Author discrimination between the Holy Quran and Prophet’s statements

Halim Sayoud
Department of Electronics and Informatics, USTHB University, Algiers, Algeria

Abstract

Author discrimination consists of checking whether two texts are written by the same author or not. In this investigation, we try to make an author discrimination between the Quran (The holy words and statements of God in the Islamic religion) and the Hadith (statements said by the prophet Muhammad). The Quran is taken in its entirety, whereas for the Prophet’s statements, we chose only the certified texts of the Bukhari book. Thus, three series of experiments are done and commented on. The first series of experiments analyses the two books in a global form (the text of every book is analyzed as a unique big text). It concerns nine different experiments. The second series of experiments analyses the two books in a segmental form (four different segments of text are extracted from every book). It concerns five different experiments. The third series of experiments makes an automatic authorship attribution of the two books in a segmental form by employing several classifiers and several types of features. 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 five different experiments. This investigation sheds light on an old enigma, which has not been solved for 14 centuries: in fact, all the results of this investigation have shown that the two books should have two different authors.

1 Introduction and Related Works

Individuals have distinctive ways of speaking and writing, as explained by Corney (2003), and there exists a long history of linguistic and stylistic investigation into authorship attribution (Holmes, 1998). In recent years, practical applications of authorship attribution have grown in areas such as intelligence (linking intercepted messages to each other and to known terrorists), criminal law (identifying writers of ransom notes and harassing letters), civil law (copyright and estate disputes), and computer security (tracking authors of computer virus source code). As reported by Madigan (2005), this activity is part of a broader growth within computer science of identification technologies, including biometrics (retinal scanning, speaker recognition, etc), cryptographic signatures, intrusion detection systems, and others.

In this research work, we deal with a religious enigma, which has not been solved for 14 centuries, as mentioned by Sayoud in 2010. In fact, several doubts on the origins of the Holy Quran do exist; some persons thought that the Holy Quran could be
an invention of the prophet Muhammad, for three purposes (Al-Shreef, 2009):

1. To facilitate his domination over his followers;
2. To frighten the unbelievers and those who disobey his orders; and
3. To permit his pleasures.

Several theologians, over time, tried to prove that this assumption was false. They were relatively logical and clever, but their proofs were not so convincing for many people, due to a lack in scientific rigor.

Similarly, for the Christian religion, there exist several disputes about the origin of some texts of the Bible. Such disputes are very difficult to solve due to the delicacy of the problem, the religious sensitivity and because the texts were written a long time ago.

One of the purposes of stylometry is authorship attribution, which is the determination of the author of a particular piece of text for which there is some dispute about its writer, as reported by Mills (2003).

Hence, it can be seen why Holmes (Mills, 2003) pinpointed that the area of stylistic analysis is the main contribution of statistics to religious studies. For example, early in the nineteenth century, Schleiermacher disputed the authorship of the Pauline Pastoral Epistle 1 Timothy (Mills, 2003). As a result, other German speaking theologians, namely, F. C. Baur and H. J. Holtzmann, initiated similar studies of New Testament books (Mills, 2003).

In such problems, it is crucial to use rigorous scientific tools and it is important to interpret the results very carefully.

Hence, knowing that authors possess specific stylistic features that make them differentiable (Jiexun, 2006), we tried to make some experiments of author discrimination between the Quran and some Prophet’s statements in order to show that the Quran was not written by the Prophet Muhammad, if the results of these techniques confirm that supposition (Al-Shreef, 2009).

The second series of experiments analyses the two books in a segmental form: four different segments of text are extracted from every book.

In the third series of experiments, the author makes an automatic authorship attribution based on a Multi-Classifier and Multi-Feature segmental analysis using the JGAAP toolbox (Juola, 2009), in order to make an automatic classification of the different segments extracted from the two books.

The article is organized as follows: Section 2 gives a description of the two books to be compared. Section 3 poses the following philosophical problem: Was the Quran invented by the prophet Muhammad? Section 4 describes the different experiments of author discrimination. Finally, an overall discussion is given at the end of the article.

