a frequency of about 5% in all Hadith segments.

Table 9.1: frequency of the character و in the different segments.

Segment	Q1	Q2	Q3	Q4	H1	H2	H3	H4
Frequency of character و	7.73	7.11	6.91	7.04	5.19	5.45	4.72	5.33

This difference in the character frequency shows that the 2 authors do not employ the character و in the same proportion.

10. Vocabulary based similarity

The tenth experiment makes an estimation of the similarity between the vocabularies (*words*) of the two books.

So, in this investigation we propose a new vocabulary similarity measure that we called VSM (*ie. Vocabulary Similarity Measure*), which is defined as follows:

VSM (text1, text2) = [number of common words between the 2 texts] / [size(text1) . size(text2)]$^{1/2}$

Typically, in case of 2 identical texts, this similarity measure will have a value of 1 (*ie. 100%*). Hence, the higher this measure is, the more similar (in terms of vocabulary) the two texts are.
We recall that there are four texts of the Quran and four texts of the Hadith that are more or less comparable in size.
The different inter-measures of similarity are represented in the following matrix (*similarity matrix*), which is displayed in table 10.1.

Table 10.1: Similarity matrix representing the different VSM similarity measures between segments.

VSM in %	H1	H2	H3	H4	Q1	Q2	Q3	Q4
H1	100	32.89	31.43	28.22	20.93	19.86	19.38	19.86
H2	32.89	100	31.37	29.23	20.84	19.99	18.63	19.45
H3	31.43	31.37	100	29.17	19.77	19.88	18.90	18.96
H4	28.22	29.23	29.17	100	19.93	18.68	18.55	18.79
Q1	20.93	20.84	19.77	19.93	100	29.73	29.56	24.49
Q2	19.86	19.99	19.88	18.68	29.73	100	34.88	25.22
Q3	19.38	18.63	18.90	18.55	29.56	34.88	100	27.09
Q4	19.86	19.45	18.96	18.79	24.49	25.22	27.09	100

We notice that all the diagonal elements are equal to 100%. We do remark also that all the Q-Q similarities and H-H similarities are high, relatively to Q-H or H-Q ones (*Q stands for a Quran segment and H stands for a Hadith segment*). This means that the 4 segments of the Quran have a great similarity in vocabulary and the 4 segments of the Hadith have a great similarity in vocabulary, too. On the other hand it implies a low similarity between the vocabulary styles of the two different books. This deduction can easily be made from the following simplified table, which represents the mean similarity measure between one segment and all the segments of a given book.

Table 10.2 gives the mean similarity according to Quran or Hadith for each segment X ($X=Q_i$ *or* $X=H_i$, $i=1..4$), which can be expressed as the average of all the similarities between segment X and the different segments of a same book. This table is displayed in order to see if a segment is more similar to the Quran family or to Hadith family.

Table 10.2: Mean VSM similarity in % between one segment and the different segments of a same book.

	Mean Similarity with H segments	Mean Similarity with Q segments
H1	30.85	20.01
H2	31.16	19.73
H3	30.66	19.38
H4	28.87	18.99
Q1	20.37	27.92
Q2	19.60	29.94
Q3	18.87	30.51
Q4	19.27	25.60

Similarly, we remark that the intra-similarities (*within a same book*) are high: between 26% and 31%; and that the inter-similarities (*segments from different books*) are relatively low: not exceeding 20%. This observation shows that all the segments of a same book appear to have a unique origin and that the two books should have two different author styles.

