Investigation on the Author's Style and the Authenticity of the Holy Quran

3rd Edition - February 2025

Pr Dr Halim Sayoud

Book Edited by Pr Halim Sayoud
Publisher EDT - SCHOLARPAGE

DOI: 10.5281/zenodo.14865663

Investigation on the Author's Style and the Authenticity of the Holy Quran

by

Pr, Dr. Halim Sayoud

Sayoud.net

Book Edited by Pr Halim Sayoud

Publisher EDT - SCHOLARPAGE

DOI: 10.5281/zenodo.14865663

February 2025

بسم الله الرحمن الرحيم

Abstract

In this book (3rd edition), the author tries to see whether the Quran could be a simple invention of the Prophet (i.e. written by the Prophet) or really a book from God (i.e. a divine book sent down by Allah) as claimed in the Islamic religion, in a scientific manner.

The book can be considered as a pure scientific investigation without any form of theological or ideological point of view. Also, the author does not discuss his personal beliefs on the subject, but only what the clear results of this investigation show.

The author describes the 19 series of scientific experiments, which were conducted during this investigation, as well as the new scientific knowledge recently discovered in the Quran. He also describes several recent scientific research works confirming the authenticity of the holy Book.

All the results reported in those experiments show that the Author's style of the Quran is completely different from the author's style of the Hadith (i.e. Prophet), which leads to the following conclusion: The Quran could not be written or invented by the Prophet.

Finally, as reported in this book, the scientific knowledge that is discovered in the Quran, and commented by several famous researchers in the field, confirms this conclusion and shows that the holy book could not be written by a human being.

About the Author

Dr. Halim Sayoud is full Professor at USTHB University and head of the EDT research team. Although, he is specialized in Electronics and Computer Sciences, his research works focus on a variety of research fields such as Natural language processing, Artificial intelligence and Biometrics.

He is the author and co-author of more than 100 scientific articles in conference proceedings, book chapters, scientific journals, etc. He is also the Editor-in-chief of a scientific journal. His official website is http://sayoud.net/

Contents *with hyperlinks*

1 Introduction

In the Islamic religion, it is well known that the holy Quran is the book of Allah (God) and is claimed to be written (i.e. created) by him. So the Author of the holy Quran is known to be Allah (God), as stated in the holy Quran and confirmed by the Prophet (pbuh).

For instance, the following Quran verses clearly confirm that authorship.

Verse	**Ref.**
أَفَلَا يَتَدَبَّرُونَ الْقُرْآنَ وَلَوْ كَانَ مِنْ عِنْدِ غَيْرِ اللَّهِ لَوَجَدُوا فِيهِ اخْتِلَافًا كَثِيرًا	4:82
Translation: Then do they not reflect upon the Qur'an? If it had been from [any] other than Allah, they would have found within it much contradiction.	

Verse	**Ref.**
وَمَا كَانَ هَذَا الْقُرْآنُ أَنْ يُفْتَرَى مِنْ دُونِ اللَّهِ وَلَكِنْ تَصْدِيقَ الَّذِي بَيْنَ يَدَيْهِ وَتَفْصِيلَ الْكِتَابِ لَا رَيْبَ فِيهِ مِنْ رَبِّ الْعَالَمِينَ	10:37
Translation: And it was not [possible] for this Qur'an to be produced by other than Allah, but [it is] a confirmation of what was before it and a detailed explanation of the [former] Scripture, about which there is no doubt, from the Lord of the worlds.	

Verse	**Ref.**
نَحْنُ نَقُصُّ عَلَيْكَ أَحْسَنَ الْقَصَصِ بِمَا أَوْحَيْنَا إِلَيْكَ هَذَا الْقُرْآنَ وَإِنْ كُنْتَ مِنْ قَبْلِهِ لَمِنَ الْغَافِلِينَ	12:3
Translation: We narrate to you the best of narratives, by Our revealing to you this Quran, though before this you were certainly one of those who did not know.	

Verse	**Ref.**
قُلْ لَئِنِ اجْتَمَعَتِ الْإِنْسُ وَالْجِنُّ عَلَى أَنْ يَأْتُوا بِمِثْلِ هَذَا الْقُرْآنِ لَا يَأْتُونَ بِمِثْلِهِ وَلَوْ كَانَ بَعْضُهُمْ لِبَعْضٍ ظَهِيرًا	17:88
Translation: Say: "If all mankind and all invisible beings would come together with a view to producing the like of this Qur'an, they could not produce its like even though they were to exert all their strength in aiding one another!"	

Verse	**Ref.**
وَإِنَّكَ لَتُلَقَّى الْقُرْآنَ مِنْ لَدُنْ حَكِيمٍ عَلِيمٍ	27:6
Translation: Indeed you receive the Quran from One who is all-wise, all-knowing.	

Verse	Ref.
إِنَّا نَحْنُ نَزَّلْنَا عَلَيْكَ الْقُرْآنَ تَنْزِيلًا	76:23
Translation: Indeed, it is We who have sent down to you, [O Muhammad], the Qur'an progressively.	

Moreover, the Prophet (Pbuh) clearly confirmed that the Quran was sent down to him by God, as expressed in the following Hadith statements – the Hadith represents the statements and speech of the Prophet.

Hadith 1

قال رسول الله – صلى الله عليه وسلم – : **نزلت علي سورة الأنعام جملة واحدة** ، وشيعها سبعون ألفا من الملائكة ، لهم زجل بالتسبيح والتحميد
Translation: The Prophet Pbuh said "The Surah Al-Anaam (chapter 6 of the Quran) was sent down to him at once (in one time) …"

Hadith 2

قال رسول الله – صلى الله عليه وسلم – : " **لقد أنزلت علي آية هي أحب إلي من الدنيا** وما فيها كلها **."**
Translation: The Prophet Pbuh said "One verse was sent down to me, which I love more than this world…"

Hadith 3

عَنْ الأَعْمَشِ عَنْ إِبْرَاهِيمَ عَنْ عَبِيدَةَ عَنْ عَبْدِ اللَّهِ قَالَ: **قَالَ لِي رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ إِقْرَأْ عَلَيَّ الْقُرْآنَ قَالَ :فَقُلْتُ يَا رَسُولَ اللَّهِ أَقْرَأُ عَلَيْكَ وَعَلَيْكَ أُنْزِلَ** قَالَ: إِنِّي أَشْتَهِي أَنْ أَسْمَعَهُ مِنْ غَيْرِي
Translation: The Prophet Pbuh said to his companion Abdullah "Recite to me some verses of the Quran" and then Abdullah replied "how could I recite the Quran in your presence while it was sent down to you (by God)?"

Hadith 4

قال رسول الله – صلى الله عليه وسلم – : **تركتُ فيكم أيُّها الناس، ما إنِ اعتصمتم به، فلن تضلُّوا أبدًا: كتاب الله، وسُنَّة نبيّه**
Translation: The Prophet Pbuh said: ""I have left two matters with you. As long as you hold to them, you will not go the wrong way. They are the Book of Allah (Quran) and the Sunna of His Prophet (Hadith)."

So, in all those statements from the Hadith, it is clear that the Prophet confirmed that the Quran was sent down to him from Allah (God).

Some unstructured doubts on the origin of the Quran

Some ignorant persons tried over the time to claim that the holy Quran is only an invention of the Prophet and that it could be written by him, instead of God, without providing any proof.

Strangely, one can retrieve such claims even in the Quran when describing some ignorant or malicious persons describing the holy Quran as a simple invention of the Prophet without any link to God (Allah), as stated in the following verses (32-3):

أَمْ يَقُولُونَ افْتَرَاهُ ۚ بَلْ هُوَ الْحَقُّ مِن رَّبِّكَ لِتُنذِرَ قَوْمًا مَّا أَتَاهُم مِّن نَّذِيرٍ مِّن قَبْلِكَ لَعَلَّهُمْ يَهْتَدُونَ

Translation: Or do they say, "He invented it"? Rather, it is the truth from your Lord, [O Muhammad], that you may warn a people to whom no warner has come before you [so] perhaps they will be guided.

Purpose of this research work

Due to some doubts on the origin of the holy Quran, from certain untrustworthy persons, it becomes necessary to undertake a scientific investigation on the matter to try discarding the doubts or at least clarifying the problem.

The solution, in the opinion of the author, is a clear scientific investigation that should shed the light on the problem and hopefully provides a final answer to all related questions.

In the age of computer sciences and artificial intelligence, it is commonly becoming possible to tackle some problems that were quite difficult to handle before, like biometrics, for instance, which has been widely used nowadays to recognize human beings in different situations. So, the key methodology should focuses on such intelligent techniques with maybe some text-mining tools to try providing a fair scientific answer to the authorship problem.

How to identify an author?

Human beings have distinctive ways of speaking and writing, as explained by Corney (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 (Madigan et al., 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.

The research field dealing with author recognition is called “Stylometry”. This research domain includes several specialties, such as: Authorship Attribution, Authorship Verification, Plagiarism Detection, Author Discrimination, Text Segmentation and so on.

There are several features that can be used: characters, vocabulary, sentences, mistakes, character N-grams, word N-grams, etc. Also, several classification techniques can be employed: distance metrics, statistics, automatic learning classifiers, neural networks, deep learning and so on.

Is stylometry efficient?

Like many biometric modalities, stylometry can be quite accurate provided that the following conditions are well respected.

Firstly, the amount of text data should be sufficiently large: a minimum of 2500 words per document should provide good performances as reported by Eder (Eder, 2010).

Secondly, to be fair, the examined texts should have the same topic and same genre.

Thirdly, the compared text documents should belong to the same period of time and should be written with the same character type.

If all those conditions are respected, the author recognition performances can reach a quite high level of accuracy. For instance, during an experiment of authorship attribution on a closed-set dataset of 100 authors, called HAT Corpus (*HAT is composed of 100 groups of Arabic texts that are extracted from 100 different Arabic books. The books were written by 100 different authors with the same topic: Travel*), we got an authorship identification accuracy of 97%, while the text documents size was only 1100 words per document. This result and many others conducted by other researchers confirm the efficiency of stylometry if the previous conditions are respected.

Tackling the religious enigma

In this research work, we deal with a religious enigma, which has not been solved for fifteen hundred years, as cited by Sayoud in (S. 2010] H. Sayoud, 2010). In fact, as mentioned previously, certain doubts on the origins of the Holy Quran do exist and some ignorant persons thought that the Holy Quran could be an invention of the prophet Muhammad, for three purposes (Al-Shreef, 2009):

- To facilitate his domination over his followers;
- To frighten the unbelievers and those who disobey his orders;
- 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.

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 (Li et al., 2006), we tried to make some experiments of author discrimination between the Quran and some Prophet's statements in order to shed the light on this enigma.

For this purpose, several investigations, experiments and definitions are presented and commented as follow:

In chapter 2, we give a brief description of the Holy Quran.

In chapter 3, we give a brief description of the Hadith.

In chapter 4, we tackle the question "Who was the Author of the Quran?"

In chapter 5, we define the task of Stylometry and Automatic Authorship Attribution

In chapter 6, we present the Quran and Hadith Corpora

In chapter 7, we describe our first Series of Experiments: A Global Analysis

In chapter 8, we describe our second series of experiments: Big Segments based Segmental analysis

In chapter 9, we explain the details of the third series of experiments: Automatic authorship attribution with several features and several classifiers

In chapter 10, we describe the fourth series of experiments: Short Segments based Segmental Authorship Attribution

In chapter 11, we describe the fifth Series of Experiment: Stylometric Comparison between the Quran and Hadith based on Successive Function Words

In chapter 12, we introduce and comment the sixth Series of Experiment: Authorship Identification of 7 Books – A Fusion Approach

In chapter 13, we present the seventh series of experiments, where we propose the authorship discrimination method using the Leave-One-Out Validation.

In chapter 14, we describe the authorship discrimination based on Gaussianity and Interpolability

In chapter 15, we talk about a mysterious numerical structure found in the Quran, which makes it different from other Human Books

In chapter 16, we propose a new method of authorship attribution based on the Interrogative Form

In chapter 17, we present an investigation on the Quran/Hadith Authorship Using Visual Analytics Approaches

In chapter 18, we propose a new method of authorship discrimination based on Word Transition Probability

In chapter 19, we explore some Embedded Scientific Knowledge in the holy book

In chapter 20, we give a general conclusion

Finally some references are given at the end of the book.

2 Description of the Holy Quran

The Quran (in Arabic: القرآن al-qur'ān, literally "the recitation"; also sometimes transliterated as Qur'ān, Koran, Alcoran or Al-Qur'ān (Wiki1, 2012, p. 1) (Nasr, 2004) is the central religious text of Islam. Muslims believe the Quran to be the book of divine guidance and direction for mankind (I. A. Ibrahim, 1997) (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 (ie. God) and transmitted to Muhammad by the angel Gibraele (Gabriel) over a period of 23 years. The beginning of Quran apparition was in the year 610 (after the birth of Christ).

Figure 2.1 Very old copy Quran (Image credit: Cadbury Research Library, University of Birmingham)

3 Description of the Hadith

Hadith (in Arabic: الحديث, transliteration: al-ḥadīth is the oral statements and words said by the Islamic prophet Muhammad (Pbuh) (Wiki2, 2012, p. 2) (Islahi, 1989). 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 (Wiki2, 2012, p. 2) (Islahi, 1989). 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).

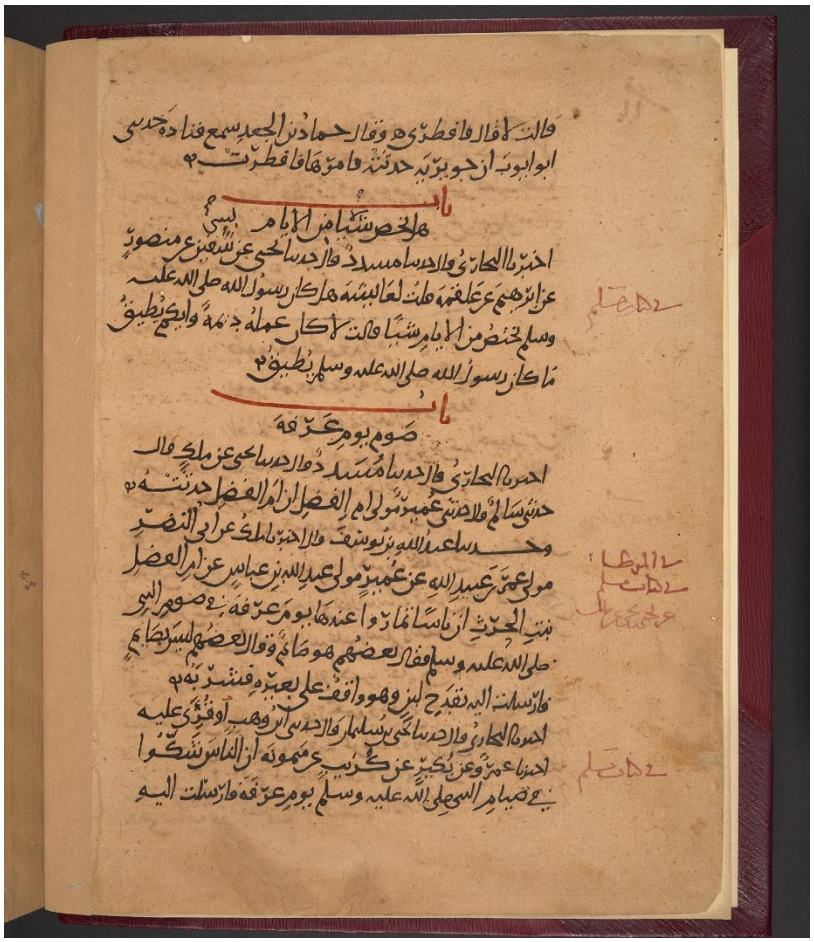


Figure 3.1 A very old manuscript of al-Bukhari's hadith collection. Courtesy of the British Library.

4 Who was the Author of the Quran?

Muslims believe that the holy Quran is from Allah (God) and 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 (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 et al., 2000) (Tambouratzis et al., 2003), regardless of the literal style or context.

5 Stylometric Analysis and Automatic Authorship Attribution

Individuals have distinctive ways of speaking and writing, as explained by Corney (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 (Madigan et al., 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.
Stylometry or author recognition is a research field that consists in recognizing the authentic author of a piece of text. It is evident that the recognition accuracy is not as high as some biometric modalities that are used in security purposes, but it has been shown that for texts with more than 2500 tokens, the recognition task becomes significantly accurate (Signoriello et al., 2005) (Eder, 2010).

Stylometry (or author recognition) can be divided into several research fields:

- Authorship Attribution (Sarwar et al., 2020), or identification, which consists in identifying the author(s) of a text;
- Authorship verification (Kestemont et al., 2020), which consists in checking if a text claimed to be written by somebody is really written by himself;
- Authorship discrimination (S. 2012] H. Sayoud, 2012), which consists in checking if two texts are written by the same author or not;
- Authorship Indexing (Zangerle et al., 2020), which consists in segmenting a multi-author text into several homogeneous segments and giving the identity of each author in those homogeneous segments;
- Plagiarism detection (Muangprathub et al., 2021) (Zouhir et al., 2021), which consists in checking if a piece of text has been picked from another author.

That is; determining the real author of a piece of text has raised several questions and problems for centuries. Problem of authorship can be of interest not only to humanities researchers, but also to politicians, historians and religious scholars in particular. Thorough investigative journalism, combined with scientific analysis (*e.g., chemical analysis*) of documents has traditionally given good results (Juola, 2006).
Furthermore, the recent development of improved statistical techniques in conjunction with the large availability of digital corpora, have made the automatic and objective inference of authorship a practical and easy task. That is why, this research field has seen an explosion of scholarship, resulting in several related works (Farringdon, 1996) (Sari et al., 2018) (Foster, 2001) (Evert, 2017) (Love, 2002) (Al-Batineh, 2019) (McMenamin, 2002) (Kalgutkar et al., 2018) (Mosteller & Wallace, 1964) (Schuster et al., 2020).
Research works on authorship attribution usually appear at several types of debates ranging from linguistics and literature through machine learning and computation, to law and forensics. Despite this interest, the field itself is somewhat in confusion with a certain sense of best practices and techniques (Juola, 2006).

Stylometry was also used in religious disputes. In fact, 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. 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 (Li et al., 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).

6 The Quran and Hadith Corpora

Herein we summarize the size of the two books in terms of words, tokens, pages, etc.