2 Description of the Two Books

Herein, we will give a brief description on the two books that are investigated in our experiments, namely: the Quran and Hadith.

2.1 The Quran

The Quran (in Arabic: القرآن, literally ‘the recitation’; also sometimes transliterated as Qur’an, Koran, Alcoran or Al-Qur’ân (Nasr, 2007; Wiki1, 2012) is the central religious text of Islam. Muslims believe the Quran to be the book of divine guidance and direction for mankind (Ibrahim, 1996) (that has been written by God), and consider this Arabic book to be the final revelation of God. Islam holds that the Quran was written by Allah (i.e. God) and transmitted to Muhammad by the angel Gibraile (Gabriel) over a period of 23 years. The beginning of Quran apparition was in the year 610 (after the birth of Christ).

2.2 The Hadith

Hadith (in Arabic: الحديث, transliteration: al-hadith is the oral statements and words said by the Islamic prophet Muhammad (Pbuh) (Islahi, 1989; Wiki2, 2012). Hadith collections are regarded as important tools for determining the Sunnah, or Muslim way of life, by all traditional schools of jurisprudence. In Islamic terminology, the term hadith refers to
reports about the statements or actions of the Islamic prophet Muhammad, or about his tacit approval of something said or done in his presence (Islahi, 1989; Wiki2, 2012). The text of the Hadith (matn) would most often come in the form of a speech, injunction, proverb, aphorism or brief dialogue of the Prophet whose sense might apply to a range of new contexts. The Hadith was recorded from the Prophet for a period of 23 years between 610 and 633 (after the birth of Christ).

3 Was the Quran Invented by the Prophet Muhammad?

Muslims believe that Muhammad was only the narrator who recited the sentences of the Quran as written by Allah (God), but not the author. See what Allah (God) says in the Quran book: ‘O Messenger (Muhammad)! transmit (the Message) which has been sent down to you from your Lord. And if you do not, then you have not conveyed his Message. Allah will protect you from people. Allah do not guide the people who disbelieve’ [5:67]. Some doubts about the origins of the Holy Quran trying to find a human source for this book do exist. Such assumptions suppose that the Holy Quran is an invention of the prophet Muhammad as reported by Al-Shreef (2009).

For a long time, different scientists tried to present strong context-based demonstrations showing that this assumption is impossible.

The purpose of our research work is to conduct a text-mining-based investigation in order to see whether the two concerned books could statistically belong to the same author or not; i.e. authorship discrimination (Mills, 2003) (Tambouratzis, 2000, 2003), regardless of the literal style or context.

4 Experiments of Author Discrimination

In this section, three series of experiments of author discrimination, between the two books, are investigated and commented.

4.1 First series of experiments: global analysis

The first series of experiments analyses the two books in a global form (the text of every book is analyzed as a unique big text). It concerns nine experiments:

4.1.1 Summary on the size of the two books

The first experiment consists in a computational investigation, which summarizes the size of the two books in terms of words, tokens, pages, etc. The different statistical characteristics of the two books are summarized as follows:

1. Number of tokens in the Quran = 87,341;
2. Number of tokens in the Hadith = 23,068;
3. Number of different words in the Quran = 13,473;
4. Number of different words in the Hadith = 6,225;
5. Ratio of the number of Quran tokens/number of Hadith tokens = 3.79;
6. Ratio of the number of Quran lines/number of Hadith lines of Bkh = 3.61; and
7. Ratio of the number of different Quran words/number of different Hadith words = 2.16.

The two books seem relatively consistent since the average number of A4 pages is 315 for the Quran book and 87 for the Hadith book. However, the two books do not have the same size; that is why we should proceed with our investigation with care.

4.1.2 Word frequency-based analysis

The second experiment is an investigation on the word frequency. Results are displayed in Fig. 1.

4.1.2.1 Notion of discriminative words. A particular interest concerns the discriminative words, as we can see in Fig. 1.