11. Fully Automatic Authorship Attribution with several features and several classifiers

The eleventh experiments, which consists in an automatic authorship attribution, analyses the two books in a segmental form by using several features (words, word n-grams, characters, character n-grams and dis-legomena) and several classifiers (Camberra distance, Cosine distance, RN cross entropy, Histogram distance, Intersection distance, Kullback Leibler distance, Manhattan distance, KS distance, LDA analysis and Naive Bayes classifier,).
The sizes of the segments are more or less in the same range: four different text segments, with approximately the same size, are extracted from every book.
It concerns two experiments:
- In the first experiment, the first segment of each book is taken as reference. Hence there will be two reference texts, one representing the Quran author and the other representing the Hadith author. The six remaining texts (*3 for each book*) have to be classified into Quran class or Hadith class.
- The second experiment is similar to the first one except that the reference texts, here, are represented by the second segments of the two books respectively.
Concerning the number of selected examples (*an example refers to a word, character, etc.*), we have considered 2 cases: in the first case, we consider all the examples and in the second case, we keep only the 50 most frequent ones.
Note: in the following paragraphs, a score of 100% means that all the Quran segments are classified as Quran class and all the Hadith segments are classified as Hadith class, without any error of attribution.

11.1 First experiment

In the following investigation, we consider the segments Q1 and H1 as reference texts for the Quran and Hadith, respectively. Then, Q2, Q3, Q4, H2, H3 and H4 will be considered as unknown texts to be classified according to Quran class or Hadith class. During the feature extraction step, two cases are possible: employing all the features or employing the most frequent ones.
In this experiment, all the text segments have to be classified into two classes: Quran class or Hadith class. Classification results (*displayed in %*) are reported in table 11.1.

Table 11.1: Precision of good classification of the different segments with several features and several classifiers

Feature / Classifier	Charac. Bigram	Charac -ter	Charac. Tetra-gram	Charac. Tri-gram	Dis Lego-mena	Word	Word Bi-gram	Word Tri-gram	Word Tetra-gram
Number of features	*All*	*all*	*all*	*all*	*All*	*all*	*50 most freq.*	*50 most freq.*	*50 most freq.*
Camberra distance	100%	50%	83%	100%	100%	100%	100%	100%	100%
Cosine distance	100%	100%	100%	100%	100%	100%	100%	100%	100%
RN cross entropy	100%	83%	100%	100%	100%	100%	100%	100%	100%
Histogram distance	83%	100%	100%	83%	100%	100%	100%	100%	100%
Intersection distance	100%	50%	100%	100%	100%	100%	100%	100%	100%
Kullback Leibler dist	83%	83%	100%	100%	100%	100%	100%	100%	100%
Manhattan distance	100%	100%	100%	100%	100%	100%	100%	100%	100%
LDA analysis	83%	100%	100%	83%	100%	100%	100%	100%	100%

This experiment employing several features (*words, word n-grams, characters, character n-grams and dis-legomena*) and several classifiers (*Camberra distance, Cosine distance, RN cross entropy, Histogram distance, Intersection distance, Kullback Leibler distance, Manhattan distance, KS distance, LDA analysis and Naive Bayes classifier*), shows clearly that the 4 Quran segments should belong to a same author, the 4 Hadith segments should belong to the same author too and that these two authors are likely to be different.

11.2 Second experiment

In the following investigation, we consider the segments Q2 and H2 as reference texts for the Quran and Hadith, respectively. Then, Q1, Q3, Q4, H1, H3 and H4 will be considered as unknown texts to be classified according to Quran class or Hadith class. As previously, during the features extraction step, two cases are possible: employing all the features or employing the most frequent ones.

Table 11.2: Precision of good classification of the different segments with several features and several classifiers

Feature / Classifier	Charac. Bigram	Charac -ter	Charac. Tetra-gram	Charac. Tri-gram	Dis Lego-mena	Word	Word Bi-gram	Word Tri-gram	Word Tetra-gram
Number of features	*All*	*all*	*all*	*All*	*All*	*all*	*50 most freq.*	*50 most freq.*	*50 most freq.*
Camberra distance	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Cosine distance	100 %	100 %	100 %	100%	100 %	100 %	100 %	100 %	100 %
RN cross entropy	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	83 %
Histogram distance	100 %	100 %	100 %	100%	100 %	100 %	100 %	100 %	100 %
Intersection distance	100 %	50 %	100 %	100%	100 %	100 %	100 %	100 %	100 %
Kullback Leibler dist	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	83 %
Manhattan distance	100 %	100 %	100 %	100%	100 %	100 %	100 %	100 %	100 %
LDA analysis	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

Also, in this experiment all the text segments have to be classified into two classes: Quran class or Hadith class. Results of good classification, displayed in %, are reported in table 11.2.