6.1 Dimension of the two religious books

In a previous work, we used the entire text of the Quran (something like 315 A4 pages) but a small collection of the Hadith (not exceeding 3 pages) only, due to the difficulty to find a book containing only the Prophet's sentences (without the comments of the narrators). In this context the author was strongly advised by some experienced stylometric researchers, who were working on Greek discourses, to try to increase the size of the Hadith text, in order to get a consistent comparison between the two investigated books. So, after a thorough investigation on the Hadith texts, the author managed to collect a confident and consistent dataset, which is organized in a form that is more convenient (book gathering pure Prophet statements, called Bukhari Hadith).

That is, the present section summarizes the size of the two new investigated books in terms of words, tokens, pages, etc. The statistical characteristics of these two books are summarized as follows:

- Quran size in terms of tokens: 87341
- Hadith size in terms of tokens (*Bukhari Hadith*): 23068
- Quran size in terms of different words = 13473
- Hadith size in terms of different words (*Bukhari Hadith*) = 6225
- Quran A4 pages in the Quran: 315 pages (*subjective size*)
- Hadith A4 pages in the Hadith (*Bukhari Hadith*): 87 pages (*subjective size*)
- Ratio of the Number of Quran Tokens / Number of Hadith Tokens = 3.79
- Ratio of the Number of Quran Lines / Number of Hadith Lines = 3.61
- Ratio of the Number of different Quran words / Number of different Hadith words = 2.16
- Ratio of the Number of Quran A4 Pages / Number of Hadith A4 Pages = 3.62

According to these size details, the two religious books seem relatively consistent, since the average number of pages is 315 for the Quran book and 87 for the Hadith book. However, since the two books do not have the same size, it will be necessary and prudent to segment these two books into segments of more or less a same size, in order to avoid unbalanced results.

6.2 Segmentation

As quoted in section 6.1, the author already conducted an authorship investigation on the two religious books by considering the whole books entirely (S. 2012] H. Sayoud, 2012). In that approach, when comparing two books, it is difficult to know any part of the book is similar to the other one or different from it. That is why 2 judicious segmentations have been proposed and applied on the different books,

which consists in segmenting those books into several text segments, where the sizes of the segments are more or less in the same range.

In the first segmentation technique, we divided every book into 4 different big text segments, where a segment size is about 10 standard pages and all the segments are distinct and separated (without intersection).

In the second segmentation technique, we used 14 different text segments for the Quran and 11 different text segments for the Hadith, with approximately the same size (about 2100 words per dsegment). In case of machine learning based classification, these segments should be organized as follows: some text segments are selected from every book to represent the training data and the remaining text segments are used during the testing step. However, in the Leave-One-Out method, all the text segments are used for classification/attribution.

The segments have more or less the same size in terms of words as it is shown in table 6.1. The medium size is about 2076 words per text.

The problem with such a size is that authorship attribution (AA) systems are usually not very accurate, since it has been shown that the minimum text size, for a good AA process, is at least 2500 words per size (Eder, 2010) (Signoriello et al., 2005).

Table 6.1: Sizes of the different text segments

Hadith text segments	Size in terms of tokens	Quran text segments	Size in terms of tokens
H1	2035	Q1	2064
H2	2096	Q2	2071
H3	2053	Q3	2086
H4	2059	Q4	2085
H5	2081	Q5	2081
H6	2073	Q6	2080
H7	2031	Q7	2087
H8	2082	Q8	2074
H9	2088	Q9	2081
H10	2097	Q10	2079
H11	2083	Q11	2078
/	/	Q12	2092
/	/	Q13	2093
/	/	Q14	2081

6.3 Word structure of the different segments

A graphical representation of the word length frequency has been made for every text segment, in order to see the overall structure of the used words in term of size. Figure 6.1 represents the smoothed word length frequency curves versus the number of characters per word. It shows that the words have more or less the same dimension frequency for both books, except for unigrams (1-character words),

trigrams, tetragrams and octograms (8-character words), where we distinguish a clear difference in their frequencies, but this observation cannot be used for objective discrimination purposes.

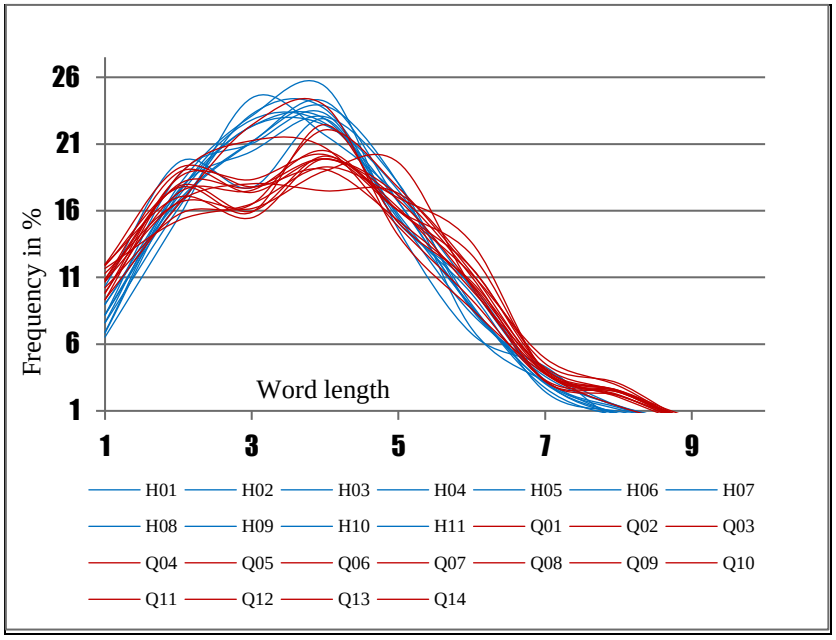


Figure 6.1: Word length frequency versus the word length (for all text segments). Q stands for a Quran segment and H stands for a Hadith segment. Curves are obtained by interpolation.

7 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 eight experiments. In this series of experiment we analyze the two books (Quran and Hadith) in a global form.

7.1 Authorship Discrimination based on Words

This experiment represents an investigation on the word frequency. Results are displayed in the following figure.

Notion of discriminative words

A particular interest concerns the discriminative words, as we can see in figure 7.1.

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 (see figure 7.1).

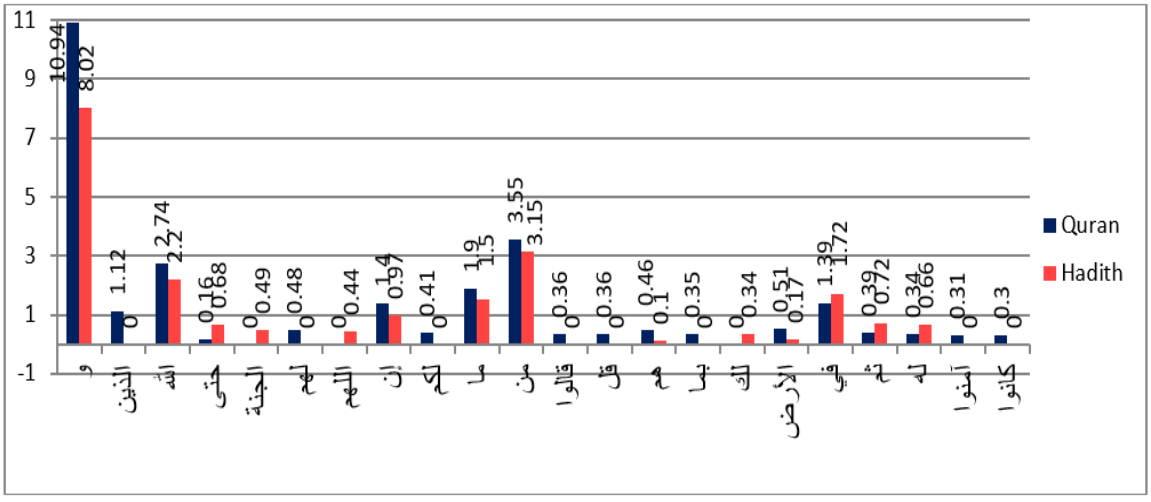


Figure 7.1: Word frequency (in %) of some of the most discriminative words: black for the Quran and grey for the Hadith.

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* "*و*"), 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 "الذين"*), meaning "*Those*" in Arabic, 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.

7.2 Authorship Discrimination based on Characters

This experiment makes a comparison between the character frequencies of the two books. From the resulting frequencies, we did a sorting of the differences between the 2 frequencies (Quran frequency and Hadith frequency), for all the characters, in a descending order. At the end we kept only the 16 first characters that have been sorted (see figure 7.2). In that figure, we have represented the character frequencies used in the two books. We can see, for example, that for the first five characters (ie. و, ن, م, ف, and أ), the difference between the utilization frequencies in the two books is appreciable. This observation implies two different writing styles for the two books.

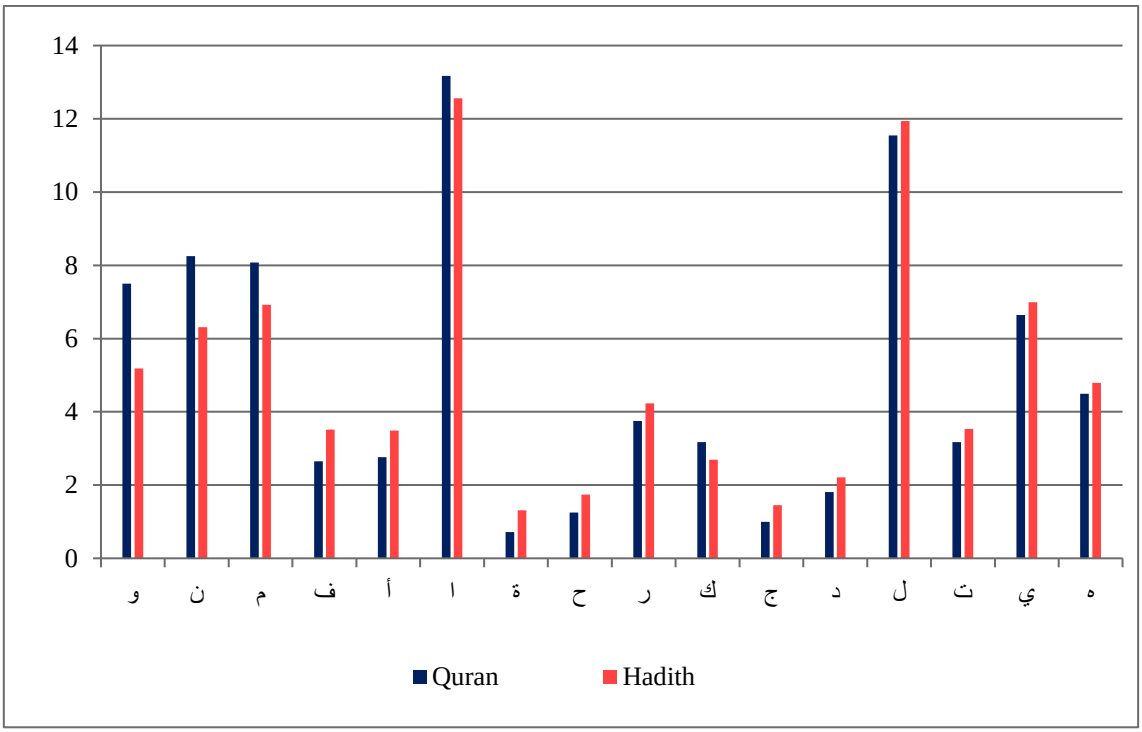


Figure 7.2: Frequency of the most discriminant Arabic characters

7.3 Authorship Analysis based on COST parameter

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:

Never say it is the end when we do believe	→	COST = 2
And never accept that you do not retrieve	→	COST = 2
Life is so short to let things kill our mind	→	COST = 2
What to do in such situations dear friend	→	COST = 4
It is true that it is hard but victory will be in hand	→	COST = 2
Do not hesitate to try if you can make any change	→	COST = 1
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 2 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 4. The same procedure is repeated for each sentence until the last one.

For concreteness, here are the COST values for some Hadith sentences (see table 7.3.1) and the COST values of some Quran sentences (see table 7.3.2).

Table 7.3.1: COST values for some Hadith sentences

Sentence No	Cost	last 2 characters	Word
1\	0	دا	أبدا
2\	0	ون	تظلمون
3\	0	ية	الجاهلية
4\	0	له	الله
5\	0	ان	شعبان
6\	0	يه	نبيه
7\	0	غت	بلغت
8\	0	هد	اشهد
9\	0	كم	لأنفسكم
10\	1	يك	عليك
11\	1	سك	لنفسك
12\	0	نم	جهنم
13\	1	ته	بركاته
14\	2	نه	أستعينه
15\	1	نا	أعمالنا
16\	1	له	له
17\	1	غه	أبلغه
18\	0	كم	عليكم

Table 7.3.2: COST values of some Quran sentences

Sentence No	Cost	last 2 characters	Word
3116\	4	ين	المسحرين
3117\	4	ين	الكاذبين
3118\	3	ين	الصادقين
3119\	1	ون	تعملون
3120\	1	يم	عظيم
3121\	2	ين	مؤمنين
3122\	2	يم	الرحيم
3123\	3	ين	العالمين
3124\	4	ين	الأمين
3125\	4	ين	المنذرين
3126\	4	ين	مبين
3127\	3	ين	الأولين
3128\	2	يل	إسرائيل
3129\	3	ين	الأعجمين
3130\	4	ين	مؤمنين
3131\	3	ين	المجرمين
3132\	1	يم	الأليم
3133\	2	ون	يشعرون
3134\	4	ون	منظرون
3135\	3	ون	يستعجلون
3136\	2	ين	سنين

According to these tables, we remark that for the Hadith mixture, there are many COST values equal to 0; 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.

7.4 Word length frequency based analysis

This experiment represents 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.

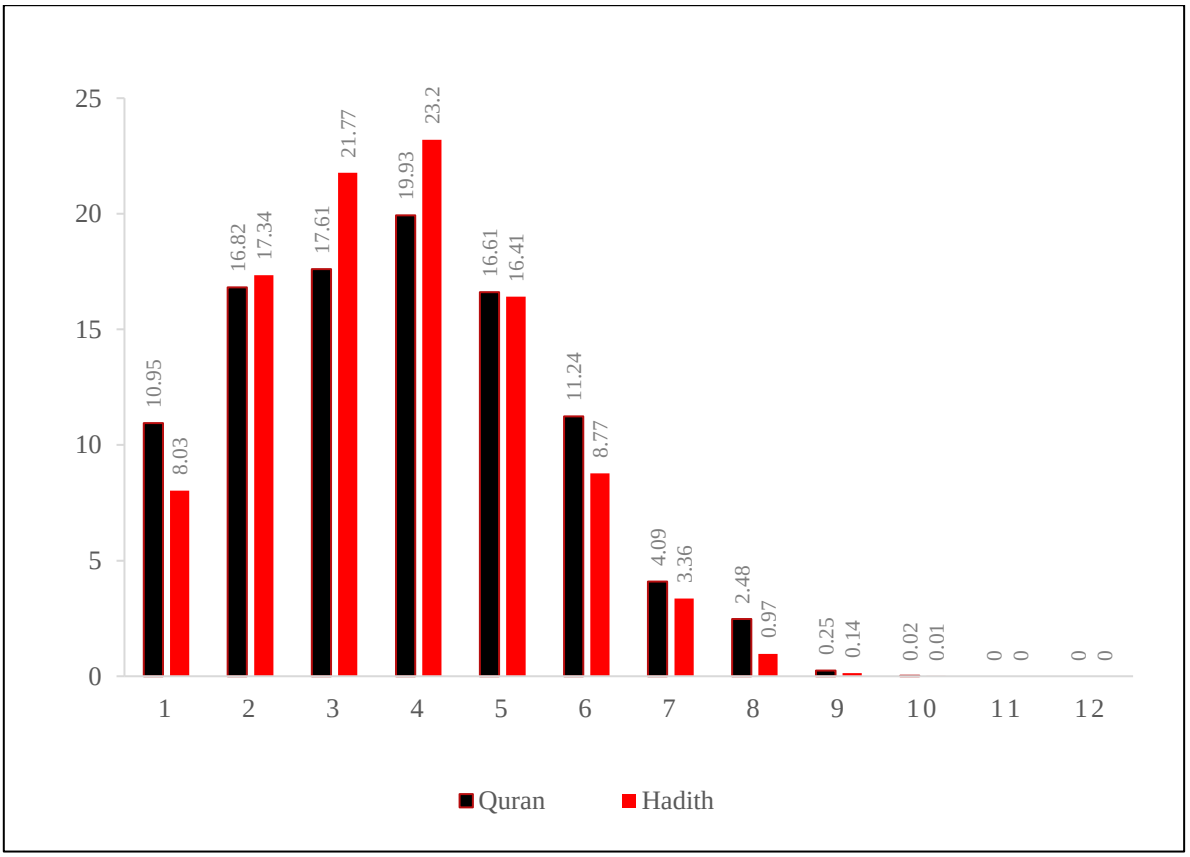


Figure 7.4: Word length frequency in histograms representation.

In figure 7.4, the two spectra are represented simultaneously, which gives an interesting way to compare the two books. So, let us assume that FQuran(j) is the frequency of the words with “j” letters in the Quran and FHadith(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: FQuran (1)=10.95%, whereas FHadith (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.