A discriminative word can be seen as a word that is frequently used in one text and rarely employed in the other, which could represent a sample word that can be used for discriminating the two texts.

For instance, suppose that two authors are asked to write a letter in a same topic. Since every author
has a set of preferred words, one should retrieve some specific words that are commonly employed by one author and almost never used by the other one.

Consequently one could distinguish the two authors (texts) by such discriminative words. That is why that type of words is investigated in this section (Fig. 1).

For the words listed above, the frequencies are relatively different showing a dissimilarity between the books vocabularies. Note that the first word in the left (namely ‘And’), meaning ‘And’ in Arabic, has a very high frequency. It is used with a frequency of about 11% in the Quran and 8% in the Hadith, which involves an average relative difference of about 30%. The second word from the left (namely ‘Those’ in Arabic), meaning ‘Those’, has a relative frequency of 1.12% in the Quran and a very low relative frequency (approximately 0%) in the Hadith, which represents an important discriminative word between these books.

The first observation of these histograms shows that the two books are written by authors using different vocabulary style.

4.1.3 COST parameter

4.1.3.1 Definition of the COST parameter. Usually, when poets write a series of poems, they make a termination similarity between the neighboring sentences of the poem, such as a same final syllable or letter. To evaluate that termination similarity, a new parameter estimating the degree of text chain (in a text of several sentences) has been proposed: the COST parameter.

Thus, the COST parameter for sentence ‘j’ is computed by adding all the occurrence marks (values) between sentence ‘j’ and its neighboring sentences (sentence ‘j−1’ and sentence ‘j+1’). In our case, the occurrence marks concern only the two last letters of the sentence.

It is interesting to note that Quran and Hadith books do not contain poems, but they consist in statements, indications, histories, questions and answers, human obligations, advices, description of God, description of the after-life, etc. The COST parameter, in this case, can give some information on the structure of the text (ending structure). In this investigation, it has been employed to see if the two texts respect certain regularities in the text structure or not and, if so, to assess the corresponding regularity ratio.

For instance, let us observe the following English poem:

(1) Never say it is the end when we do believe → COST = 2
(2) And never accept that you do not retrieve → COST = 2
(3) Life is so short to let things kill our mind → COST = 2
(4) What to do in such situations dear friend → COST = 4
(5) It is true that it is hard but victory will be in hand → COST = 2
(6) Do not hesitate to try if you can make any change → COST = 1
(7) Yes it is worth trying even if it is the last chance → COST = 1

If we consider the fourth sentence (ending with ‘nd’), we notice that the previous and next sentences...
(sentence 3 and 5) are ended with the same last two characters (i.e. 'nd').

So by counting the number of similar characters [i.e. \((1+1) + (1+1) = 4\)], we get a COST value of four. The same procedure is repeated for each sentence until the last one.

For concreteness, here are the COST values for some Hadith sentences (Table 1) and the COST values of some Quran sentences (Table 2).

According to these tables, we remark that for the Hadith mixture, there are many COST values equal to zero; and when the COST is non-null, it has very small values: the average COST is only 0.46.

For the Quran, we notice that the COST is almost never null and the corresponding values are relatively high: the average COST of the Quran is approximately 2.52.

This fact means that the structure of the Quran is very different from the Hadith one. Consequently, the two books must have two different author styles.

4.1.4 Word length frequency-based analysis

The fourth experiment is an investigation on the word length frequency. Herein, we must define some technical terms employed in our article:

1. The word length is the number of letters composing that word.

2. 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.

In Fig. 2, the two spectra are represented simultaneously, which gives an interesting way to compare the two books. So, let us assume that \(F_{\text{Quran}}(j)\) is the frequency of the words with ‘\(j\)’ letters in the Quran and \(F_{\text{Hadith}}(j)\) 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:

1. Length 1: \(F_{\text{Quran}}(1) = 10.95\%\), whereas \(F_{\text{Hadith}}(1) = 8.03\%\); 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 \(\chi^2\) (uncorrected for continuity) regarding this result is 167.54, involving a probability of consistency \(P < 0.0001\), consequently results related to one-word frequency appear to be significant.
(2) 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 $\chi^2$ (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.