As in the first investigation, this experiment employing several features (*words, word n-grams, characters, character n-grams and dis-legomena*) and several classifiers (*Camberra distance, Cosine distance, RN cross entropy, Histogram distance, Intersection distance, Kullback Leibler distance, Manhattan distance, KS distance, LDA analysis and Naive Bayes classifier*), shows clearly that the 4 Quran segments should belong to the same author, the 4 Hadith segments should belong to the same author too and that these two authors are very probably different.

Discussion on these two experiments: According to these two experiments, we can clearly see that the classification accuracy for the two books is 100% with almost all features and all classifiers. Consequently, we can statistically state that the two investigated books have two different authors or at least two different styles.

12. Hierarchical clustering based classification

In this section we will use a hierarchical clustering (with an average linkage) and try to separate the 8 segments into different appropriate clusters (*hopefully 2 clusters*), thanks to different features, which are denoted in this investigation as follows:

- *F1= frequency of the word (الذين)*
- *F2= frequency of the word (الأرض)*
- *F3= frequency of words with a length of 1 character*
- *F4= frequency of words with a length of 2 characters*
- *F5= frequency of words with a length of 3 characters*
- *F6= frequency of words with a length of 4 characters*
- *F7= frequency of words with a length of 5 characters*
- *F8= frequency of words with a length of 6 characters*
- *F9= frequency of words with a length of 7 characters*
- *F10= frequency of words with a length of 8 characters*
- *F11= frequency of words with a length of 9 characters*
- *F12= frequency of the character (و)*

- *F13= COST value*
- *F14= Average Vocabulary Similarity Measure with regards to the Quran*
- *F15= Average Vocabulary Similarity Measure with regards to the Hadith*

The employed distance is the cosine distance and all the previous result scores are fused into a global feature vector for each segment. These feature vectors are used as input vectors for the hierarchical clustering.
The result of this hierarchical clustering is given by the following dendrogram (*see figure 12.1*), which illustrates the different possibilities of clustering with their corresponding distances in a graphical way.

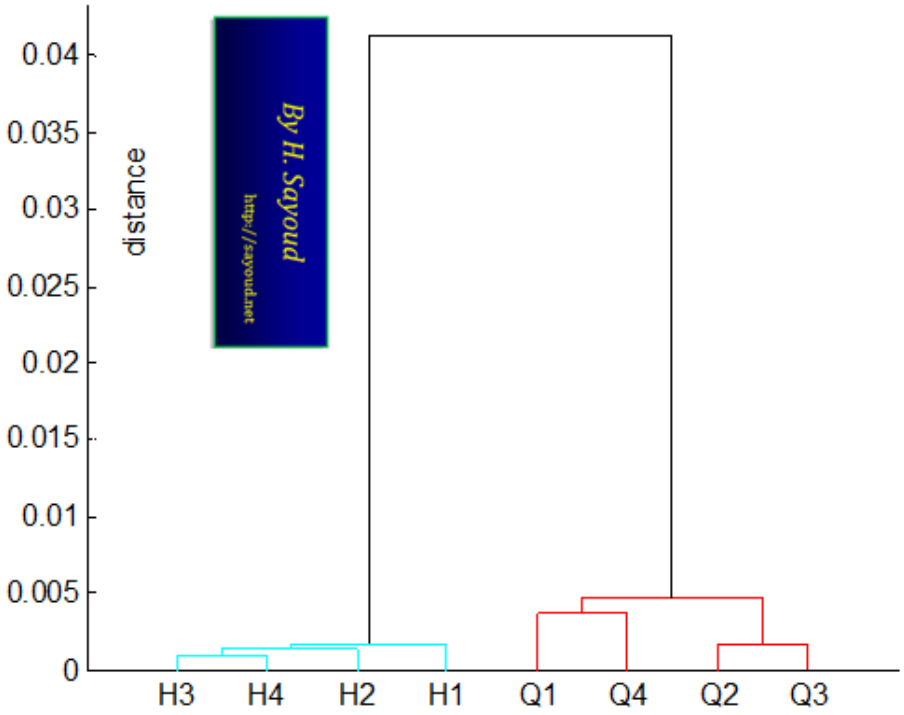

Figure 12.1: *Dendrogram of the different clusters with the corresponding distances.*
Q1..Q4 represent the 4 segments of the Quran and H1..H4 represent the 4 segments of the Hadith. Note that the last cluster (big one) is inconsistent, which involves the existence of two main classes.