7.5 Discriminative words based analysis

In this 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 7.5.1 and 7.5.2). Such tokens are called *Discriminant Words (we chose this appellation due to the proposed application of discrimination).*

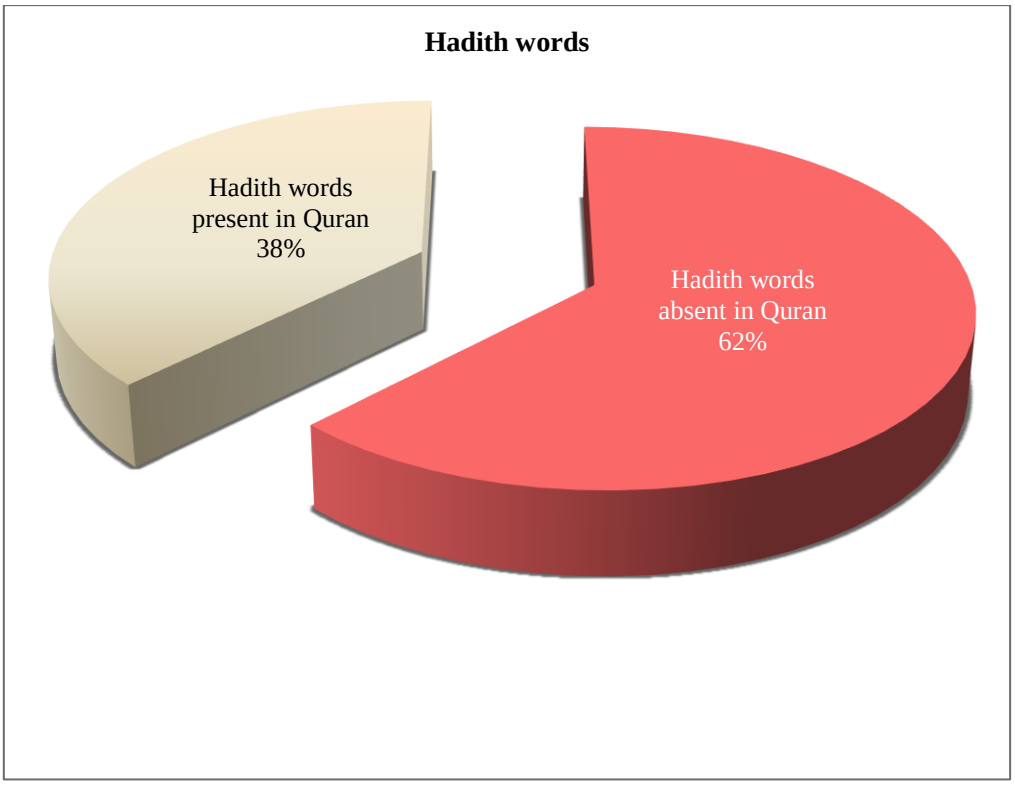


Figure 7.5.1: Hadith words never used in Quran : 3885 different words (over 6225 total different words contained in Bukhari Hadith) : 3885/6225 =62.41% of words absent in Quran.

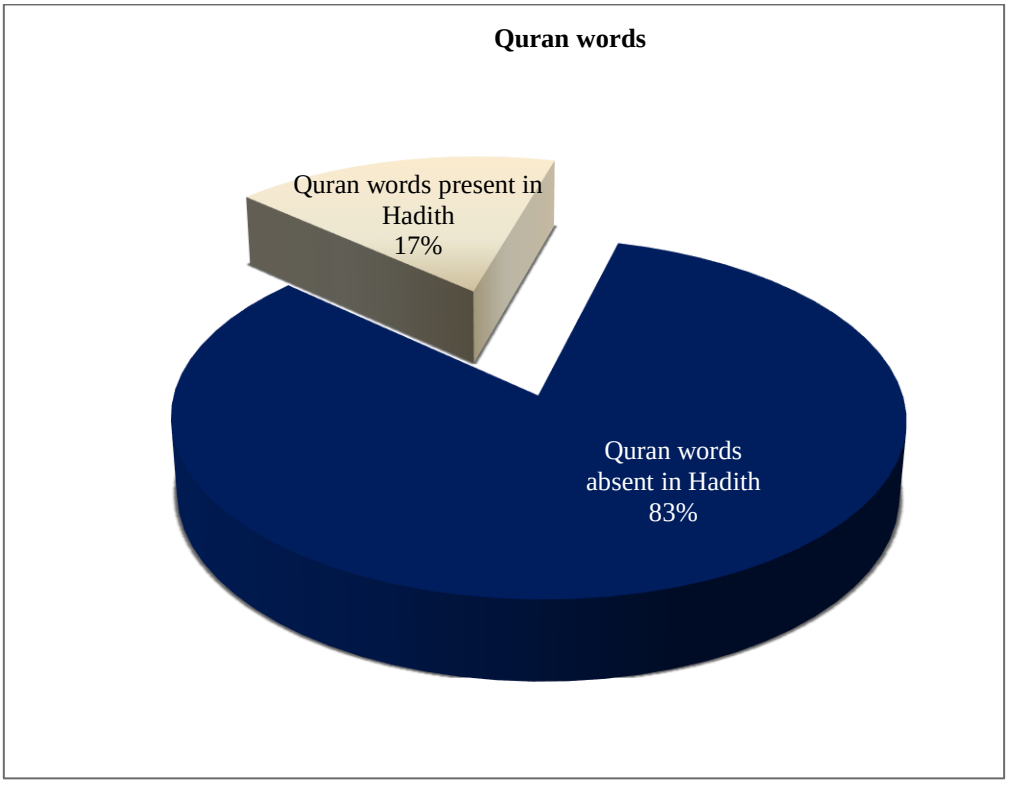


Figure 7.5.2: Quran words never used in Bukhari Hadith: 11133 different words (over 13473 total different words contained in Quran) : 11133/13473=82.63% of words absent in Hadith.

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.

7.6 Numbers citation based analysis

This experiment investigates the citation of numbers in the text: How many times a specific number (0.. 9) has been used in the books?

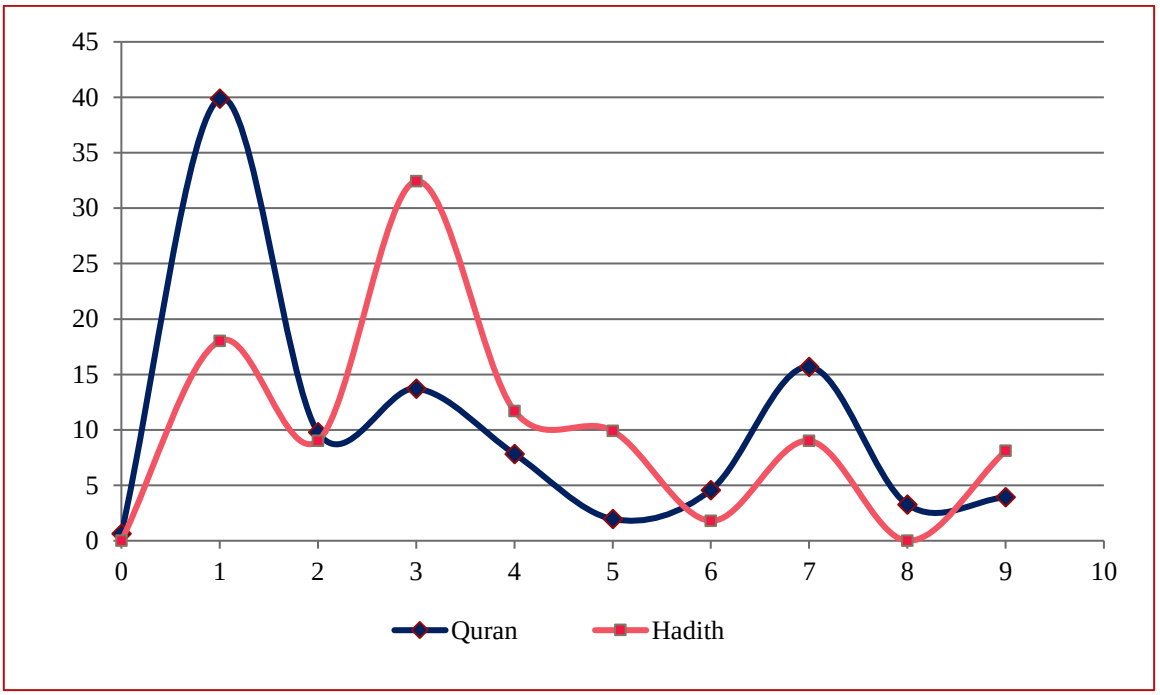


Figure 7.6: Frequency of numbers citation with interpolated curves.

In the above graph (figure 7.6), we 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.

7.7 Animal citation based analysis

The eighth 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)

Example:

freq in Quran = 100.(frequency of occurrence / 155), since the total number of animal citations was 155.

freq in Hadith = 100.(frequency of occurrence / 94), since the total number of animal citations was 94.

First observation:

The following table 7.7.1 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 7.7.1: 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 7.7.2, 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 7.7.2: Citation frequency of animals that are quoted in the Quran but completely absent in the Bukhari Hadith.

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 7.7.3, 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 7.7.3: Citation frequency of animals that are quoted in the Bukhari Hadith but completely absent in the Quran.

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.

7.8 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 character 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:

"ـين-"
"ون-"
"يم-"
"وم-"

"ـهم"
"ـكم"

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

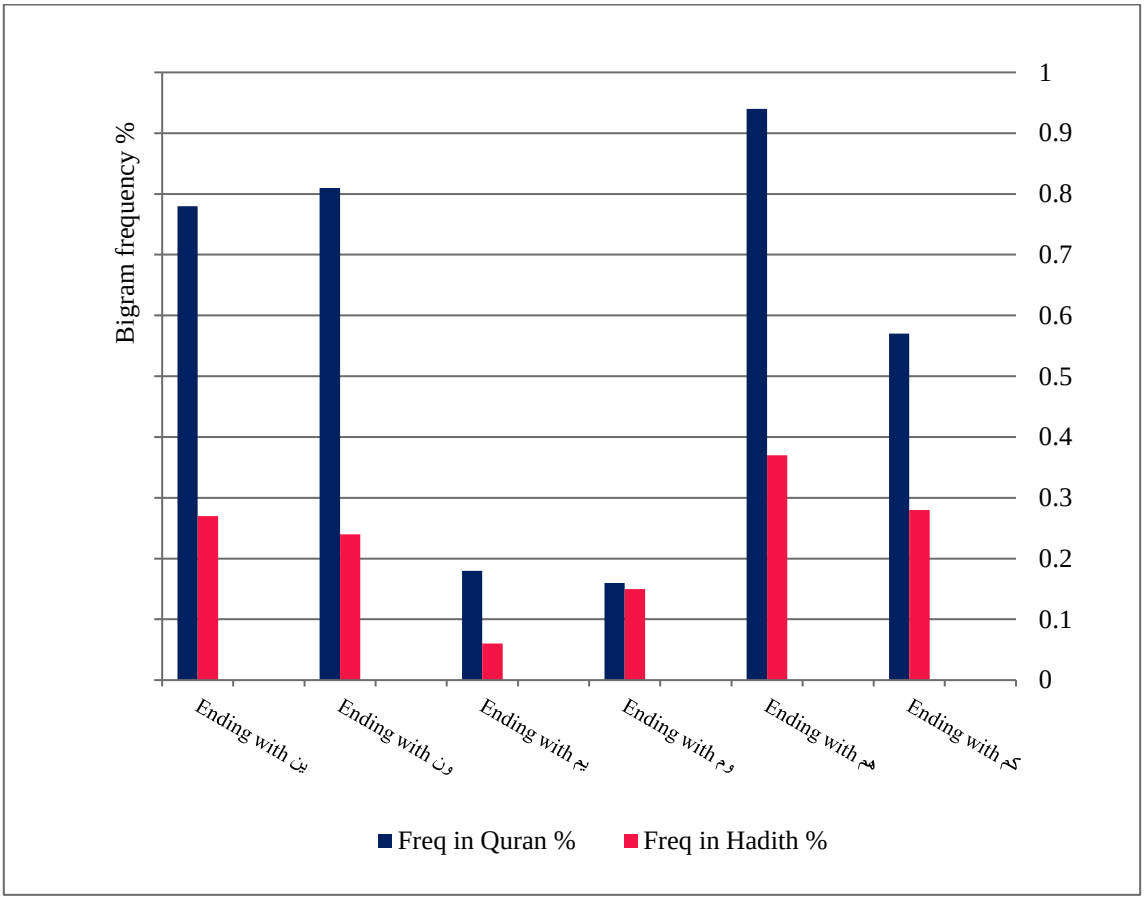


Figure 7.8: Frequency of some ending bigrams.

We notice, in figure 7.8, 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.

7.9 Discussion on the first series of experiments

This research work consists in an investigation of authorship discrimination between two old Arabic books: the Quran and Bukhari Hadith.

In this first series of experiments, which consists of 8 different experiments, we have analyzed the two books in a global form. And amazingly, the 8 corresponding experiments have led to the same conclusion: the two investigated books should have different authors.

Consequently, this first series of experiments shows that the Quran could not be written by the Prophet Muhammad.

8 Second series of experiments: Big Segments based Segmental analysis

The second series of experiments analyses the two books in a segmental form with big text segments: 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 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 et al., 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 et al., 2003) have proposed to analyze documents as sequences of characters. For example, the character 4-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 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.

8.1 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: 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 8.1).

Table 8.1: 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.

8.2 Word length frequency based analysis

The second experiment is an investigation on the word length frequency. In the following figure (figure 8.2), 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%.

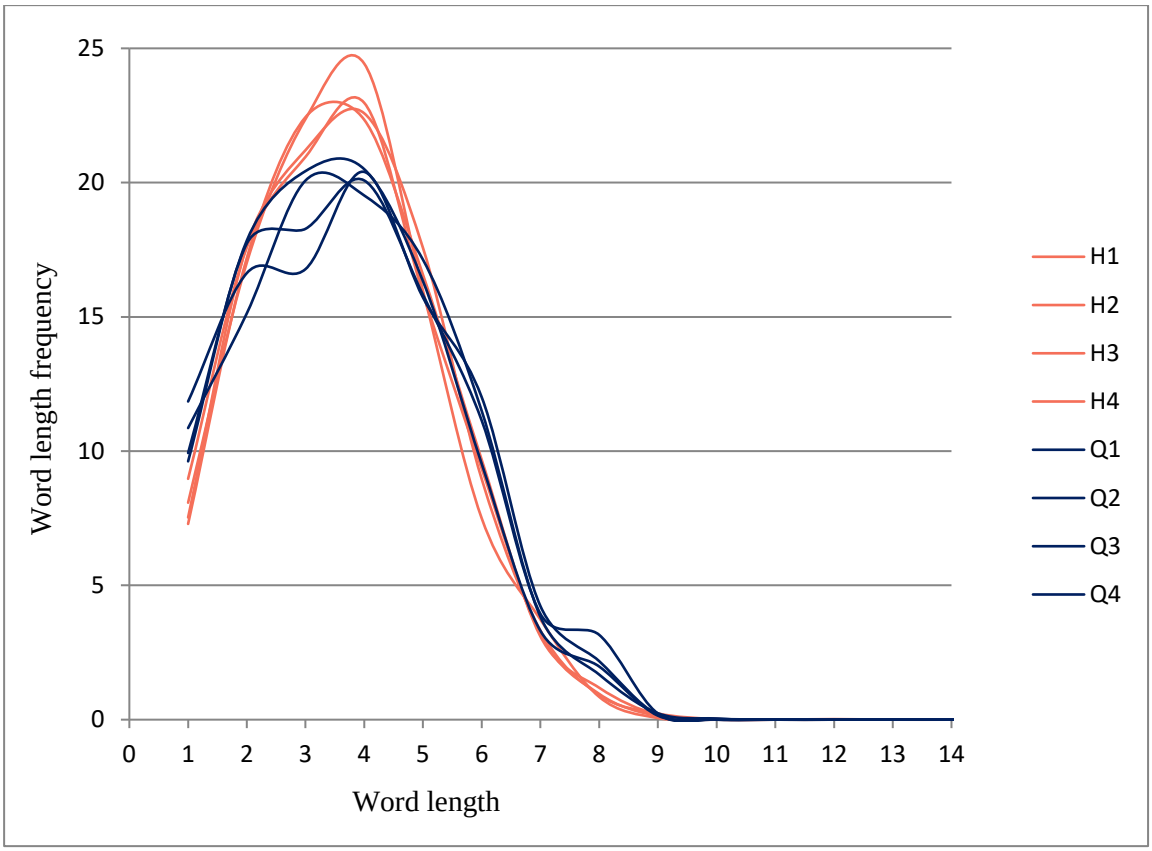


Figure 8.2: Word length frequency (smoothed lines).

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.

8.3 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 8.3. The COST parameter is explained in section 7.3.

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.

The following table shows the average COST values of the 8 different segments.

Table 8.3: 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).

8.4 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 " 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 8.4: 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.

8.5 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 (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 8.5.1.

Table 8.5.1: 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 8.5.2 gives the mean similarity according to Quran or Hadith for each segment X (X=Qi or X=Hi , 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 8.5.2: 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.

8.6 Discussion on the second series of experiments

This second series, consisting of 5 experiments, analyzed the two books in a segmental form by using statistical techniques of stylometry, where the segments are quite big (four different big segments of texts are extracted from every book).

Once again, the 5 corresponding experiments have led to the same conclusion: the authors' styles of the two investigated books are very different.

Consequently and once again, this second series of experiments shows that the Quran could not be conceived by the Prophet Muhammad.

9 Third series of experiments: Automatic authorship attribution with several features and several classifiers

The next series of experiments, which consists in an automatic authorship attribution (Sanderson & Guener, 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 & Sharp, 2003) 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,) (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 8.5).

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.

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

9.1 First experiment

In the first 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 9.1.

Table 9.1: 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.

9.2 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 9.2: Precision of good classification of the different segments with several features and several classifiers

Feature / Classifier	Charc Bigram	Charc-ter	Charc 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 9.2.
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 (*9.1 and 9.2*): 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.

9.3 Discussion on the third series of experiments

The third series of experiments could be considered as a continuation of the 2nd series of experiments since it uses the same text segments (i.e. same segmented corpus).
It consists of 2 experiments analyzing the two books in a segmental form by using nearest neighbor techniques of classification and by employing several types of features.
This series of experiments is interesting since it shows a score of 100% in almost all the experimental tests, which means that all Qi segments are similar one another by representing the Quran class, and all Hi segments are similar one another by representing the Hadith class, while the Qi segments are completely different from the Hi segments. This difference is probably due to a distinction between the 2 styles. Hence this result clearly confirms that the two books Quran and Hadith should have different authors.

10 Fourth series of experiments: Short Segments based Segmental Authorship Attribution

In this research part, the religious books are finely split and segmented into 25 different (relatively short) medium documents.

In fact, 14 text segments are extracted from the Quran book and 11 text segments are extracted from the Bukhari Hadith. These segments have more or less the same size in terms of words and the medium size is about 2080 words per text segment. The Quran is taken in its entirety, whereas for the Prophet's statements, we chose only the certified texts of the Bukhari book. That is, four series of experiments are done and commented. The first experiment concerns several experiments of authorship attribution using different state of the art features and classifiers, the second experiment analyses the different texts by using a new parameter called COST, the third experiment consists in an authorship discrimination using the frequency of a particular word ("الذين " meaning those/who in English) and the fourth experiment performs a hierarchical clustering on the 25 text segments, in order to assess the real number of clusters (author styles) and to see if the hypothesis of a unique author is possible.

The text segments are organized as follows: three text segments are selected from every book to represent the training data and the remaining text segments are used during the testing step. In the other cases, all the text segments are used for classification/attribution. The segments have more or less the same size in terms of words as it is shown in table 10.1. The medium size is about 2076 words per text. The problem with such a size is that authorship attribution (AA) systems are usually not accurate, since it has been shown that the minimum text size, for a good AA process, is at least 2500 words per size (Eder, 2010) (Signoriello et al., 2005).