(3) 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 $\chi^2$ (uncorrected for continuity) regarding this result is 198.3, involving a probability of consistency $P < 0.00001$, consequently results related to long-word frequency appear to be significant.

(4) Lengths 9 and 10: The Quran contains approximately a double number of words with nine and ten letters than the Hadith. This fact shows that the Quran vocabulary contains more very-long words (very-long stands for more than eight letters) than the Hadith. The Pearson $\chi^2$ (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.

4.1.5 Character frequency-based analysis

The fifth experiment makes a comparison between the character frequencies of the two books (Fig. 3). From Figure 3, we did a sorting of the differences between the two frequencies (Quran frequency and Hadith frequency), for all the characters, in a descending order. At the end we kept only the sixteen first characters that have been sorted (Fig. 4). In Figure 3, we have represented the character frequencies used in the two books. We can see, for example, that for the first five characters (i.e. श च ण न त and /) the difference between the utilization frequencies in the two books is appreciable. This observation implies two different writing styles for the two books.
4.1.6 Discriminative words

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

4.1.6.1 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 23,068 tokens and 6,225 different words. The Quran contains 87,339 tokens and 13,473 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 (Figs 5 and 6). Such tokens are called Discriminant Words (we chose this appellation due to the proposed application of discrimination).

4.1.6.2 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.

4.1.7 Numbers citation-based analysis

The seventh experiment investigates the citation of numbers in the text: How many times a specific number (zero to nine) has been used in the books?

In Figure 7, we can easily notice that the most frequently cited number, in the Quran, is
the number ‘1’, whereas for the Hadith it is the number ‘3’.

We also notice that both books use more odd numbers than even ones, except for the Quran book regarding the number ‘5’, where it appears to be the less used non-null number, which is not the case for the Hadith (its corresponding frequency is about 10%).

In this experiment, the difference regarding the use of numbers, between the two books, is so different that we can state that the authors should probably be different.

4.1.8 Animal citation-based analysis

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

The animal citation frequency (Freq.) is defined as follows:

\[
\text{Freq. in \%} = 100 \times \frac{\text{frequency of occurrence}}{\text{total number of animal citations}}
\]

Example:

Freq. in Quran = 100 \times \frac{\text{frequency of occurrence}}{155}, since the total number of animal citations was 155.
Freq. in Hadith = 100 \times \frac{\text{frequency of occurrence}}{94}, since the total number of animal citations was 94.

### First observation.

The following Table 3 shows that for the seven following animals, the difference in citation between the two books is relatively great:

- The name **اذاعة** (general name of kamels, cows, and sheeps) is cited thirty three times in the Quran, whereas in the Bukhari Hadith it is cited only two times;
- The name **كلب** (dog) is cited only five times in the Quran, whereas in the Bukhari Hadith it is cited thirteen times;
- The name **شاعر** (sheep) is completely absent in the Quran, whereas in the Bukhari Hadith it is cited ten times;
- The name **ذابة** (animal) is cited seventeen times in the Quran, whereas in the Bukhari Hadith it is cited only three times;
- The name **البلاء** (camel) is cited only two times in the Quran, whereas in the Bukhari Hadith it is cited seven times;
- The name **جلاء** (calf) is cited ten times in the Quran, whereas in the Bukhari Hadith it is completely absent; and
- The name **لحاء** (fish) is cited only four times in the Quran, whereas in the Bukhari Hadith it is cited eight times.

### Fig. 5

Hadith words never used in Quran: 3,885 different words (over 6,225 total different words contained in Bukhari Hadith): \(\frac{3,885}{6,225} = 62.41\%\) of words absent in Quran.

### Fig. 6

Quran words never used in Bukhari Hadith: 11,133 different words (over 13,473 total different words contained in Quran): \(\frac{11,133}{13,473} = 82.63\%\) of words absent in Hadith.