The inconsistencies of the different clusters are respectively (*from the first cluster to the last one*): 0, 0.7, 0, 0.7, 0, 0.85 and 1.15.
As we can see, the last cluster has an inconsistency parameter greater than 1 (*inconsistency of 1.15*), while all the other clusters do not exceed 0.85.
Moreover, even by observing the dendrogram of figure L.2, we can easily notice that the hierarchical clustering has revealed two main classes: one class grouping the 4 Hadith segments (*in the left side of the dendrogram: ligh blue*) and a second class grouping the 4 Quran segments (*in the right side of the dendrogram: red color*).

This new result involves two important conclusions:
- First, the two books Q and H should have different authors;
- Second, the 4 segments of each book seem to have a same author (*the presumed author of the book*).

13. Authorship Attribution Results on 25 Text Segments: a segmental analysis

In this investigation, there are 25 different text segments of about 2080 words each, consisting of 11 Hadith segments and 14 Quran segments. In these experiments, 3 segments of the Hadith and 3 other segments of the Quran are used for the training and the remaining segments (*8 Hadith segments and 11 Quran segments*) are used for the testing. Therefore, there are 19 different segments to identify according to 2 referential Authors (*Quran Author or Hadith Author*).

Note: in the following paragraphs, an attribution error of 0% means that all the Quran segments are classified as "*Quran class"* and all the Hadith segments are classified as "*Hadith class"*, without any error of attribution. In fact the attribution error is defined as the ratio of the number of false attributions over the total number of testing segments (*see equation 13.1*).

$$attribution\ error = \frac{number\ of\ false\ attributions}{total\ number\ of\ testing\ segments} \quad (13.1)$$

Table 13.1: Attribution error in % for the different text segments.
There are 11 segments for the Hadith (8 testing + 3 reference) and 14 for the Quran (11 testing + 3 reference)

Feature / Classifier	Charac. Bigram	Charac. Tri-gram	Charac. Tetra-gram	Word Bi-gram	Word Tri-gram	Word Tetra-gram	Word	Rare words (freq=1.. 3)
Number of features	All	All	All	50 most freq.	50 most freq.	50 most freq.	All	All
SMO-SVM	0%	0%	0%	0%	0%	0%	0%	0%
Linear Regression	0%	0%	0%	0%	0%	0%	0%	0%
MLP	0%*	0%*	0%*	0%	0%	0%	0%*	0%*
Stamatatos distance	0%	0%	0%	0%	0%	5.3%	0%	0%
Canberra distance	0%	0%	0%	0%	0%	10.5%	0%	0%
Cosine distance	0%	0%	0%	0%	0%	0%	0%	0%
RN cross entropy	0%	0%	0%	0%	5.3%	0%	0%	0%
Intersection distance	-	0%	0%	0%	0%	0%	5.3%	0%
Manhattan distance	0%	0%	0%	0%	0%	0%	0%	0%

** : means that only the 500 most frequent features are employed*
- : means a classification failure

By observing the above table (*table 13.1*), we can notice that all Quran segments are attributed to the referential "*Quran Author*" and all Hadith segments are attributed to the referential "*Hadith Author*". That is, the 19 different text segments are classified into 2 main classes: "*Quran class"* and "*Hadith class"*, with 0% classification error. From this result, we can deduce that the 2 religious books should have 2 different authors (*or at least 2 different writing styles*) and that every book should be written by one author (*or at least one writing style*).