Table 10.1: Sizes of the different text segments

Hadith text segments	Size in terms of tokens	Quran text segments	Size in terms of tokens
H1	2035	Q1	2064
H2	2096	Q2	2071
H3	2053	Q3	2086
H4	2059	Q4	2085
H5	2081	Q5	2081
H6	2073	Q6	2080
H7	2031	Q7	2087
H8	2082	Q8	2074
H9	2088	Q9	2081
H10	2097	Q10	2079
H11	2083	Q11	2078
/	/	Q12	2092
/	/	Q13	2093
/	/	Q14	2081

Word structure of the different segments

A graphical representation of the word length frequency has been made for every text segment, in order to see the overall structure of the used words in term of size. Figure 10.1 represents the smoothed word length frequency curves versus the number of characters per word. It shows that the words have more or less the same dimension frequency for both books, except for unigrams (1-character words), trigrams, tetragrams and octograms (8-character words), where we often distinguish a certain difference in their frequencies, but this observation cannot be used for objective discrimination purposes.

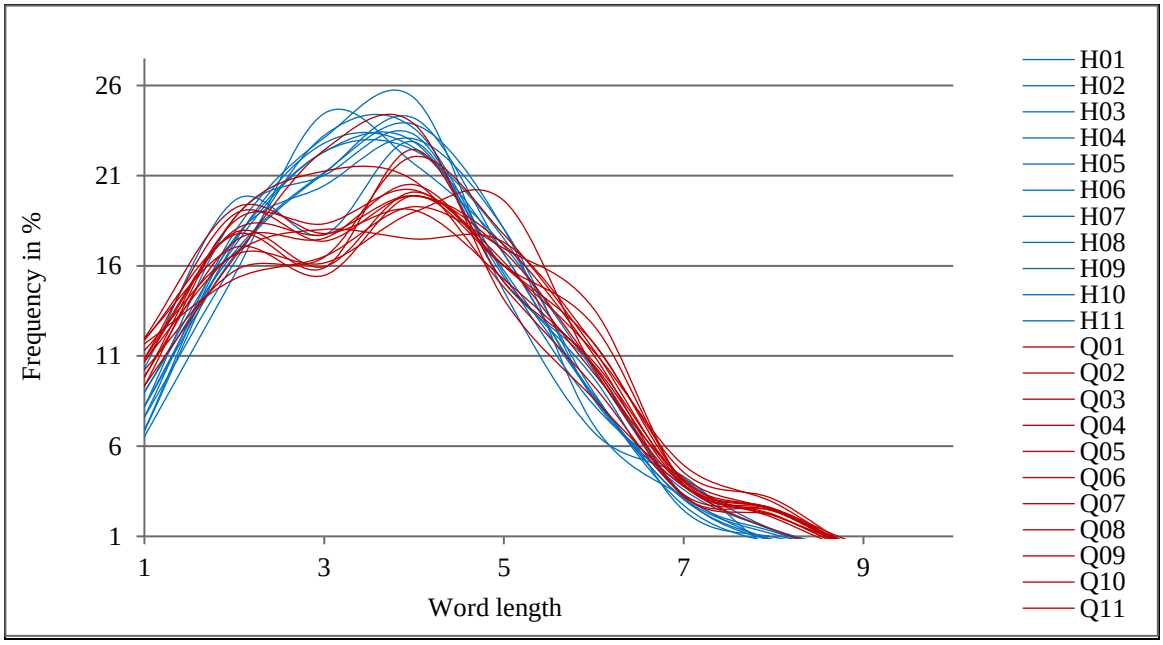


Figure 10.1: Word length frequency versus the word length (for all text segments). Curves are obtained by interpolation.

In the following sections, we will describe the four experiments that have been conducted on the two religious books for a purpose of authorship discrimination.

10.1 Experiments of Authorship Attribution using a Hierarchical clustering

In order to represent the stylistic similitude between the different texts, in a graphical way, a hierarchical clustering (Sayoud 2012-b), using cityblock distance, has been performed on all text segments by using the following features: COST parameter (see section 7.3) and frequency of the word "الذين " meaning THOSE (or WHO in a plural form) in English. The resulting dendrogram is displayed in figure 10.1, where we can see the different possible clusters and their costs (distances). The smaller the cost is, the more similar the segments are (in the same cluster).

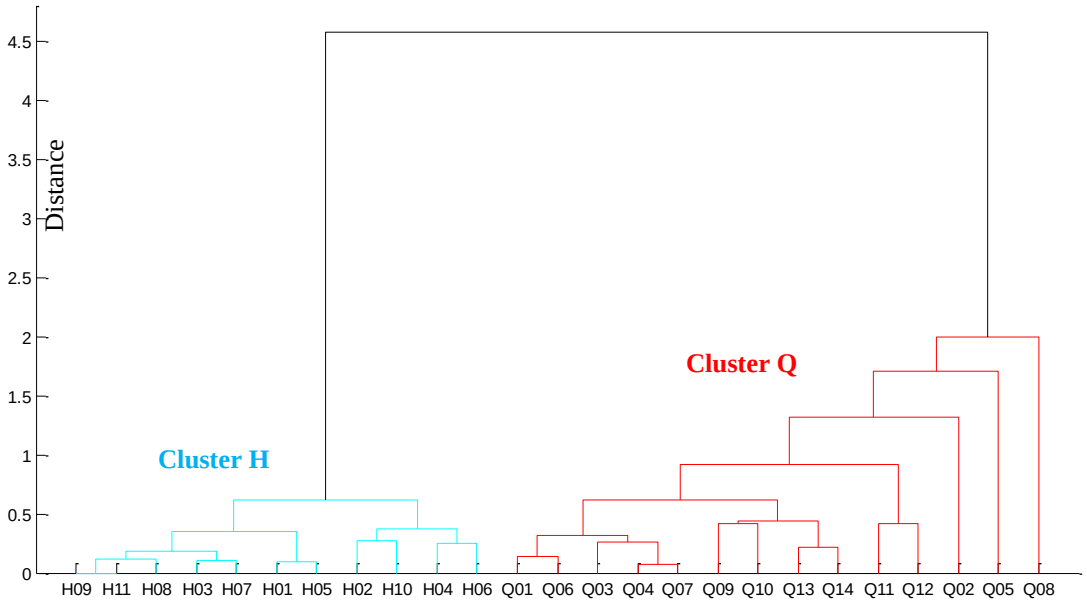


Figure 10.2: Dendrogram of a hierarchical clustering corresponding to the 25 text segments.

As we can see in figure 10.2, the segments have been automatically divided into 2 main clusters: “cluster Q” (in dark red) grouping all the text segments of the Quran and “cluster H” (in light blue) gathering all the text segments of the Hadith. We can notice that the last clustering into one cluster (big line at the top) is inconsistent for two reasons: first, because the corresponding distance of this last cluster is more than 4.5, which is relatively very large; and second, because we do not retrieve any link between heterogeneous segments at all (clusters grouping different label types such as Qj-Hk). This result confirms the previous conclusions stating that the different text segments should belong to 2 different authors, or at least 2 different author styles. It also shows that Quran texts are relatively similar (low intra-variability with distances less than 2) and that Hadith texts are relatively similar too (low intra-variability with distances less than 1).

10.2 Experiments of Authorship Attribution using different state of the art features and classifiers

This series of experiments, which consists in an authorship attribution (Sanderson & Guener, 2006), analyses the two books in a segmental form by using several features (word n-grams, character n-grams and rare words) (Clement & Sharp, 2003) and several classifiers: Stamatatos distance, Canberra distance, Cosine distance, RN cross entropy distance, Intersection distance, Manhattan distance, SMO-SVM (Sequential Minimal Optimization based Support Vector Machine) classifier, Linear Regression classifier and MLP (Multi Layer Perceptron) classifier.

10.2.1. Brief description of the different classifiers

Short definitions of the different classifiers are given below:

- **Manhattan distance**

This distance (S. 2012] H. Sayoud, 2012) is very reliable in text classification. The Manhattan distance between two vectors f and g is given by the following formula:

$$\sum_{i=1}^{n} |f_i - g_i| \qquad (10.1)$$

where n is the length of the vector.

- **Cosine distance**

Cosine similarity is a measure of similarity between two vectors that measures the cosine of the angle between them. The technique is also used to compare documents in text mining. The cosine of two vectors can be derived by using the Euclidean dot product formula:

$$f.g = \|f\|\|g\| \cos\theta \qquad (10.2)$$

Given two vectors of attributes, f and g, the cosine similarity, cos(θ), is represented using a dot product and magnitude as:

$$similarity = \cos\theta = \frac{f.g}{\|f\|\|g\|} = \frac{\sum_{i=1}^{n} f_i * g_i}{\sqrt{\sum_{i=1}^{n} (f_i)^2} * \sqrt{\sum_{i=1}^{n} (g_i)^2}} \qquad (10.3)$$

where ‖f‖ denotes the magnitude of vector f and n is its length (Wiki_COS, 2013).

- **Stamatatos distance**

This distance was proposed by Stamatatos (Stamatatos, 2007). The Stamatatos distance between two vectors f and g is given by the following formula:

$$\sum_{i=1}^{n} [2(f_i - g_i)/(f_i + g_i)]^2 \qquad (10.4)$$

where n is the length of the vector.

- **Canberra distance**

Canberra distance is a numerical measure of the distance between pairs of points in a vector space. It is more or less similar to Manhattan distance. It is mostly used for data scattered around the origin. The Canberra distance between two vectors f and g is given by the following formula:

$$\sum_{i=1}^{n} \left| \frac{(f_i - g_i)}{f_i + g_i} \right| \qquad (10.5)$$

where n is the length of the vector.

- **Cross Entropy Distance**

The Cross entropy distance, where f and g are supposed independent (Juola, 2006), is given by:

$$H(f,g) = -\sum_{i=1}^{N} f(x_i).log_2 g(x_i) \qquad (10.6)$$

It has been widely used (improved version) by Juola (Juola, 2006) in his released software.

- **Intersection distance**

The intersection distance, which measures the dissimilarity between two sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the intersection-to-union ratio from 1:

$$I(f,g) = 1 - \frac{|f \cap g|}{|f \cup g|} \qquad (10.7)$$

- **Multi-Layer Perceptron (MLP)**

The MLP is a classic neural network classifier that uses the errors of the output to train the neural network (H. Sayoud, 2003). The MLP can use different back-propagation schemes to ensure the training of the classifier. The MLP is trained by the two first texts for every author, whereas the remaining text (the third one) is used for the testing task. Usually the MLP is efficient in supervised classification, however in case of local minima, we could get some errors of classification.

- **The Sequential Minimal Optimization based Support Vector Machine (SMO-SVM)**

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples into one category or the other. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The SVM is a very accurate classifier that uses bad examples to form the boundaries of the different classes (Witten et al., 1999). The SMO algorithm is used to speed up the training of the SVM (Keerthi et al., 2001).

- **Linear Regression**

The Linear Regression is the oldest and most widely used predictive model. The method of minimizing the sum of the squared errors to fit a straight line to a set of data points was published by Legendre in

1805 and by Gauss in 1809. Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norms (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression (Wiki_REG, 2013) (Huang & Pan, 2003).

10.2.2. Authorship Attribution Results

As quoted previously, 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 10.8).

$$attribution\ error = \frac{number\ of\ false\ attributions}{total\ number\ of\ testing\ segments} \qquad (10.8)$$

Table 10.1: Attribution error 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 10.1), 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. Once again, 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).

10.3 Experiments of Authorship Attribution using the COST parameter

The COST parameter has already been explained in section 7.3 with several examples. This parameter measures the degree of similarity between sentences endings.

According to the previous results of section 7.3, we could remark that for the Hadith text, there were many COST values equal to 0; and when the COST is non-null, it has very small values: the average COST is only 0.46. For the Quran, we noticed 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 interesting fact suggests the application of this type of experiment on the different text segments in order to see if there exists a stylistic difference between those segments. The resulting average COST values are represented in figure 10.3.

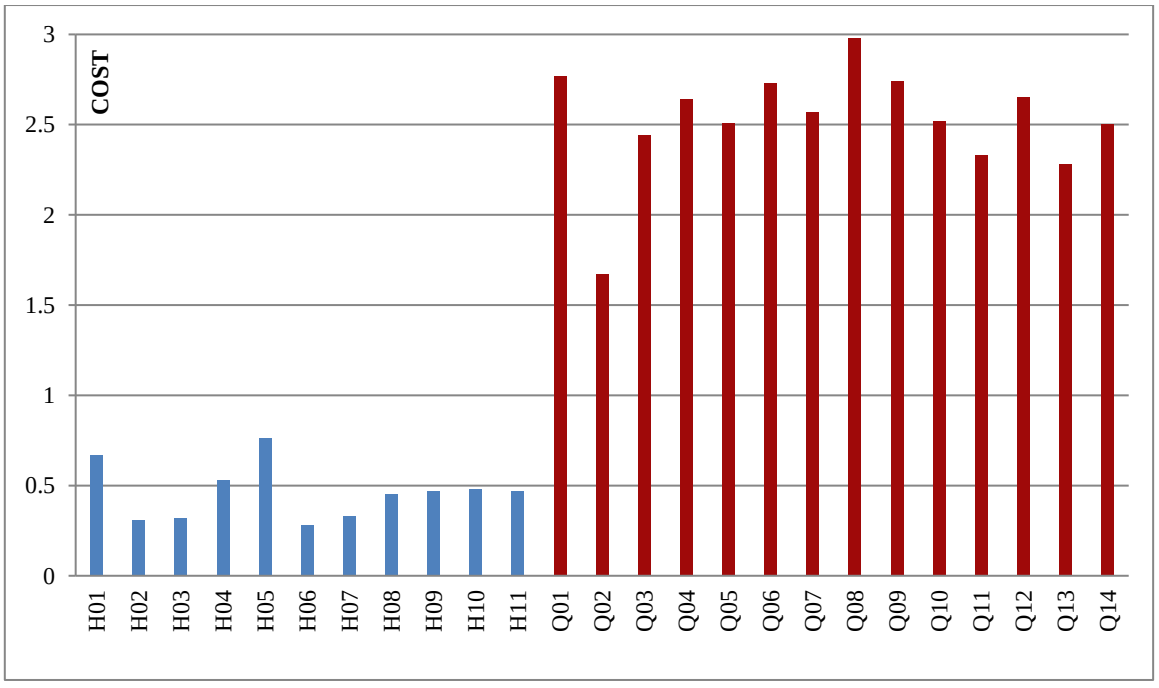


Figure 10.3: Average COST values for all text segments

Figure 10.3 shows a sharp difference between the Quran segments, which present relatively high COST values, and the Hadith segments, for which the COST values are very small. This fact implies that the structures of Quran and Hadith are different. Consequently, and since we deal with the same topic (i.e. religion), the two books should have two different author styles.

Furthermore, in order to assess the significance of the previous results, a statistical investigation on the consistency of the discrimination between the two types of segments, is made by using the Fisher's statistical exact test (Lowry 2012). Results show a two-tailed P probability that is less than 0.0001. This result shows that the association between the style and COST parameter is statistically significant.

10.4 Experiments of Authorship Attribution using the frequency of the word "الذين"

This experiment investigates the use of some words that are very commonly used in only one of the books (H. Sayoud, 2012). In practice, we remarked that the word الذين (in English: THOSE or WHO in a plural form) is very commonly used in the Quran; whereas, in the Hadith, this word is rarely used, as we can see in the following figure. Its occurrence frequency is between 0.63% and 2.02% for Quran segments, but it is between 0% and 0.29% for Hadith segments (see figure 10.4). Its average occurrence frequency is 1.3% for Quran segments and it is only 0.09% for Hadith segments (namely almost the 1/14th of the average Quran frequency).

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

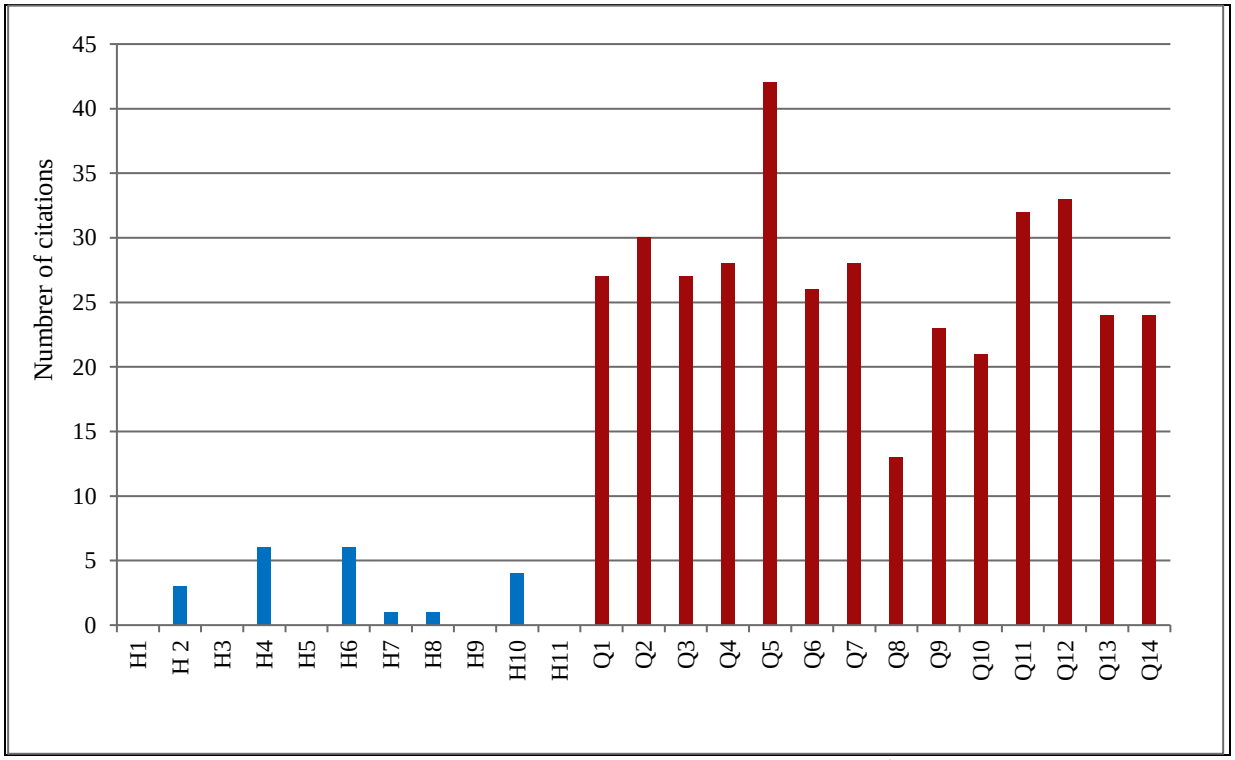


Figure 10.4: Number of citations of the word "الذين"

As previously, in order to evaluate the statistical significance of these results, a Fisher's statistical exact test (Lowry 2012) has been made to compute the discrimination consistency. We get a two-tailed P probability that is less than 0.0001. This result means that the association between style and citation number of "الذين" is considered to be statistically significant.