---

434 Literary and Linguistic Computing, Vol. 27, No. 4, 2012
Table 3 Citation frequency of some animals appearing more frequently in one book than in the other

<table>
<thead>
<tr>
<th>Animal</th>
<th>Translation</th>
<th>Citation in Quran</th>
<th>Citation in Hadith</th>
<th>Frequency in Quran (%)</th>
<th>Frequency in Hadith (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calf</td>
<td>Ant</td>
<td>33</td>
<td>2</td>
<td>21.3</td>
<td>2.13</td>
</tr>
<tr>
<td>Fish</td>
<td></td>
<td>4</td>
<td>8</td>
<td>2.6</td>
<td>8.51</td>
</tr>
<tr>
<td>Sheep</td>
<td>Cat</td>
<td>5</td>
<td>13</td>
<td>3.2</td>
<td>13.83</td>
</tr>
<tr>
<td>Thirsty camel</td>
<td>Miraclous type of horse</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Lice</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Bee</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Mosquito</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Type of horse</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>General name</td>
<td>(kamels, cows, sheeps)</td>
<td>33</td>
<td>2</td>
<td>21.3</td>
<td>2.13</td>
</tr>
</tbody>
</table>

4.1.8.2 Second observation. In Table 4, we quote the animals that are quoted in the Quran but completely absent in the Bukhari Hadith. There are twenty nine such animal names.

Table 4 Citation frequency of animals that are quoted in the Quran but completely absent in the Bukhari Hadith

<table>
<thead>
<tr>
<th>Animal</th>
<th>Translation</th>
<th>Citation in Quran</th>
<th>Citation in Hadith</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calf</td>
<td></td>
<td>10</td>
<td>Absent</td>
</tr>
<tr>
<td>Deer</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Frog</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Ant</td>
<td></td>
<td>3</td>
<td>Absent</td>
</tr>
<tr>
<td>Monkey</td>
<td></td>
<td>3</td>
<td>Absent</td>
</tr>
<tr>
<td>Female sheep</td>
<td></td>
<td>3</td>
<td>Absent</td>
</tr>
<tr>
<td>Snake</td>
<td></td>
<td>2</td>
<td>Absent</td>
</tr>
<tr>
<td>Fly</td>
<td></td>
<td>2</td>
<td>Absent</td>
</tr>
<tr>
<td>Spider</td>
<td></td>
<td>2</td>
<td>Absent</td>
</tr>
<tr>
<td>Grasshopper</td>
<td></td>
<td>2</td>
<td>Absent</td>
</tr>
<tr>
<td>Crow</td>
<td></td>
<td>2</td>
<td>Absent</td>
</tr>
<tr>
<td>Lion</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Hoopoe</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Witch</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Lice</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Mosquito</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Bee</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Lamb</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Goat</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Yet</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Frog</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Thirsty camel</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>General name</td>
<td>(kamels, cows, sheeps)</td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Maybe: type</td>
<td>birds</td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Lions</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Earthworm</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Pregnant camel</td>
<td>(+/-)</td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Wild animals</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Type of wild monkeys or maybe zebras</td>
<td>1</td>
<td>Absent</td>
<td></td>
</tr>
</tbody>
</table>

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

Table 5 Citation frequency of animals that are quoted in the Bukhari Hadith but completely absent in the Quran

<table>
<thead>
<tr>
<th>Animal</th>
<th>Translation</th>
<th>Citation in Quran</th>
<th>Citation in Hadith</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheep</td>
<td></td>
<td>0</td>
<td>Absent</td>
</tr>
<tr>
<td>Bull</td>
<td></td>
<td>3</td>
<td>Absent</td>
</tr>
<tr>
<td>Cat</td>
<td></td>
<td>2</td>
<td>Absent</td>
</tr>
<tr>
<td>Bird</td>
<td></td>
<td>2</td>
<td>Absent</td>
</tr>
<tr>
<td>Horse</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Sheep</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Rooster</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Hen</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
<tr>
<td>Miraculous type of horse (Buraq)</td>
<td></td>
<td>1</td>
<td>Absent</td>
</tr>
</tbody>
</table>

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

4.1.8.3 Third observation. In Table 5, we quote the animals that are quoted in the Bukhari Hadith but completely absent in the Quran. There are eleven such animal names.