14. INVESTIGATION BASED ON A CORPUS OF SEVEN RELIGIOUS BOOKS

In this investigation, there are seven different books written by seven different authors: 5 books are contemporary and 2 others are ancient (*dating from the 6th century*). We called this dataset: *SAB-1* (*Seven Arabic Books – dataset One*). These books are described as follows:

1st book: the holy Quran (*author: God (Allah)*), it is considered as the divine book of Islam. The Quran is written by Allah (God) and only sent down to his Prophet Muhammad fourteen centuries ago. This divine book has been delicately conserved by the different scholars over the time. The holy Quran is considered as the first reference of Islam since it contains the authentic speech and statements of God (Allah);

Fig. 14.1: Old pages of the *holy Quran*

2nd book: the Hadith (*author: the Prophet Muhammad*) contains the statements of the Prophet Muhammad in different situations. Muhammad was born in Mecca in the 6th century, became Prophet at the age of 40 and died at the age of 63. In this investigation, we used the Bukhari Hadith book, which is considered as one of the most confident book of the Hadith;

Fig. 14.2: Old pages of the *Hadith*

3rd book: text collection of Alghazali (*Author: Mohammed al-Ghazali al-Saqqa*): it contains some articles and dissertations of Alghazali. This author is a contemporary Egyptian religious scholar, who is born in 1917 and died in 1996. Sheikh al-Ghazali held the post of Chairman of the Academic Council of the International Institute of Islamic Thought in Cairo.

Fig. 14.3: The third author: *Alghazali*

4th book: text collection of Alquaradawi (*Author: Yusuf al-Qaradawi*): it contains some articles and dissertations of Alquaradawi. This author is a contemporary Egyptian/Qatari religious scholar, who is born in 1926. He is the head of the European Council for Fatwa and Research, an Islamic scholarly entity based in Ireland. He also serves as the chairman of International Union for Muslim Scholars (*IUMS*).

Fig. 14.4: The fourth author: *Alquaradawi*

5th book: text collection of Abdelkafy (*Author: Omar Abdelkafy*). This text collection contains some articles and dissertations of Dr. Omar Abdelkafy, who was born in Almenia, Egypt on May 1, 1951. He memorized the Holy Quran completely when he was ten years old. Dr. Abdelkafy also memorized Sahih Al-Bukhary and Muslim with full references. Abdelkafy studied Islamic Theology and Arabic Linguistics from clever scholars and started serving the Islamic Dawah in 1972.

Fig. 14.5: The fifth author: *Abdelkafy*

6th book: text collection of Al-Qarni (*Author: Aaidh ibn Abdullah al-Qarni*). This text collection contains some articles and dissertations of Shaykh Aaidh ibn Abdullah al-Qarni, who was born in 1960. He is a Saudi religious scholar and author of a famous book. Al-Qarni is best known for his distinguished book "La Tahzan" (*in English: Don't Be Sad*), which had a lot of success over the time.

Fig. 14.6: The sixth author: *Al-Qarni*

7th book: text collection of Amr Khaled (Author: Amr Mohamed Helmi Khaled). Several articles and dissertations of Amr Khaled have been collected into a unique text. This author was born in 1967 in Egypt. He is an Egyptian Muslim activist and television preacher. He is often described as "the world's most famous and influential Muslim television preacher." Amr Khaled has recently been chosen as one of the world's 100 most influential people by Time Magazine.

Fig. 14.7: The seventh author: *Amr Khaled*

The experimental results concerning the authorship attribution of those 7 books, by using some fusion techniques, are reported on the following tables (table 14.1 and 14.2).

Table 14.1: Author identification score using the *feature-based fusion*

Total Identification score on the 7 books	The holy Quran book	The Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
100%	100%	100%	100%	100%	100%	100%	100%

Table 14.2: Author identification score using the *classifier-based fusion*

Total Identification score on the 7 books	The holy Quran book	The Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
100%	100%	100%	100%	100%	100%	100%	100%

These fusion results show that the 7 books should be written by 7 different authors (or at least different styles).