10.5 Discussion on the fourth series of experiment

As a continuation of the previous experiments on the authorship analysis of the holy Quran, in the present investigation we have performed a segmental analysis for the task of authorship discrimination (Tambouratzis et al., 2000) (Tambouratzis et al., 2003) between the two old Arabic religious books.

That is, four series of experiments have been made:

The first series of experiments performs a hierarchical clustering on the 25 text segments, in order to see how many possible clusters really exist and if the hypothesis of a unique author is possible.

The second series of experiments consists in an authorship attribution task, which analyses the different text segments by using several state-of-the-art features and classifiers.

The third series of experiments concerns the new COST parameter, which appears non-null only in the Quran. This parameter estimates the degree of similarity between the endings of sentences.

The fourth series of experiments investigates the use of some words that are very commonly employed in only one of the books. In this research work, the word: الذين (in English, it is translated into: THOSE or WHO in its plural form) has been chosen and analysed in the different text segments.

After observing all the experimental results and since the two books appear to have the same theme (i.e. religion), it would be reasonable to deduce the following conclusions:

- The two segmented books should have different authors (or at least two different author styles);
- All the 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.

According to these two important results, we should be able to extend our conclusions to the entire books from which the concerned segments were extracted. In fact the styles of these text segments represent the style of their corresponding original books (i.e. statistically). Consequently, it appears that the two investigated books should have different authors. Without entering in theological debates, the present investigation gives us a new scientific way to analyze and check the authorship authenticity of old or disputed documents.

11 Fifth Series of Experiment: Stylometric Comparison between the Quran and Hadith based on Successive Function Words

Usually, when one writes a text document, several function words are put within that text to make a logical link between words and to give more explanation to the major idea of the paragraph.

However, it is not very common to see two Successive Function Words (SFW) put together in a sentence, even if it is not really incorrect. For instance, instead of writing “I left my room after our discussion”, one can say in Arabic: “I left my room from-after our discussion”, which is correct in this language.

In this work, those successive function words are investigated in the Quran, Hadith and five other religious books, in order to see whether the corresponding styles are similar (i.e. Stylometric comparison between the different books).

The first results of this investigation show that the use of SFW is very discriminative between the Quran and Hadith.

11.1 Used function words

Function words are very used in Arabic (El-Zohairy, 2008) (García-Barrero et al., 2012) (Almujaiw, 2017), however successive pairs of function words cannot be translated directly to English, because of the complex translation process and the different morphologies in Arabic and in English. So, for instance, the succession of the following function words: من بعد (rough translation: From After) cannot be translated into English by keeping both prepositions, which may lead to an ambiguous meaning (i.e. From after), while the correct and most suitable translation is simply: After.

That is, 10 different types of Successive Function Words (SFW) have been selected from the Quran, Hadith and five other religious books by computing their corresponding frequencies, in a purpose of authorship discrimination (Sayoud, 2012) (Tambouratzis et al., 2004). Those successive words are represented in table 11.1 and their corresponding frequencies are given in table 11.2 and 11.3.

Table 11.1: The used Arabic successive words (SFW)

prep +prep	Translation
و إذ	And Then-When-Since
و مما	And From-what
و لقد	And Already
أَوَ مَنْ	And-do-you-think-that Who
من بعد	After / From After
ذَلِكَ لِمَنْ	That (is) for (whom/who)
ذَلِكَ بِأَنَّ	That (is) because
وَ لَئِنْ	And If
مَا لَا	What (is) Not
مِنْ قَبْل	Before / From Before

11.2 Analysis of the books based on the SFW

11.2.1. Experiment 1 - Comparison between the Quran and Hadith

The main goal of this investigation is to see whether the two religious books Quran (Nasr, 2004) (Ibrahim., 1997) and Hadith (Islahi, 1989) come from the same author or not, which could shed some light on the authorship origin of the holy Quran.

The occurrence frequencies of the quoted pairs of function words, between the Quran and Hadith, are given in table 11.2, where one can see that for the Quran the frequencies values are relatively high, while for the Hadith, the frequencies are very low and sometimes null. This difference can be better noticed in figure 11.1.

Table 11.2: Frequencies of the Arabic successive words (SFW) in the Quran and Hadith.

SFW	Frequency in the Quran $x10^{-4}$	Frequency in the Hadith $x10^{-4}$
و إذ	8.24	0.00
و مما	1.26	0.00
و لقد	14.77	0.43
أَوَ مَنْ	0.23	0.00
من بعد	15.23	0.87
ذَلِكَ لِمَنْ	0.69	0.00
ذَلِكَ بِأَنَّ	1.03	0.00
وَ لَئِنْ	3.43	0.43
مَا لَا	4.35	0.43
مِنْ قَبْلِ	10.65	0.00
Mean value	5.99	0.22

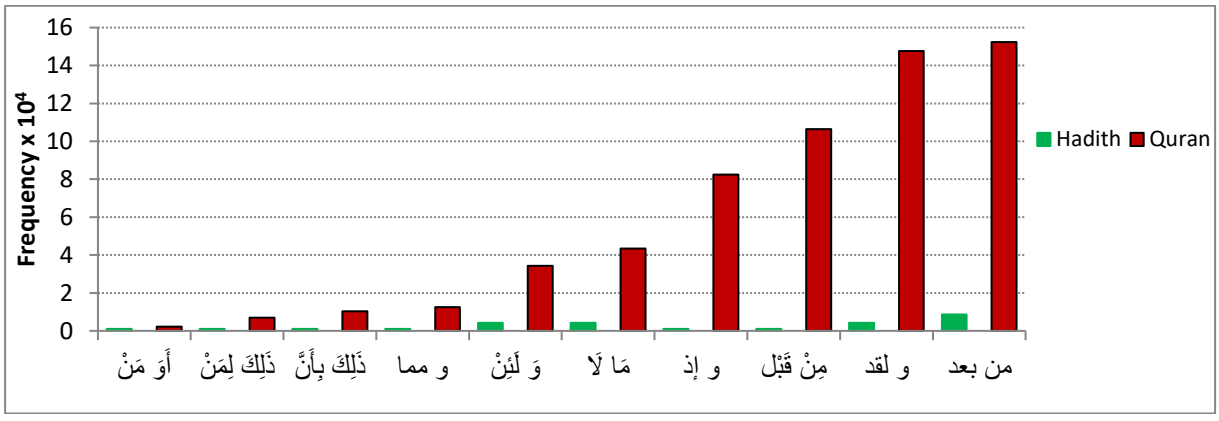


Figure 11.1: Graphical representation of the frequencies of the Arabic successive words (SFW)

We notice that the use of SFW is very different between the Quran and Hadith, showing the simple and easy style of the Hadith while for the Quran the style appears to be more sophisticated and richer by using much more successive function words.

11.2.2. Experiment 2 - Comparison of seven different books: Quran, Hadith and five other books

In this experiment, we try to compare the SFW frequencies of 7 different book belonging to different authors. The occurrence frequencies of the quoted pairs of function words, between the different books: Quran, Hadith and 5 other mixed Arabic books written by 5 religious scholars (i.e. *Abdalkafy*, *Alghazali*, *Alqaradawi*, *Alqarn* and *Amrokhaled*) (Hadjadj & Sayoud, 2018), are given in table 11.3. One can see that for the Quran, the frequencies values are relatively high, while for all other books, the frequencies are very low and sometimes null. This difference is easily noticeable in figure 11.2.

Table 11.3: Frequencies of the Arabic successive words (SFW) in the seven books.

SFW	Frequency in the Quran x10^{-4}	Frequency in the Hadith x10^{-4}	Medium Frequency in the 5 other books x10^{-4}
و إذ	8.24	0.00	0.00
و مما	1.26	0.00	0.59
و لقد	14.77	0.43	1.32
أوَ مَنْ	0.23	0.00	0.00
من بعد	15.23	0.87	0.92
ذَلِكَ لِمَنْ	0.69	0.00	0.00
ذَلِكَ بِأَنَّ	1.03	0.00	0.00
وَ لَئِنْ	3.43	0.43	0.00
مَا لَا	4.35	0.43	0.92
مِنْ قَبْل	10.65	0.00	1.76
Mean value	5.99	0.22	0.55

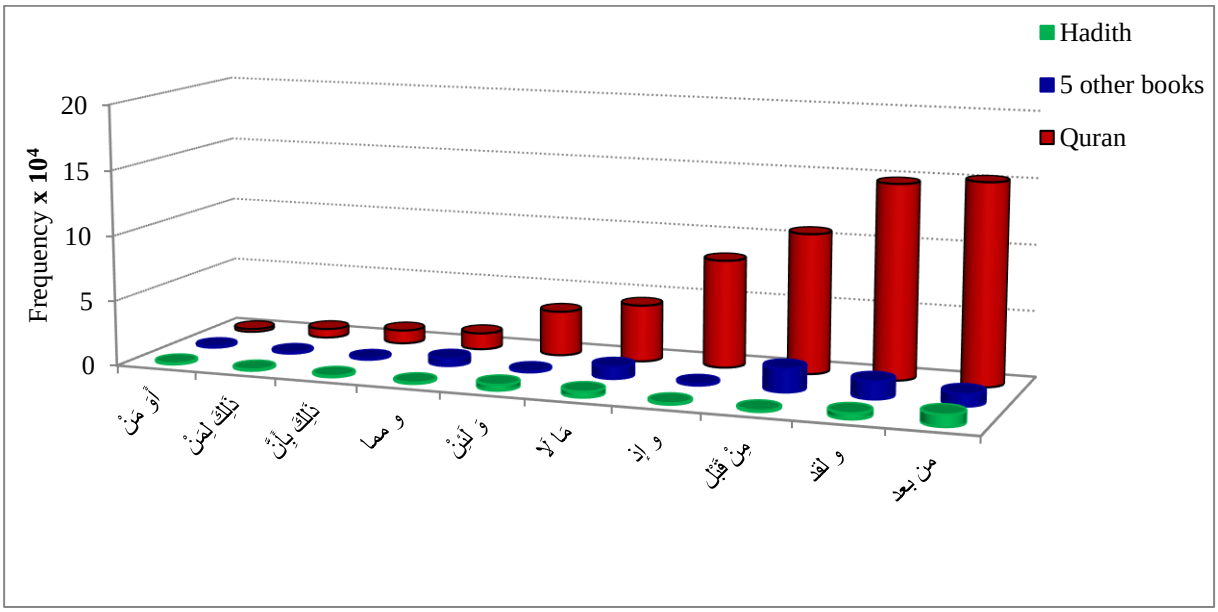


Figure 11.2: Graphical representation of the frequencies of the Arabic successive words (SFW)

As an overall consequence, one can conclude that the Quran style uses much more frequently the combination of function words than all other books (i.e. the Hadith and 5 religious books). In other

words, the SFW frequency is much higher in the holy Quran that in all other books that have been analyzed in this investigation, which involves a particular and distinct style for the holy Quran.

11.3 Discussion on this investigation

By observing figure 11.1, one can notice that the utilization of SFW in the Hadith is very rare: the mean frequency is only 0.22 10-4, while in the case of the holy Quran, the frequency is much higher: mean frequency of about 6 10-4, which is 27 times higher than the previous one.

Similarly, by observing figure 11.2, one can notice that the utilization of SFW in the 6 religious books (i.e. the Hadith and 5 other books) is very rare too. In fact, the mean frequency does not exceed 0.55 10-4 for those books, while in the case of the holy Quran, the frequency is much higher: mean frequency of about 6 10-4, which is 11 times higher than the mean frequency of the 5 religious books and 27 times higher than the frequency of the Hadith.

This noticeable difference in the use of SFW has two important interpretations:

- Firstly, the two writing styles of the Quran and Hadith appear different, with regards to the use of SFW.
- Secondly, the writing style of the Quran is even different from all the human religious books that have been investigated, with regards to the SFW.

Consequently, it appears that the holy Quran's style is not only different from the Hadith style, and then different from the Prophet style, but also different from all the Arabic styles that have been studied in this investigation. This result suggests that probably the Quran Author style is beyond the traditional human religious styles that are usually employed in the religious Arabic literature (ancient or contemporary).

12 Sixth Series of Experiment: Authorship Identification of 7 Books – A Fusion Approach

In this investigation, we conduct some experiments of automatic authorship attribution on seven Arabic religious books, namely: the holy Quran, Hadith and five other books written by five religious scholars. The Arabic styles are almost the same (i.e. Standard Arabic) for the seven books. The genre is quite the same and the topics of the different books are also the same (i.e. Religion).

The authorship characterization is based on four different features: character trigrams, character tetragrams, word unigrams and word bigrams. The task of authorship identification is ensured by four conventional classifiers: Manhattan distance, Multi-Layer Perceptron, Support Vector Machines and Linear Regression. Furthermore, a fusion approach has been proposed to enhance the performances of authorship attribution, with two fusion techniques.

A particular application is dedicated to the authorship discrimination between the Quran and Hadith, in order to see if the two books could have the same author or not.

12.1 Corpus of the seven religious books

As cited previously, there are seven different books written by seven different authors: the holy Quran, Hadith and 5 other books written by 5 religious scholars. We recall that the Arabic styles are almost the same (i.e. Standard Arabic) for the 7 books, the genre of the books is the same and the topics are also the same (i.e. Religion). 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)*). The Quran is considered to be written by Allah (God) and only sent down to the Prophet Muhammad fourteen centuries ago.

2nd book: the Hadith (*author: the Prophet Muhammad*) contains the statements of the Prophet Muhammad. As previously, in this investigation, we used the Bukhari Hadith.

The 5 other books: represent books and texts collections written by 5 religious scholar, namely: *Mohammed al-Ghazali al-Saqqa* (al-Ghazali, 2021), *Yusuf al-Qaradawi* (Alqaradawi, 2013), *Omar Abdelkafy* (Abdelkafy, 2013), *Aaidh ibn Abdullah al-Qarni* (Al-Qarni, 2021) and Amr Mohamed Helmi Khaled (Amr-Khaled, 2013).

Those seven books are preprocessed and segmented into different and distinct text segments. Every segment is about 2900 tokens each. Here are the numbers of segments by book:

Table 12.1: Books specifications of ***SAB-1*** dataset.

Book/Author	Number of segments by book*	Big/ Small parameter#	Training set size	Testing set size
1st book: the holy Quran	30 segments	Big	7	23
2nd book: the Hadith	8 segments	Small	4	4
3rd book: books of Alghazali	39 segments	Big	7	32
4th book: books of AlQuaradhawi	13 segments	Small	4	9
5th book: books of Abdelkafy	10 segments	Small	4	6
6th book: books of Aid Alkarny	23 segments	Big	7	16
7th book: books of Amrokhaled	9 segments	Small	4	5

**Each segment is composed of 2900 tokens.*
#Big/Small is a logical parameter (i.e. binary value).

The corpus is decomposed into 2 parts: training part and testing part, and since the different books have different sizes, an optimal logical rule has been established: 4 text segments are used for the training of small books and 7 text segments are employed for the training of big books. The main reasons for this choice are explained here below.

The choice of the training dataset size is defined by a particular logical (binary) parameter we called Big/Small, which gives a qualitative estimation on the size of the book. That is, if the size of the book is over 20 segments, then it is considered as a big dataset otherwise it is considered small. The value or the threshold 20 is equal to the half size of the biggest dataset (ie. 39 segments for Alghazali book, which implies a threshold of 39/2 ≅20). This scheme permits us to have different possible sizes for the training dataset.

By observing the small books, we notice that "4 text segments" should be a good choice for the small books. In fact, the value 4 is equal to the half size (50%) of the smallest book (ie. the smallest book contains only 8 segments).

By observing the seven books, we notice that "7 text segments" should be a good choice too for the big books. In fact, the value 7 is equal to the maximum size of the training set for the smallest book (ie. a maximum of 7 segments for the training, since we require at least 1 segment for the testing).

These two training rules could be applied to the different books with regards to the parameter Big/ Small. But even though, the value 7 is a limit that we cannot exceed (and could be seen as a fixed choice), we cannot say that the value 4 is optimal for small texts: why not 3 or 5 text segments, for instance.

In order to check if this choice was judicious or not, (experimentally speaking), we did some experiments of authorship attribution on another corpus consisting of 7 different books (from a second different dataset called SAB-2) denoted by A, B, C, … G and where the sizes of the books are very similar to those of the previous one: SAB-1 dataset (see table 12.2). The used classification technique was based on the Manhatan Centroid distance.

Table 12.2: Features of the second dataset: ***SAB-2****

		Case 1	Case 2	Case 3
Book designation	Big/ Small dataset	Training set size	Training set size	Training set size
Book A	Big	**7**	**7**	**7**
book B	Small	**3**	**4**	**5**
book C	Big	**7**	**7**	**7**
book D	Small	**3**	**4**	**5**
book E	Small	**3**	**4**	**5**
book F	Big	**7**	**7**	**7**
book G	Small	**3**	**4**	**5**

** Note that the corpus SAB-2 will no longer be utilized in the next sections.*

Hence, three cases are investigated:

- Case 1: **3** text segments are used for the training of small books and 7 text segments are employed for the training of big books;
- Case 2: **4** text segments are used for the training of small books and 7 text segments are employed for the training of big books;
- Case 3: **5** text segments are used for the training of small books and 7 text segments are employed for the training of big books.

The different results of authorship attribution, got on this second dataset, are summarized in the following table:

Table 12.3: Results of authorship attribution got on the second dataset: ***SAB-2****

Training size			Score of good attribution in % (experiments conducted on another corpus)							
Case	Big	Small	Char. Bi-gram	Char. Tri-gram	Char. Tetra-gram	Word	Word Bi-gram	Word tri-gram	Word Tetra-gram	Average performance in %
Case 1	7	3	74.74	83.83	89.89	94.94	94.94	32.32	33.33	63.88
Case 2	7	4	76.84	89.47	91.57	93.68	97.89	54.73	31.57	69.47
Case 3	7	5	76.92	85.71	89.01	95.6	97.8	35.16	32.96	65.38

** Note that the corpus SAB-2 will no longer be utilized in the next sections.*

According to this table (table 12.3), the case 2, corresponding to 4 training texts for the small books, seems to be the most interesting case. That is, by observing the average performance given by Manhattan distance, we can easily see that the best average score is 69.47%, which corresponds to the second case (ie. 4 text segments for small books and 7 ones for big books). According to this result, the chosen training configuration seems to be judicious and interesting for the authorship attribution experiments conducted on the first dataset.