A particular observation can be done about the name شاة (sheep), which is cited ten times in the Bukhari Hadith and which is completely absent in the Quran.
Discussion: results show that there are different animal name citations in the two books. That is, two cases are possible:

1. the two books could be related to two topics that are contextually different, citing a contextual type of animal consequently;
2. 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.

4.1.9 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:

- ‘ح-’
- ‘و-’
- ‘م-’
- ‘م-’
- ‘م-’
- ‘م-’

Usually, these bigrams (except the third and fourth ones) are often related to the plural form in Arabic.

We notice, in Fig. 8, 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.

4.2 Second series of experiments: segmental analysis

The second series of experiments analyses the two books in a segmental form: four different segments of texts are extracted from every book and the different texts are analyzed 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 of these features:

Vocabulary-based features: a simple way to confirm or refute authorship is to look for something that completely settles the authorship question (Juola, 2006). It is clear, then, that the individual words an author uses can reveal his or her identity. The problem with such features is that the data can be faked easily. A more reliable method would be able to take into account a large fraction of the words in the document (Juola, 2006) as the average sentence length.

Syntax-based features: one reason that function words perform well is because they are topic-independent (Juola, 2006). A person’s preferred syntactic constructions can be cues to his authorship. One simple way to capture this is to tag the relevant documents for part of speech or other syntactic constructions (Stamatatos, 2001) using a tagger.

Orthographic-based features: one weakness of vocabulary-based approaches is that they do not take advantage of morphologically related words. A person who writes of ‘work’ is also likely to write of ‘working’, ‘worker’, etc. (Juola, 2006).

Characters-based features: some researchers (Peng, 2003) have proposed to analyze documents as sequences of characters. For example, the character four-gram ‘work’ is shared in the previous example by all the words. That is why this type of parameter can replace several other high-level linguistic features. Furthermore, several experiments showed that character n-gram is one of the most reliable and robust features in authorship attribution (Stamatatos, 2009).

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 ten 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.

4.2.1 Discriminative words

This first 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: (in English: THOSE or WHO in a plural form) and (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 Table 6.

For 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 one-tenth of the Quran frequency).

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency (%) in the Quran segments</th>
<th>Frequency (%) in the Hadith segments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quran 1</td>
<td>Quran 2</td>
</tr>
<tr>
<td>الدين</td>
<td>1.35</td>
<td>1.02</td>
</tr>
<tr>
<td>الأرض</td>
<td>0.34</td>
<td>0.63</td>
</tr>
</tbody>
</table>
For 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.

### 4.2.2 Word length frequency-based analysis

The second experiment is an investigation on the word length frequency. In Fig. 9, the different curves (smoothed curves), representing the ‘word length frequency’ versus the ‘word length’, show the following two important points:

1. 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).
2. 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%.

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.

### 4.2.3 COST parameter

The third experiment concerns the new COST parameter which appears non-null only in the Holy Quran, as we can see in Table 7. The COST parameter is explained in Section 4.1.3.

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.

Table 7 shows the average COST values of the eight different segments.

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).

4.2.4 Discriminative characters
The fourth 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 'ṣ', 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 Table 8, 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.

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

4.2.5 Vocabulary-based similarity
The fifth 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 (i.e. vocabulary similarity measure), which is defined as follows:

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

Typically, in case of two identical texts, this similarity measure will have a value of one (i.e. 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 9.