15. Visual Analytics: FCM and Hierarchical Clustering based Authorship Classification

Recent work in progress (when I wrote this report)…

Here are our first results we got by using a Fuzzy FCM clustering and a Hierarchical clustering, which show the presence of 2 clusters, when analyzing the 25 text segments consisting of: -14 Quran text segments and -11 Hadith segments. We can easily see 2 clusters (in figure 15.1 and 15.2): the right one in red grouping all the Quran segments and the left one in blue grouping all the Hadith segments.

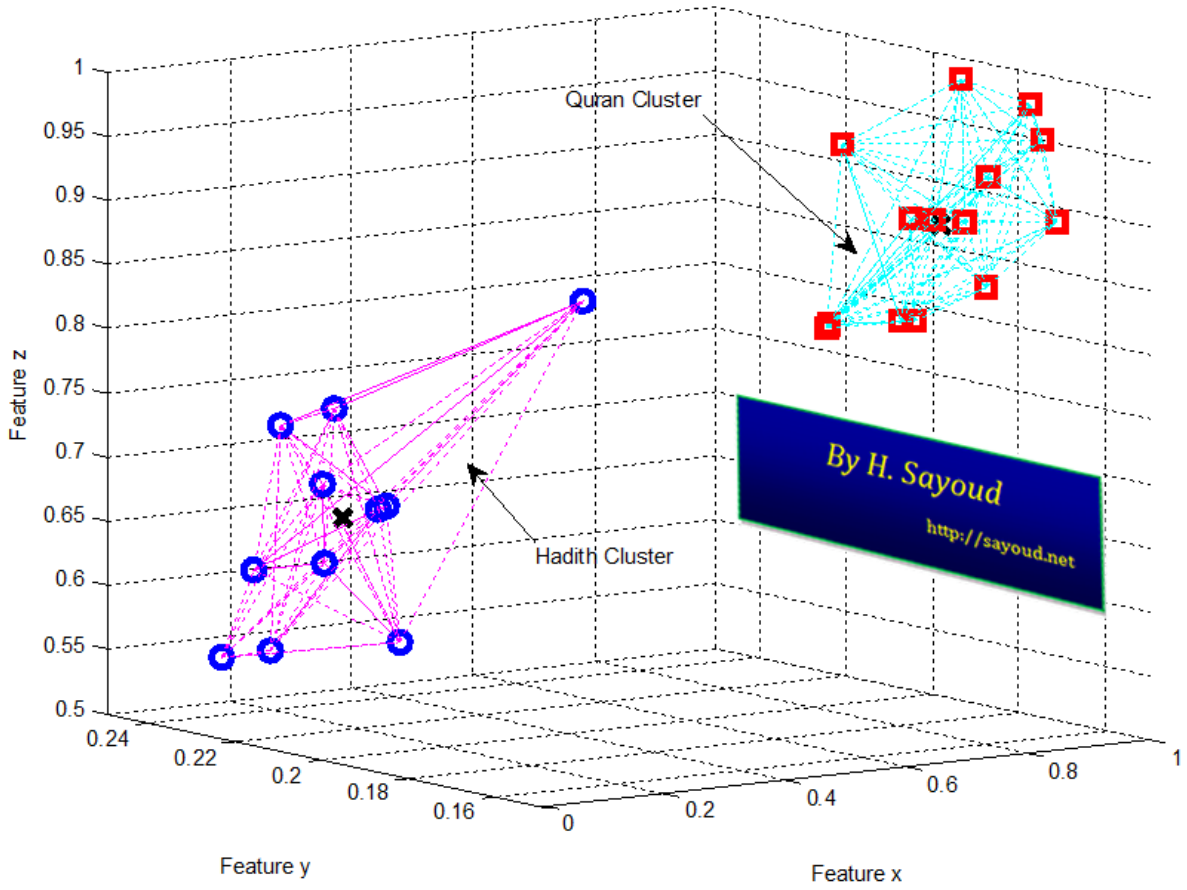

Figure 15.1: Fuzzy FCM Clustering. We can easily see 2 clusters: the right one in red grouping all the Quran segments and the left one in blue grouping all the Hadith segments.