However, we should note that we cannot expand this result to other classifiers like machine learning ones, especially those which need a great amount of training data, such as neural networks or support vector machines, for instance.

12.2 Authorship Attribution Methods

Several experiments of authorship attribution are conducted on the 7 segmented religious books.

For a purpose of feature selection and evaluation (Hawashin et al., 2013), four types of characteristics are employed: character-trigram, character tetra-gram, word and word-bigram. Two of these features are based on characters and the two others are typically lexical.

Also, four different classifiers are used for the automatic authorship classification (into ideally 7 different classes), where every class should represent one particular author. The different classifiers are defined as follows:

- Manhattan centroid distance;
- Multi Layer Perceptron;
- SMO based Support Vector Machines;
- Linear Regression.

Furthermore, a Fusion approach is proposed to try enhancing the attribution accuracy of the conventional classifiers/features.

12.2.1 Conventional Classifiers

The 4 conventional classifiers are described here below.

- **Manhattan distance**

This distance (H. Sayoud, 2012) is very reliable in text classification. The corresponding distance between two vectors X and Y is given by the following formula:

$$d_{X,Y} = \sum_{i=1}^{n} |X_i - Y_i| \quad (12.1)$$

where n is the length of the vector.

In this investigation, the different samples of the training are employed to build the centroid vector, which will be used, as reference, to compute the required distance with the previous formula (also called KNN method). Manhattan distance is simple to implement and very efficient for text classification.

- **Multi-Layer Perceptron (MLP)**

The MLP (Multi-Layer Perceptron) is a classical neural network classifier that uses the errors of the output to train the neural network (H. Sayoud, 2003). The MLP can use different back-propagation schemes to ensure the training of the classifier. It is trained by the different texts of the training set, whereas the remaining texts are used for the testing task. Usually the MLP is efficient in supervised classification, however in case of local minima; we usually can get some errors of classification.

- **Sequential Minimal Optimization based Support Vector Machine (SMO-SVM)**

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, which are used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples into one category or the other. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

The SVM is a very accurate classifier that uses bad examples to form the boundaries of the different classes (Witten et al., 1999). Concerning the Sequential Minimal Optimization (SMO) algorithm, it is used to speed up the training of the SVM (Keerthi et al., 2001).

- **Linear Regression**

Linear Regression is the oldest and most widely used predictive model. The method of minimizing the sum of the squared errors to fit a straight line to a set of data points was published by Legendre in 1805 and by Gauss in 1809. Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the “lack of fit” in some other norms (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression (Wiki_REG, 2013) (Huang & Pan, 2003).

12.2.2 The Fusion approach

In order to enhance the authorship attribution performance, we have proposed the use of several classifiers, which are combined in order to get a lower identification error: this combination is technically called Fusion (Tchechmedjiev et al., 2013).

Theoretically, the fusion can be performed at different hierarchical levels and forms. A very commonly encountered taxonomy of data fusion is given by the following techniques (Jain et al., 2004) (Dasarathy, 1994) (Verlinde, 1999):

- Feature level where the feature sets of different modalities are combined. Fusion at this level provides the highest flexibility but classification problems may arise due to the large dimension of the combined (*concatenated*) feature vectors.

- Score (*matching*) level is the most common level where the fusion takes place. The scores of the classifiers are usually normalized and then they are combined in a consistent manner.

- Decision level where the outputs of the classifiers establish the decision via techniques such as majority voting. Fusion at the decision level is considered to be rigid for information integration (Stylianou & et al., 2005), but it is not complicated in implementation.

In this investigation, we propose the use of the third technique, namely the decision level based fusion. Furthermore, two types of combinations are employed: combination of features, called FDF or Feature-based Decision Fusion, and combination of classifiers, called CDF or Classifier-based Decision Fusion.

- Feature-based Decision Fusion (FDF): In the first proposed fusion (*combination of several features*), three different features are employed:

-Character-tetragram;

-Word;

-Word Bigram.

The fusion technique fuses the different corresponding scores of decision into one decision (*the final decision*). The chosen classifier is *Manhatan centroid* because it has shown excellent performances during the previous experiments.

It is called Feature-based Decision Fusion or FDF (*see figure 12.1*) and consists in fusing the outputs of the different classifiers according to a specific vote provided by their different decisions: each decision concerns one feature *Fj*.

The fused decision D_f of N features is given by the following equation:

$$Decision = D_f,\ with\ f = argmax_j(freq(D_j)) \qquad (12.2)$$

freq denotes the occurrence frequency of a specific decision and j=1..N.

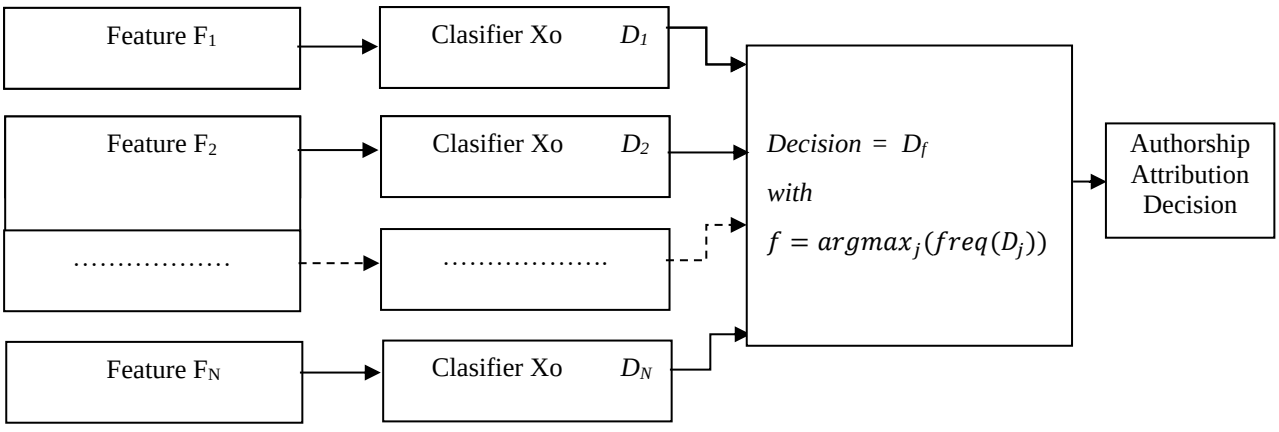


Figure 12.1: Principle of the Feature-based Decision Fusion (FDF)

- Classifier-based Decision Fusion (CDF): In the second proposed fusion (*combination of several classifiers*), three different classifiers are employed:

-Manhattan centroid;

-SMO-SVM;

-MLP.

As previously, the fusion technique fuses the different corresponding scores of decision into one decision (the final decision). Concerning the choice of the features, the word descriptor has been used because it has been shown that this type of feature presented relatively good performances during our experiments.

It is called Classifier-based Decision Fusion or CDF (*see figure 12.2*) and consists in fusing the outputs of the different classifiers according to a specific vote provided by their different decisions: each decision concerns one classifier *Cj*.

The fused decision D_f of M classifiers is given by the following equation:

$$Decision = D_f,\ with\ f = argmax_i(freq(D_i)) \quad (12.3)$$

freq denotes the occurrence frequency of a specific decision and i=1..M.

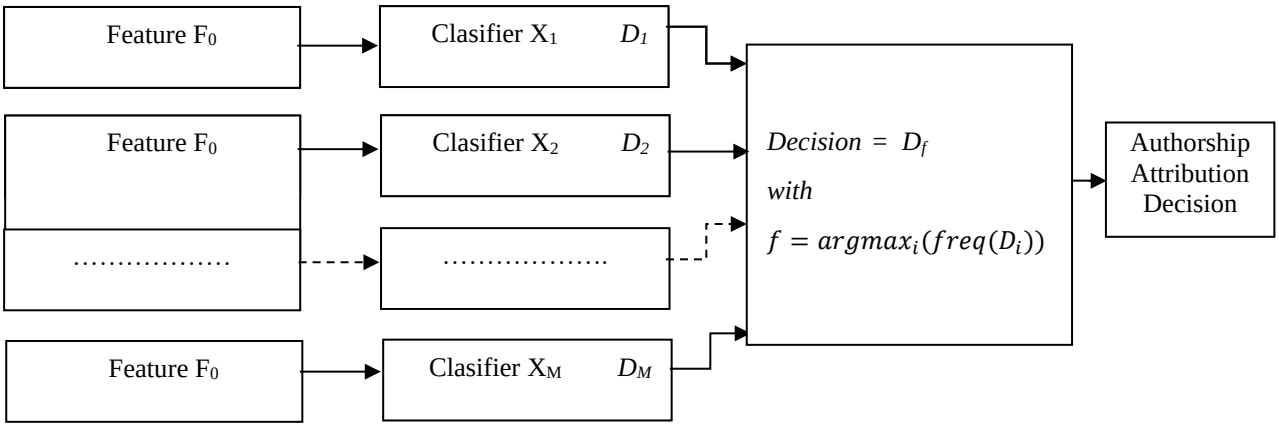


Figure 12.2: Principle of the Classifier-based Decision Fusion (CDF)

All the results of the fusion approach are represented in tables 12.12 and 12.13, summarizing the corresponding AA scores of the first and second fusion techniques respectively.

12.3 Experiments of authorship attribution

As mentioned previously, seven Arabic religious books are investigated and analyzed in order to make a classification of the text documents per author: the experimented corpus is called ***SAB-1***. We also recall that several features and several classifiers are used in the experiments of authorship attribution.

12.3.1 Experiments of authorship attribution using conventional features and classifiers

In this section we report the different results obtained by using conventional classifiers and features. The different experimental results are organized into 4 tables (*table 12.8, 12.9,12.10 and 12.11*):

- Table 12.8 displays the different results obtained with the Character-trigram feature;
- Table 12.9 displays the different results obtained with the Character-tetragram feature;
- Table 12.10 displays the different results obtained with the Word (*Word-unigram*) feature;
- Table 12.11 displays the different results obtained with the Word-bigram feature**.**

Those tables display the errors of authorship attribution given by the 4 classifiers: Manhatan centroid, MLP, SMO-SVM and Linear Regression. Furthermore, a column untitled "Total identification error" summarizes the overall error of attribution for the 7 books. This indication gives us an interesting idea on the overall performances of authorship attribution (*corresponding to a specific feature*).

Table 12.8: Identification Error in % using the feature: *Character-trigram*, on *SAB-1* dataset.

		Total Identification error on the 7 books	**The holy Quran book**	**The Hadith book**	**Aaid's book**	**Abd-elkafy's book**	**Algha-zali's book**	**Alquara-dawi's book**	**Amro-Khaled's book**
Date / Century			Ancient : 6th century	Ancient: 6th century	Recent: 20th century	Recent: 20th century	Recent : 20th century	Recent: 20th century	Recent: 20th century
Classifier	**Manhatan centroid distance**	**4.2%**	0%	0%	12.5%	0%	0%	22.2%	0%
	MLP classifier	**3.1%**	0%	0%	0%	16.7%	0%	22.2%	0%
	SMO-SVM classifier	**4.2%**	0%	0%	0%	33.3%	0%	22.2%	0%
	Linear Regression	**4.2%**	0%	0%	6.25%	16.7%	0%	22.2%	0%

In table 12.8, we can see that the best classifier is the MLP, which gives an error of only 3.1% (look at the 1st columbn), the other classifiers have the same performances (*total identification errors of 4.2%*). The two authors: Abdelkafy and Alquaradawi present some problems of authorship attribution, with respectively 16.7% and 22.2.% in the case of the MLP. These two authors are often confused with other authors. Note that the Quran and Hadith books are attributed without any error (*error of 0%*).

Table 12.9: Identification Error in % using the feature: *Character-tetragram*, on *SAB-1* dataset.

		Total Identification error 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**
Date / Century			Ancient 6th century	Ancient 6th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century
Classifier	**Manhatan centroid distance**	**1.05%**	0%	0%	0%	0%	0%	11.1%	0%
	MLP classifier*	**2.1%**	0%	0%	6.25%	16.7%	0%	0%	0%
	SMO-SVM classifier*	**3.1%**	0%	0%	12.5%	16.7%	0%	0%	0%
	Linear Regression*	**2.1%**	0%	0%	6.25%	16.7%	0%	0%	0%

**: 500 most frequent features only.*

In table 12.9, we can see that the best classifier is Manhattan distance, which gives an error of only 1.05%, the other classifiers present different performances (*total identification errors ranging between 2.1% and 3.1*). The three authors: Aaid-Alkarni, Abdelkafy and Alquaradawi present some problems of authorship attribution depending on the choice of the classifier. These two first ones are often confused with other authors. As previously, we can note that the Quran and Hadith books are attributed without any error (*error of 0%*).

Table 12.10: Identification Error in % using the feature: *Word*, on *SAB-1* dataset.

		Total Identification error 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**
Date / Century			Ancient 6th century	Ancient 6th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century
Classifier	**Manhatan centroid Distance**	**1.05%**	0%	0%	6.25%	0%	0%	0%	0%
	MLP classifier*	**1.05%**	0%	0%	0%	16.7%	0%	0%	0%
	SMO-SVM classifier*	**2.1%**	0%	0%	0%	0%	33.3%	0%	0%
	Linear Regression*	**2.1%**	0%	0%	6.25%	16.7%	0%	0%	0%

**: 500 most frequent features only.*

In table 12.10, we can see that the best classifiers are the MLP and Manhattan distance, which give an error of only 1.05%, the other classifiers present the same performances (*total identification errors of 2.1%*). The two authors: Aaid-Alkarni and Abdelkafy present some problems of authorship attribution depending on the choice of the classifier. These two particular authors are often confused with other authors. Also, as in the previous results, we can note that the Quran and Hadith books are attributed correctly (*error of 0%*).

Table 12.11: Identification Error in % using the feature: *Word Bigram*, on *SAB-1* dataset.

		Total Identification error on the 7 books	**The holy Quran book**	**The Hadith book**	**Aaid's book**	**Abd-elkafy's book**	**Algha-zali's book**	**Alquara-dawi's book**	**Amro-Khaled's book**
Date / Century			Ancient 6th century	Ancient 6th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century
Classifier	**Manhatan centroid distance**	**1.05%**	0%	0%	0%	0%	3.1%	0%	0%
	SMO-SVM classifier#	**3.1%**	0%	0%	12.5%	16.7%	0%	0%	0%
	MLP classifier#	**4.2%**	0%	0%	12.5%	33.3%	0%	0%	0%
	Linear Regression #	**4.2%**	0%	0%	12.5%	16.7%	0%	0%	20%

#: 600 most frequent features only.

In table 12.11, we can see that the best classifier is Manhattan distance, which gives an error of only 1.05%, the other classifiers present different performances (*total identification errors ranging between 3.1% and 4.2%*). The three authors: Aaid-Alkarni, Abdelkafy and Alghazali present some problems of authorship attribution depending on the choice of the classifier. Again, these two first ones are often confused with other authors. Once again, as in the previous results, we can note that the Quran and Hadith books are attributed without any error (*error of 0%*).

Note: we notice that Manhatan centroid distance, which is a relatively simple statistical classifier, outperforms the other machine learning classifiers in many cases. However we do know that these last ones are usually better than the distance based classifiers especially for the SVM classifier, which is considered as the state-of-the-art classifier in many research fields. The main possible reason is the low dimensionality of the training dataset, which usually leads to a weak training process (*note that some books are too small with only 8 or 9 texts per book: this fact makes difficult to get a big training dataset*).

12.3.2 Experiments of authorship attribution using fusion techniques

In order to further enhance the authorship attribution performances, two fusion techniques have been proposed and implemented: the FDF and CDF fusion techniques. We can see in tables 12.12 and 12.13 the corresponding results of those two fusion techniques respectively.

Table 12.12: Error of identification using the *feature-based fusion (FDF)*

Total Identification error on the 7 books	The holy Quran	The Hadith	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
0%	0%	0%	0%	0%	0%	0%	0%

Table 12.13: Error of identification using the *classifier-based fusion (CDF)*

Total Identification error on the 7 books	The holy Quran	The Hadith	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
0%	0%	0%	0%	0%	0%	0%	0%

As we can see in tables 12.12 and 12.13, the authorship attribution error is equal to zero for every author. Te total identification score is 100%, showing the superior performances of the fusion techniques over the conventional classifiers as expected in theory. This result is very interesting since it shows that a combination of different features and/or classifiers can lead to high authorship attribution performances.

12.4 Discussion and comments

By observing the different experimental results, we can see that the 7 different books have been discriminated (*let us say*) correctly with regards to the writer/author: the corresponding text segments have been attributed to the correct authors with a small error of identification. Moreover, by using the fusion approach the attribution error have been reduced to only 0%. This important result shows that the classical features and classifiers that are usually employed in English and Greek languages got good results for the Arabic language too and appear to be utilizable for the authorship attribution of texts that are written in Arabic.

The first conclusion we can state is that the fusion approach is quite interesting in multi-classifier or multi-feature authorship attribution.

Another important conclusion, one can deduce, is that the two religious books Quran and Hadith appear to have two different Authors. In fact, all the experiments have shown that the Quran segments and Hadith segments are identified and classified separately without any attribution error, with or without the use of fusion.

Now, for a reason of statistics, let us have a look at table 12.1. This table shows the following indications on the cross-combinations of the authorship identification tests: the total number of possible cross-combinations tests $Ntest_{QH}$, between Quran segments and Hadith segments, is equal to:

$$Ntest_{QH} = (23 \times 4) + (4 \times 7) = 92 + 28$$
$$= 120 \text{ different combinations.} \quad (12.4)$$

So, by supposing that at the 121th combination we will get an error of attribution between one Quran segment and one Hadith segment, then the probability to get an error of attribution would be 1/121= 0.8%.
Since this is only a supposition and not a real attribution error, then we can say that the probability of authorship confusion between the two religious books is less than 0.8% in the present experimental conditions.

Furthermore, since there are no cross-errors of attribution between the Quran and Hadith texts (*each other*) and more generally, since there was not any cross-error of attribution for the Quran texts or Hadith texts, with regards to the 6 other investigated books, we can state that these 2 books are completely different in style each other and also different from all the other investigated books. Consequently, it appears that the Quran and Hadith should have two different authors or at least two different author styles.