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 four segments of the Quran have a great similarity in vocabulary and the four 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

<table>
<thead>
<tr>
<th>Segment</th>
<th>Quran 1</th>
<th>Quran 2</th>
<th>Quran 3</th>
<th>Quran 4</th>
<th>Hadith 1</th>
<th>Hadith 2</th>
<th>Hadith 3</th>
<th>Hadith 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSTaverage</td>
<td>2.2</td>
<td>2.6</td>
<td>2.6</td>
<td>2.38</td>
<td>0.46</td>
<td>0.47</td>
<td>0.43</td>
<td>0.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segment</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of character ṣ</td>
<td>7.73</td>
<td>7.11</td>
<td>6.91</td>
<td>7.04</td>
<td>5.19</td>
<td>5.45</td>
<td>4.72</td>
<td>5.33</td>
</tr>
</tbody>
</table>
two different books. This deduction can easily be made from Table 10, which represents the mean similarity measure between one segment and all the segments of a given book.

Table 10 gives the mean similarity according to Quran or Hadith for each segment X (\(X = Q_i\) or \(X = H_i\), \(i = 1\) to \(4\)), which can be expressed as the average of all the similarities between segment X and the different segments of a same book. Table 10 is displayed in order to see if a segment is more similar to the Quran family or to the Hadith family.

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.

### 4.3 Third series of experiments: automatic authorship attribution with several features and several classifiers

The third series of experiments, which consists in an automatic authorship attribution (Sanderson, 2006), analyses the two books in a segmental form by using several features (words, word n-grams, characters, character n-grams, and dis-legomena) (Clement 2003) and several classifiers: Canberra distance, cosine distance, Ryan-Noecker (RN) cross-entropy, histogram distance, intersection distance, Kullback-Leibler distance, Manhattan distance, Linear Discriminant Analysis (LDA) and Naïve Bayes classifier (Juola, 2009).

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 (the same dataset as in experiment 4.2.5).

It concerns two experiments:

1. 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 (three for each book) have to be classified into Quran class or Hadith class.

2. 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.

In this series of experiments, the author employs the JGAA software (Juola, 2009) to make an

---

**Table 9** Similarity matrix representing the different VSM similarity measures between segments

<table>
<thead>
<tr>
<th>VSM (%)</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>100</td>
<td>32.89</td>
<td>31.43</td>
<td>28.22</td>
<td>20.93</td>
<td>19.86</td>
<td>19.38</td>
<td>19.86</td>
</tr>
<tr>
<td>H2</td>
<td>32.89</td>
<td>100</td>
<td>31.37</td>
<td>29.23</td>
<td>20.84</td>
<td>19.99</td>
<td>18.63</td>
<td>19.45</td>
</tr>
<tr>
<td>H3</td>
<td>31.43</td>
<td>31.37</td>
<td>100</td>
<td>29.17</td>
<td>19.77</td>
<td>19.88</td>
<td>18.90</td>
<td>18.96</td>
</tr>
<tr>
<td>H4</td>
<td>28.22</td>
<td>29.23</td>
<td>29.17</td>
<td>100</td>
<td>19.93</td>
<td>18.68</td>
<td>18.55</td>
<td>18.79</td>
</tr>
<tr>
<td>Q1</td>
<td>20.93</td>
<td>20.84</td>
<td>19.77</td>
<td>19.93</td>
<td>100</td>
<td>29.73</td>
<td>29.56</td>
<td>24.49</td>
</tr>
<tr>
<td>Q2</td>
<td>19.86</td>
<td>19.99</td>
<td>19.88</td>
<td>18.68</td>
<td>29.73</td>
<td>100</td>
<td>34.88</td>
<td>25.22</td>
</tr>
<tr>
<td>Q3</td>
<td>19.38</td>
<td>18.63</td>
<td>18.90</td>
<td>18.55</td>
<td>29.56</td>
<td>34.88</td>
<td>100</td>
<td>27.09</td>
</tr>
<tr>
<td>Q4</td>
<td>19.86</td>
<td>19.45</td>
<td>18.96</td>
<td>18.79</td>
<td>24.49</td>
<td>25.22</td>
<td>27.09</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 10** Mean VSM similarity in % between one segment and the different segments of a same book