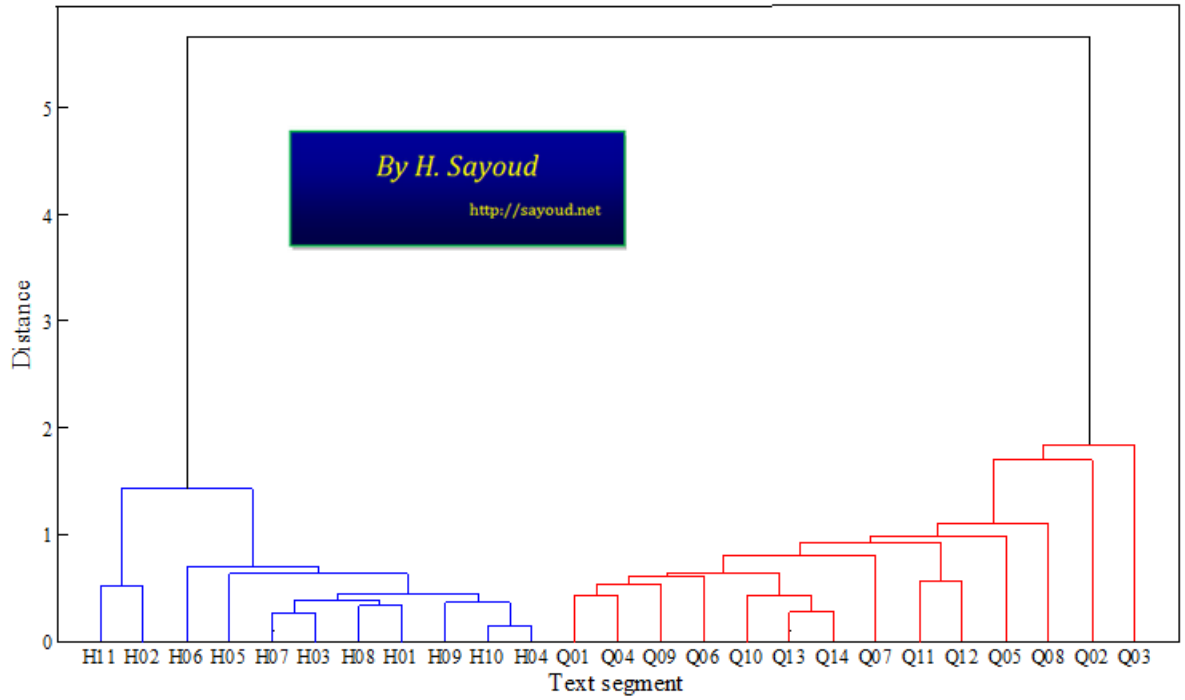

Figure 15.2: Automatic Hierarchical Clustering. We can easily see 2 clusters: the right one in red grouping all the Quran segments and the left one in blue grouping all the Hadith segments.

These last results are in progress and not published yet (*when I wrote this report*)…

16. Discussion and Conclusion

The present report describes the different investigations we performed on the Quran and Bukhari Hadith. It summarizes some of our previous research works on the topic of Quran/Hadith authorship discrimination, in a brief and simplified way.
That is, several series of experiments have been made and 15 investigations are reported in this draft report.
After observing all the experimental results and since the two books appear to have the same genre and theme, it would be reasonable to deduce the following conclusions:

- The two books should have two different authors (or at least two different author styles);
- All the text segments that have been extracted from a unique book (from the Quran only, or from the Hadith only) should probably belong to the same author.

Consequently, the supposition that the Quran was written by the Prophet Muhammad is statistically rejected. Muslims believe that it is written by Allah (God) and sent to his messenger (the prophet Muhammad). Without entering in theological debates, the present investigation gives us a new scientific result reinforcing what had been stated by Muhammad: "the Quran has been transmitted to him".

***Note**: All comments are welcome...*
By Prof H. Sayoud *http://sayoud.net*

August 2014