13 Seventh Series of Experiments: Authorship Discrimination using the Leave-One-Out Validation

In this survey, we employ the Leave-One-Out cross-validation (LOO) in the stylometric analysis of the two religious books (i.e. Quran and Hadith). In fact, although conventional classification approaches were widely employed in the literature, they still present a lack of statistical consistency. The LOO and LTO cross-validation techniques consist in different experiments of authorship attribution that are carried out in a rotating manner, excluding every time one or two new samples.
Hence, two experiments of authorship classification are made on the two religious books, which are expected to be fair and significant. The books are segmented into distinct text segments of 2900 tokens each, and the used features are composed of character-tetragrams, which are known to be quite efficient in stylometry.
The proposed discrimination approach is based on the Leave-One-Out (LOO) and Leave-Two-Out (LTO) cross-validation techniques using a Support Vector Machines (SVM) based classifier.

13.1 The Leave-One-Out Method

The Leave-One-Out Method is a jackknife method for evaluating the classification accuracy (Vehtari 2016). It was proposed by Lachenbruch in 1967 (Lachenbruch, 1967). His approach was based on discriminant analysis; it has been named the leave-one-out (LOO) method (Huberty, 1994). This technique has two steps:

-First, the template is built in the samples with one observation removed,

-Then the resulting estimate parameters (of the training) are used to classify the single removed observation.

The main process is repeated M times so that each observation was removed and classified once (see Figure 13.1), where M represents the number of samples (Kroopnick et al., 2010).

Eventually, the proposed measure of good classification is given by the number of times that the removed observation was correctly classified (Huberty, 1994) (Kroopnick et al., 2010).

To evaluate the LOO method, Lachenbruch conducted a small Monte Carlo simulation with 300 replications for a two group discriminant analysis. His results showed the efficiency of Lachenbruch's LOO technique (Kroopnick et al., 2010).

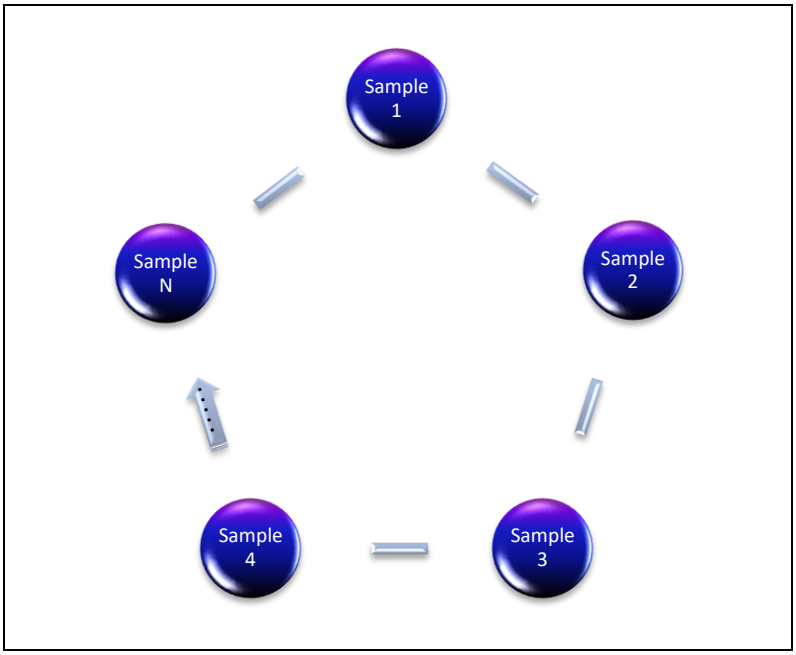


Figure 13.1.a. Set of the N samples to classify (Sample 1, Sample 2, … Sample N).

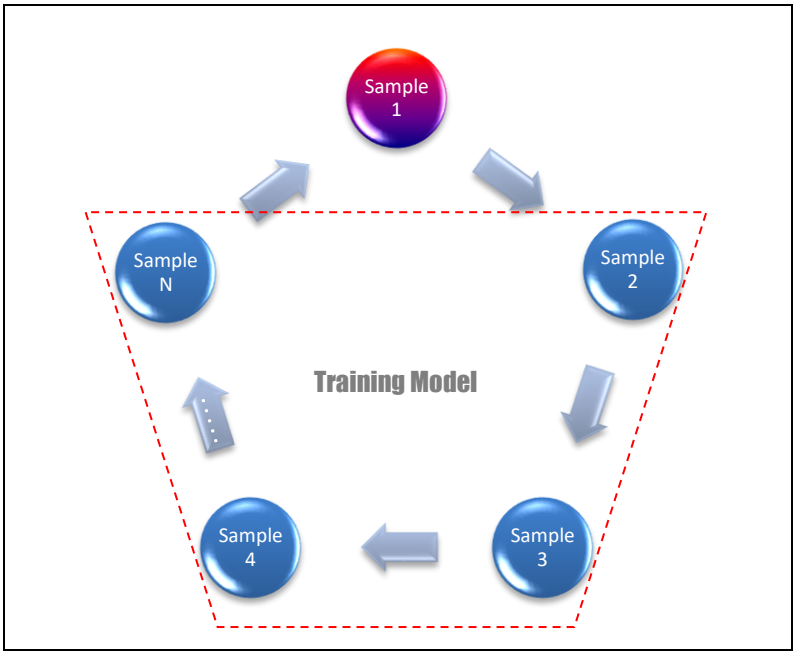


Figure 13.1.b. The Leave-One-Out algorithm applied to Sample 1 (start of the algorithm).

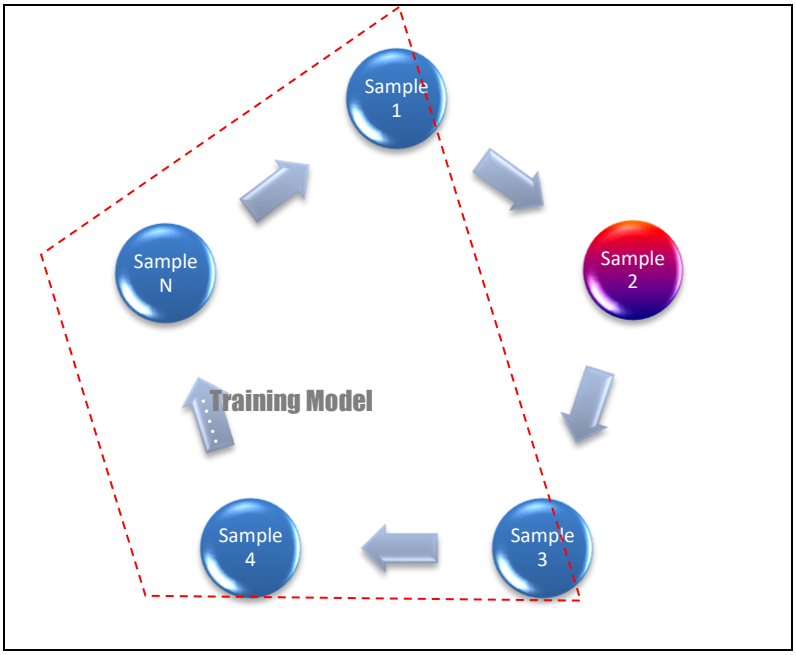


Figure 13.1.c. The Leave-One-Out algorithm applied to Sample 2 and moving to the next sample.

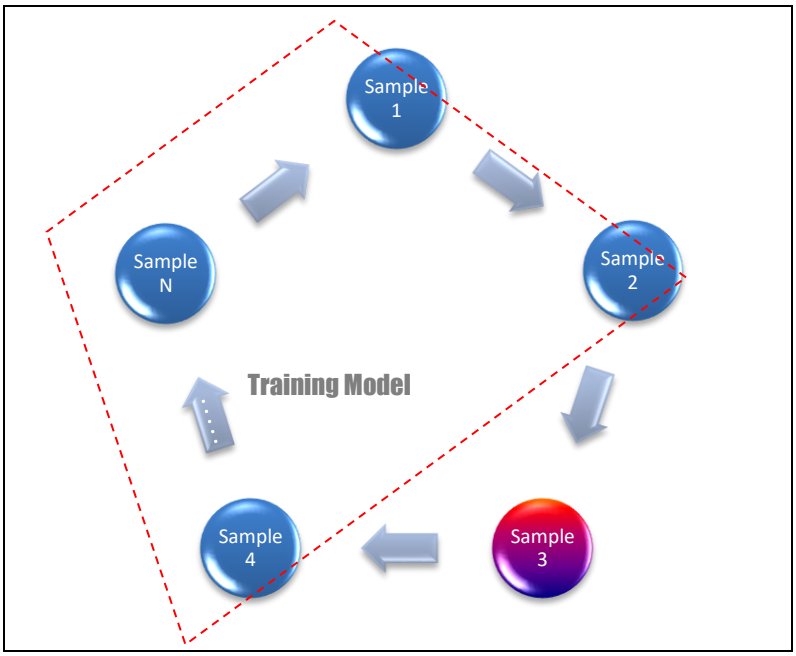


Figure 13.1.d. The Leave-One-Out algorithm applied to Sample 3 and moving to the next sample.

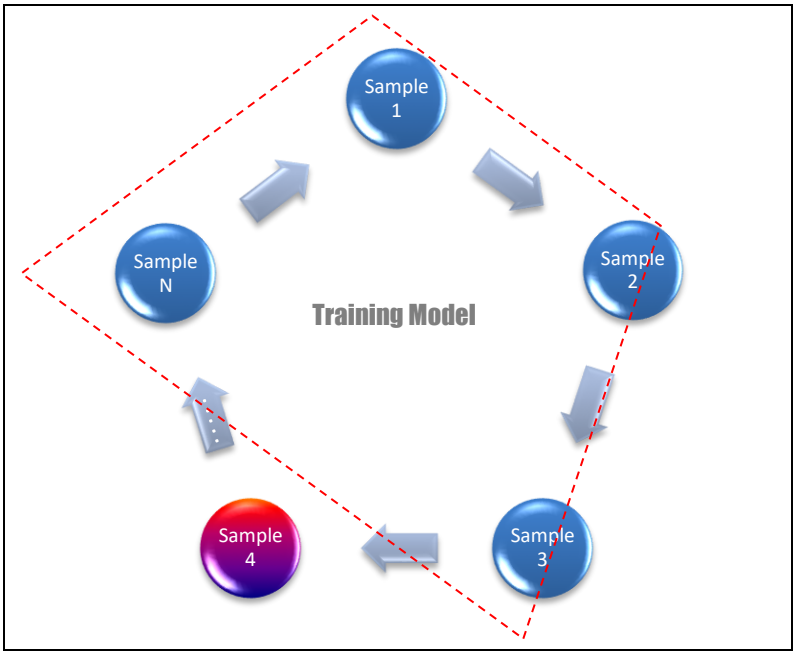


Figure 13.1.e. The Leave-One-Out algorithm applied to Sample 4 and moving to the next sample.

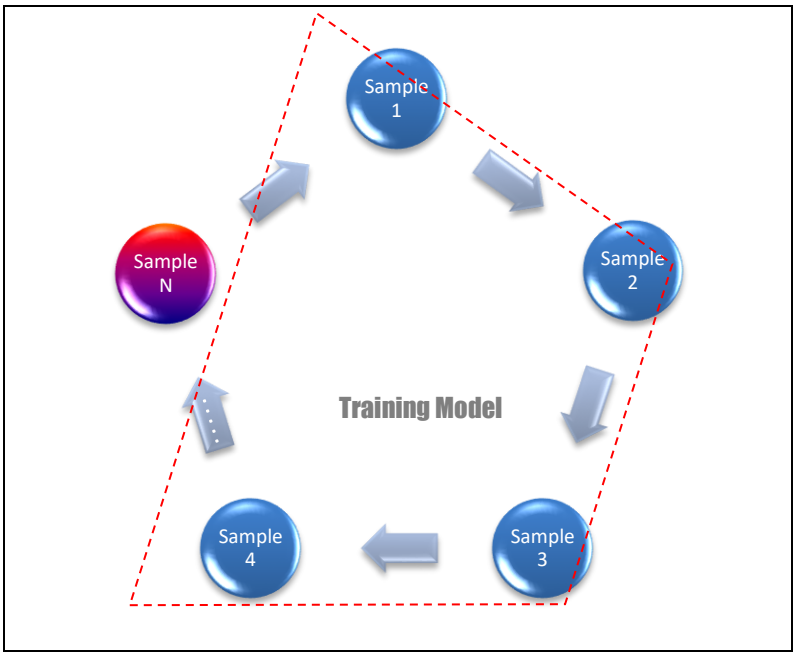


Figure 13.1.f. The Leave-One-Out algorithm applied to Sample N (end of the algorithm).

Similarly, the Leave-Two-Out is an extension of the Leave-One-Out algorithm, where instead of excluding only one sample, we should exclude two samples simultaneously. It is referred to by the abbreviation LTO.

13.2 On the Choice of the Classifier and Feature

Concerning the task of authorship attribution (Ouamour et al., 2013), we previously investigated the authorship of several short historical texts that were written by ten ancient Arabic travelers: called AAAT dataset. Several experiments of authorship attribution were conducted on those Arabic texts, by using seven different classifiers, namely: Manhattan distance, Cosine distance, Stamatatos distance, Camberra distance, Multi-Layer Perceptron (MLP), Sequential Minimal Optimization based Support Vector Machine (SMO-SVM) and Linear Regression. The results showed that the best performances of authorship attribution were given by the SVM, which outperformed the other investigated classifiers.

For this reason, and knowing the good performances of the SVM in discrimination, we have decided to use this classifier for the task of authorship discrimination.

Concerning the features, in the present investigation, we have used the character-tetragrams.

13.3 Text Segmentation

A text segmentation is applied in order to construct individual documents with the same size. In fact, when comparing two books with different sizes, it is difficult to know if a specific part of the book is similar to another one or different. That is why a smart segmentation has been proposed and applied to the different books.

The sizes of the segments are more or less in the same range: we obtained 29 different text segments for the Quran and 8 different text segments for the Hadith, with approximately the same size. So, we got 37 different text segments of about 2900 words each in the whole dataset. Our chosen configuration seems to be correct and suitable to the different AA experiments.

The segmented dataset is decomposed into 2 rotating Leave-K-Out parts: the training part containing all the text samples except K (i.e. one or two), and the testing part consisting in the removed ones.

13.4 Experiments of AA using the LOO and LTO techniques

We recall that we used the character-tetragram feature by keeping only the 500 most frequent features, and employed the SMO-based SVM classifier.

Since there are 37 samples, we will also have 37 experiments of rotating classification, where in every experiment one sample (or two) is removed and put in testing set, in order to be identified through the remaining samples that represent the training model.

In the following table, we represent the scores of good classification obtained by the first series of experiments of LOO, which corresponds to the 37 LOO cross validation experiments.

Table 16.1: Results of AA using the LOO technique

Experiment Number	Tested document	Accuracy
1.	Q1	100%
2.	Q2	100%
3.	Q3	100%
4.	Q4	100%
5.	Q5	100%
6.	Q6	100%
7.	Q7	100%
8.	Q8	100%
9.	Q9	100%
10.	Q10	100%
11.	Q11	100%
12.	Q12	100%
13.	Q13	100%
14.	Q14	100%
15.	Q15	100%
16.	Q16	100%
17.	Q17	100%
18.	Q18	100%
19.	Q19	100%
20.	Q20	100%
21.	Q21	100%
22.	Q22	100%
23.	Q23	100%
24.	Q24	100%
25.	Q25	100%
26.	Q26	100%
27.	Q27	100%
28.	Q28	100%
29.	Q29	100%
30.	H1	100%
31.	H2	100%
32.	H3	100%
33.	H4	100%
34.	H5	100%
35.	H6	100%
36.	H7	100%
37.	H8	100%

In the following table, we represent the scores of good classification obtained by the second series of experiments of LTO, which corresponds to the 19 LTO cross validation experiments.

Table 16.2: Results of AA using the LTO technique

Experiment Number	Tested documents	Accuracy
1.	Q1,2	100%
2.	Q3,4	100%
3.	Q5,6	100%
4.	Q7,8	100%
5.	Q9,10	100%
6.	Q11,12	100%
7.	Q13,14	100%
8.	Q15,16	100%
9.	Q17,18	100%
10.	Q19,20	100%
11.	Q21,22	100%
12.	Q23,24	100%
13.	Q25,26	100%
14.	Q27,28	100%
15.	Q29,30	100%
16.	H1,2	100%
17.	H3,4	100%
18.	H5,6	100%
19.	H7,8	100%

It will be now interesting to compute the average accuracy, corresponding to the overall performances of classification. This entity can be evaluated by using equation 16.1.

$$\textbf{Average Accuracy} = \frac{\sum_{i=1}^{N} CrossVal_i}{N} \qquad (16.1)$$

where *N* represents the number of cross-validation experiments (denoted by *CrossVal*).

According to table 16.1, the average accuracy of all LOO experiments is equal to **100%**.

And, according to table 16.2, the average accuracy of all LTO experiments is equal to **100%**.

13.5 Discussion

The two ancient religious books (i.e. Quran and Hadith) have been analysed by a discriminative authorship analysis using the Leave-One-Out and Leave-Two-Out validation techniques. The features consist in character-tetragrams, while the used classification is based on an SMO-SVM classifier. The dataset is composed of 37 text documents, where the size of a single segment is about 2900 tokens.

As we could see in the results section, the accuracy of every cross-validation step, for all LOO and LTO experiments, was 100%, leading to an average cross-validation score of 100% too.

From these results, one can deduce the following important conclusions:

- Firstly, the two books Quran and Hadith appear to have two different author styles;
- The segments of every book are quite similar in terms of style within a single book;
- The LOO and LTO cross validation techniques have shown that this result (discrimination score of 100%) is quite significant, since the same score has been obtained 37+19 times, namely 56 times during the tests of cross-validation and with different configurations.

Consequently and according to this investigation, it is clear that the two ancient books: Quran and Hadith possess two different styles and should probably come from two different Authors. Finally, and once again, one can deduce that the Quran could not be written by the Prophet.

14 Eighth Series of Experiments: Authorship Discrimination based on Gaussianity and Interpolability

It is well known that every physical phenomenon in the universe respects some specific rules, as we can scientifically observe and measure. For instance, the Gaussianity rule characterizes every physical phenomenon respecting the "Large Numbers" condition. On the other hand, the Interpolability can be noticed for almost every discrete curve representing a natural physical phenomenon. Those two rules are respected by a wide variety of data present in the Universe or let us say simply in our daily life.

In the case of textual data, the two rules should theoretically be respected, and indeed, they have been verified during this investigation.