<table>
<thead>
<tr>
<th>Segment</th>
<th>Mean similarity with H segments</th>
<th>Mean similarity with Q segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>30.85</td>
<td>20.01</td>
</tr>
<tr>
<td>H2</td>
<td>31.16</td>
<td>19.73</td>
</tr>
<tr>
<td>H3</td>
<td>30.66</td>
<td>19.38</td>
</tr>
<tr>
<td>H4</td>
<td>28.87</td>
<td>18.99</td>
</tr>
<tr>
<td>Q1</td>
<td>20.37</td>
<td>27.92</td>
</tr>
<tr>
<td>Q2</td>
<td>19.60</td>
<td>29.94</td>
</tr>
<tr>
<td>Q3</td>
<td>18.87</td>
<td>30.51</td>
</tr>
<tr>
<td>Q4</td>
<td>19.27</td>
<td>25.60</td>
</tr>
</tbody>
</table>
automatic classification of the eight texts by using different features and different classifiers.

Concerning the number of selected examples (an example refers to a word, character, etc.), we have considered two cases: in the first case, we consider all the examples and in the second case, we keep only the fifty 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.

4.3.1 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.

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, LDA analysis, and Naïve Bayes classifier), shows clearly that the four Quran segments should belong to a same author, the four Hadith segments should belong to the same author too and that these two authors are likely to be different.

4.3.2 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.

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 12.

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, LDA analysis and Naïve Bayes classifier), shows clearly that the four Quran segments should belong to the same author, the four Hadith segments should belong to the same author.
too and that these two authors are very probably different.

Discussion on these two experiments (Sections 4.3.1 and 4.3.2): 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.

5 Discussion

In this research work, we have made an investigation of authorship discrimination (Tambouratzis, 2000, 2003) between two old Arabic books: the Quran and Bukhari Hadith.

For that purpose, we made three series of experiments:

The first series of experiments, which consists of nine experiments, analyses the two books in a global form.

The second series, consisting of five experiments, analyses the two books in a segmental form by using statistical techniques of stylometry.

The third series of experiments, which consists of two experiments, analyses the two books in a segmental form by using state-of-the-art techniques of stylometry and by employing several classifiers and several types of features.

Results of all experiments have led to two main conclusions:

1. First, the two investigated books should have different authors;
2. Second, all the segments that are extracted from a unique book appear to have a certain stylistic similarity.

Consequently, we can conclude, according to this investigation, that the Quran was not written by the Prophet Muhammad and that it belongs to a unique author too. Muslims believe that it is written by Allah (God) and sent to his messenger (the prophet Muhammad). We will not extend our research work into an etymological point of view: this is not the main topic of this work, but we think that it may be interesting to mention this point.

The result of this investigation rejects, then, the doubts assuming that the Quran was written by the Prophet and it confirms what has been stated by Muhammad: that the Quran has been transmitted to him (by Allah).

Acknowledgements

The author wishes to thank Dr George Tambouratzis and Dr Marina Vassiliou from...
the ILSP institute of Athens and Dr Efstathios Stamatatos from the Aegaean University of Samos for their great help and advice during the elaboration of this investigation. He also wishes to thank Dr Patrick Juola from Duquesne University for the useful program that he has generously released (JGAAP) and for his pertinent comments regarding this article. I do not forget to express many thanks to Prof. Michael Oakes from Sunderland University for his constructive suggestions and help during the finalization of this article. Also, I would like to express many thanks to the anonymous reviewers for their time and valuable comments. Finally, in order to enhance the quality of this research work, the author welcomes all the comments and suggestions of the readers regarding this topic.

References


within a Specific Register. In *Proceedings of the Workshop on Text Processing for Modern Greek: From Symbolic to Statistical Approaches* (held in conjunction with the 6th International Conference of Greek Linguistics). Rethymno, Greece, 20 September, pp. 1–10.