However, in the case of the holy Quran, this study shows that the Gaussianity and interpolability of the data curve do not appear to be respected. Furthermore, we notice an inexplicable and strange statistical structure in the holy book, without any (prior) scientific interpretation.

14.1 Introduction

Most of existing natural data obey a certain set of physical rules that seem to be quite simple, mathematically speaking. For instance, the gravity equation is basically simple (W=m.g); the energy equation is also simple ($E=mc^2$) and the electric voltage equation is further simpler (U=RI). This mathematical simplicity in most of the natural or physical data is very often verified.

On the other hand, a famous theorem in statistics, called "Central Limit Theorem" or sometimes "Theorem of Large Numbers", stipulates (after a rigorous demonstration) that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed (i.e. Gaussian distribution), regardless of the underlying distribution [Siegrist, 2016] [Rice, 1995]. That is, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to the normal distribution (i.e. Gaussian distribution) (Contributors, 2015).

The central limit theorem has an interesting history. The first version of this theorem was postulated by the French mathematician De Moivre who, in a remarkable article published in 1733, used the normal distribution to approximate the distribution of the number of heads resulting from many tosses of a fair coin. This finding was far ahead of its time, and was nearly forgotten until the French mathematician Laplace rescued it from obscurity in his work called "Théorie analytique des probabilities", which was published in 1812. Laplace expanded De Moivre's finding by approximating the binomial distribution with the normal distribution. But as with De Moivre, Laplace's finding received little attention in his own time. It was until the end of nineteenth century that the importance of the central limit theorem was discerned, when, in 1901, the Russian mathematician Aleksandr Lyapunov defined it in general terms and proved precisely how it worked mathematically. Nowadays, the central limit theorem is considered to be the sovereign of probability theory (Contributors, 2015).

Sir Francis Galton described the Central Limit Theorem as (Galton, 1889): <<…It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The larger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along>> (Contributors, 2015).

Another important aspect is the natural continuity of every physical or natural phenomenon. In other words, the graphical representation of a measured physical data should present a certain continuity and regularity (i.e. the curve shape respects some well-known graphical models). Again, taking discrete samples from that measured physical data, will lead to a discrete curve from which it should be possible to interpolate and fit with usual interpolation or fitting functions.

Hence, it could be seen why Michael Whiteman stated (Whiteman, 1967): "To speak of running through 'all the points' of a curve would therefore be inadmissible. Nevertheless, there are exact concepts of continuity and gradient, which are applicable to conceptually defined curves, and thence are applicable approximately to physical. curves. Thus, in any particular case a physical curve may be tested for continuity, and an approximate measure of its gradient may be found. Likewise a physical trajectory does not consist of isolated events. Nevertheless, by selecting points, the exact concept of velocity may be applied so as to obtain an approximate measure of velocity at any physical point of the trajectory" (Whiteman, 1967).

Now, by considering some natural/physical curves present in the universe, we observe continuity in the graphical representation of the measured dimension (continuity of the dimension). Moreover, we should even find some continuity in its first derivative (continuity of the variation). For instance, let us observe the temperature curve of the weather. Not only, the temperature should vary continuously and smoothly, but also its derivative does so, and this with an extreme respect of the physical and natural well-known laws.

Again, by observing the natural curves present in real life, one remark that the form of the curves is quite identifiable by a simple visual observation (as experts do), and the curves are easily fitted by usual mathematical functions (e.g. Gaussian, Linear, Polynomial, Sinusoidal, Exponential, etc.).

For concreteness, if we take the text data for instance, we may remark that it is composed of several characters (i.e. A, B, C, … Z, for the English), several words (i.e. vocabulary), several numbers (e.g. 1, 2, 3 etc.) and so on.

That is, if a large amount of textual data is analyzed by computing the frequency of some features, we should usually retrieve a Gaussian distribution of the data, when the features are represented from the most frequent to the least frequent.

As it can be seen in the next section, this particularity has been checked with 7 different books by taking the following feature: "Word Length Frequency".

Contrariwise, for the holy Quran those mathematical rules do not seem to be respected, without any possible interpretation. Moreover, by analyzing another feature (i.e. Number citation frequency), we strangely noticed that this last feature does not respect the previous laws either, while for the case of the Hadith book, the mathematical laws are well respected for all those features.

14.2 Fitting and interpolation definitions

Given a set of data that results from an experiment or from a physical scenario, we show (in Mathematics) that there is some function that passes through the data points and optimally represents the area of interest at all present and absent points.

With interpolation (Milne, 2012), we look for a function that allows us to approximate the values between the original data points [William Edmund Milne]. The interpolating function should pass exactly through the original data points. In our experiments, we chose two different techniques: the Piecewise cubic Hermite interpolation (PCHIP), which preserves the monotonicity and shape of the data and the Bezier interpolation technique, which preserves the curve shape in a graphical way.

On the other hand, with curve fitting (Milne, 2012), we simply want a function that represents a good fit to the original data points with a minimum estimation error. With curve fitting the approximating function does not have to pass through the original data points. It should respect the overall data with the best possible fitting, and respect the chosen mathematical expression type as well.

For instance, in table 14.1, we have represented four columns : the first one represents the sample order, the second one represents the original data corresponding the real samples values, the third column represents the interpolation based values and the fourth one represents the fitting based values.

Table 14.1. Example of Interpolation and Fitting.

Sample	Original Samples	Interpolation	Fitting
1	1.00	1.00	1.07
2	absent	2.01	2.04
3	3.00	3.00	2.99
4	absent	3.97	3.98
5	5.00	5.00	5.03

14.3 Investigation on the Word-Length frequency

14.3.1. Definition of the Word Length Frequency (WLF)

Since the first part of our experiment concerns the Word Length Frequency (WLF) (H. Sayoud, 2012), we think that it could be useful to define some technical terms employed in this work:

-The "Word Length" represents the number of letters composing the 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.

14.3.2. Graphical representation of the Word Length Frequency

In this section we will graphically represent the WLF of the two books : holy Quran and Hadith. Furthermore, we will represent the WLF of 6 other books written by 6 different authors to make a general comparison between their features.

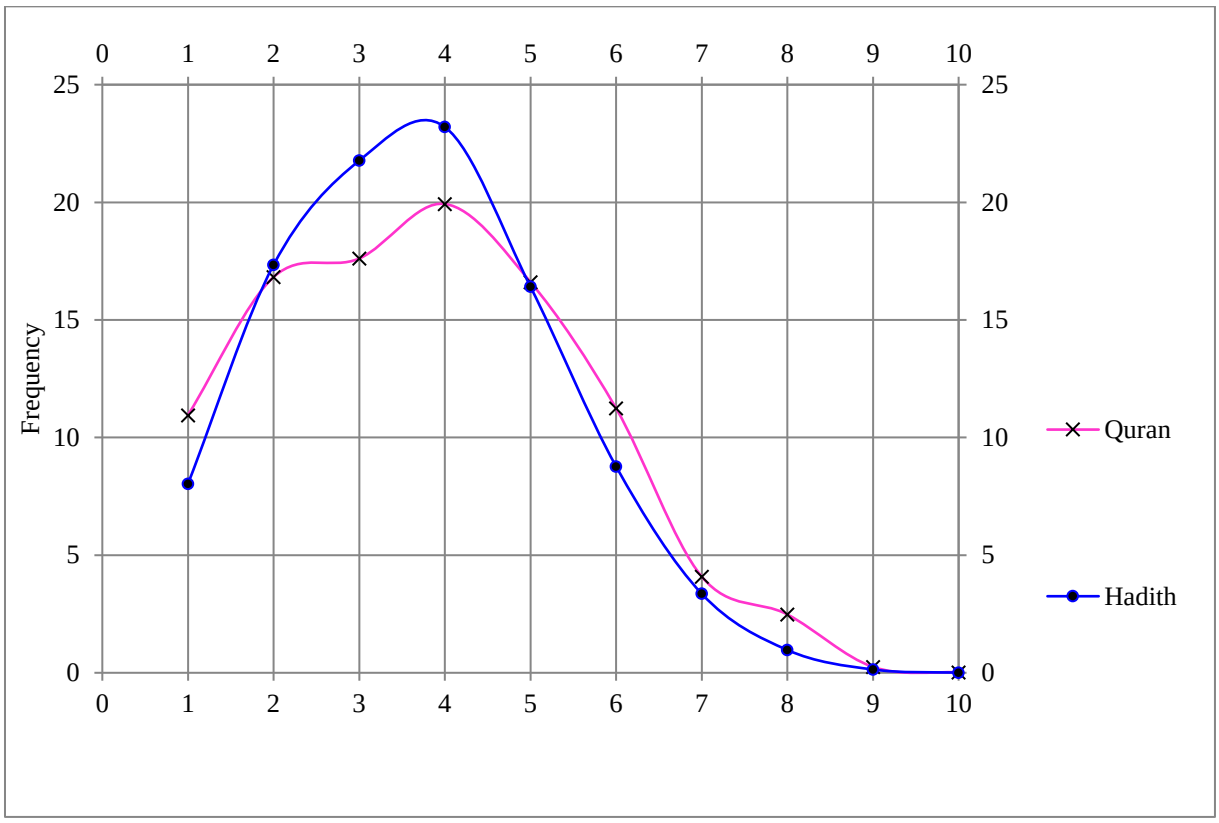


Figure 14.1: Word length frequency of the Quran (in red) and Hadith (in blue), obtained with Bezier interpolation. We can notice that only the Hadith curve presents a Gaussian (or log-normal) shape.

As we can see from figure 14.1, in a visual point of view, the Hadith curve (in blue) respect the Guassianity, whereas the Quran curve (in red) does not seem to respect any Guassianity or at least any semi-Gaussianity (i.e. Gaussianity in one side).
By the term Gaussianity, we mean a smooth bell curve, which more or less resemble to a mono-Gaussian form in one or both sides.

We also visually notice that the ideal curve that could, maybe, ensure a certain Guassianity for the Quran is disconnected from the real Quran curve at abscises 3 and 7.
Due to that strange observation, we have decided to test if the Guassianity is also respected with other long Arabic texts (eg. testing other books/authors) or not. So, we have basically drawn the WLFs of 6 other books written by 6 different authors to see if there is any possible Gussianity or at least a log-normal shape. Hence, several experiments of Word length frequency have been conducted on the holy Quran and the books of 7 other religious Authors, namely: the Prophet (PBUH), Dr Abd AlKafy, Dr Amro Khaled, Dr Al Ghazali, Dr Al Arifi, Dr Al Qarqdqwi and Dr Hassan. The 6 last ones represent contemporary authors from the 20th and 21th century, for which the main topic is also religion. Results are represented in figure 2.

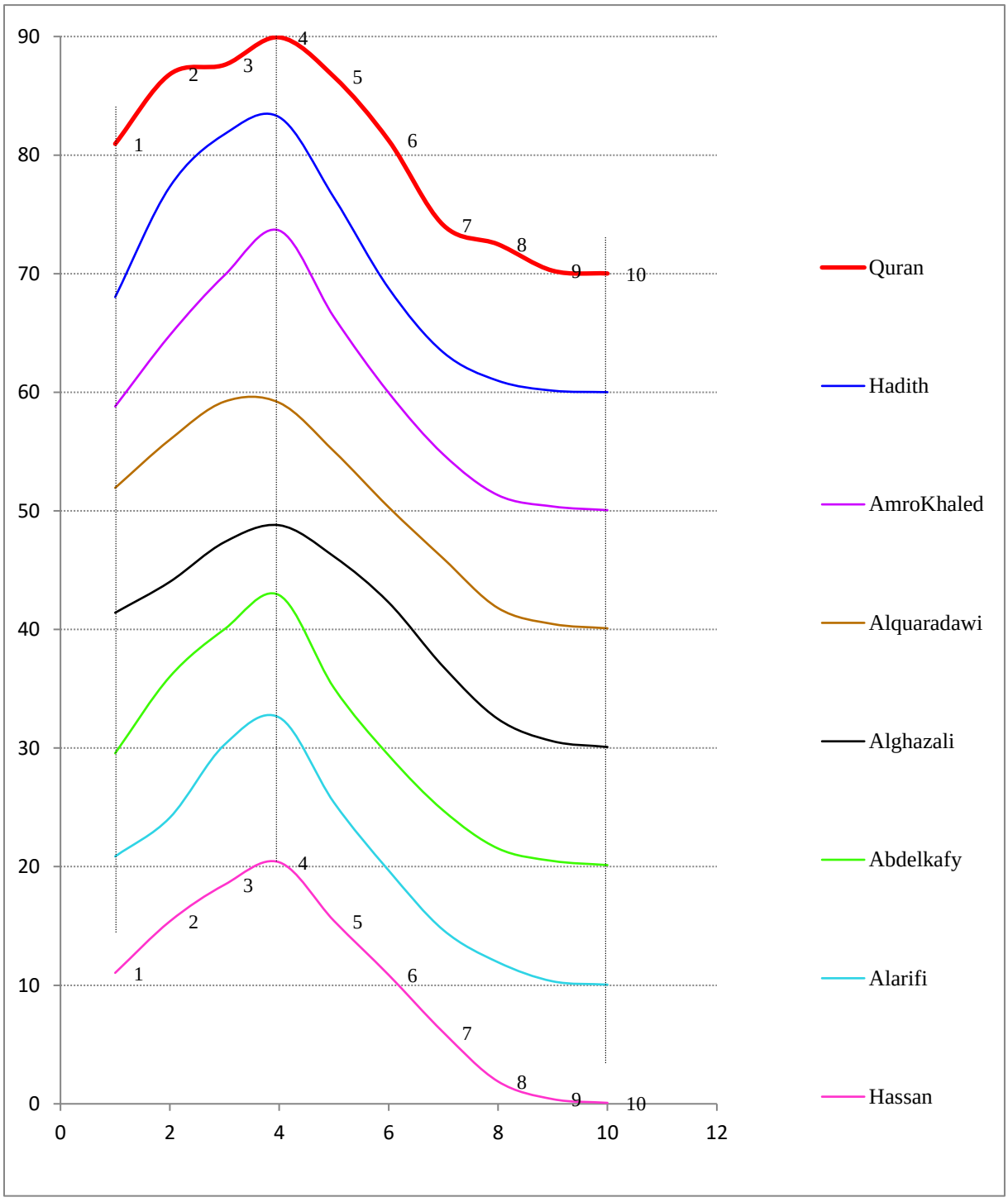


Figure 14.2: The word length frequency of 8 different books: the Quran (in red), Hadith (in blue) and 6 other books written by 6 different contemporary authors, obtained with Bezier smoothing. We notice that all the curves, except the Quran, present a Gaussian or log-normal shape, at least for one of the two sides.

By observing figure 14.2, we notice that all the seven WLF curves present a certain gaussianity (for at least one side), except the Quran one, which has a strange graphical shape that does not seem to respect any form of gaussianity.

In a separate representation, we can see the Quran WLF with more details (see figure 14.3), where it clearly appears that it does not respect any Gaussianity or Interpolability. Moreover, we notice that the general form of the curve is quite strange (not familiar).

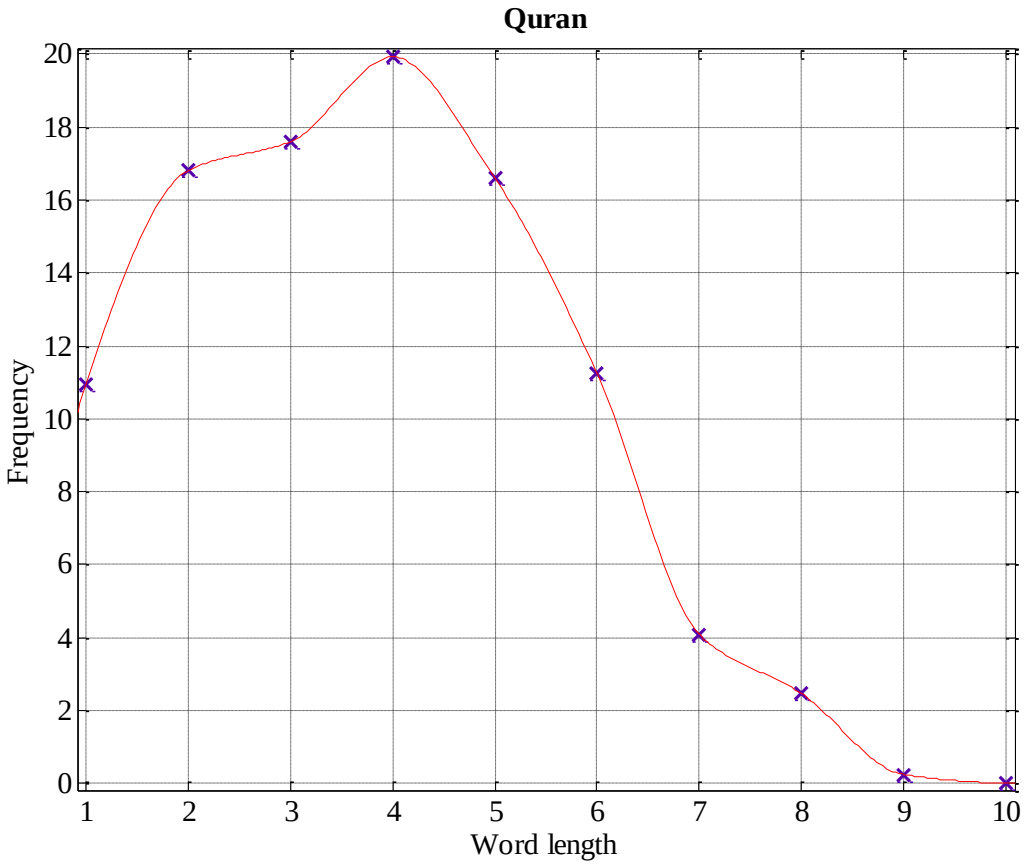


Figure 14.3: The word length frequency of the holy Quran, obtained with PCHIP interpolation: we notice that the curve shape is not Gaussian and not log-normal either, in neither the left nor the right side. We also remark that it is not interpolable with conventional interpolation functions. In fact, two exceptions, precisely in 3 and 7, make the Gaussianity and Interpolability not respected.

In the previous figures (14.1 and 14.3), a strange form is noticed for the Quran WLF and a question could be asked then: is it the result of a mixture of two (or more) styles? In other terms, does the holy Quran result from a mixture of several authors? That question could be statistically stated as: is the Quran representation multi-gaussian?

To answer that question, we simulated several text mixtures and computed the corresponding approximated WLFs.

In the first experiment, we simulated the text mixture of the two authors: Dr Hassan and Dr Amro-Khaled. See figure 14.4.a.

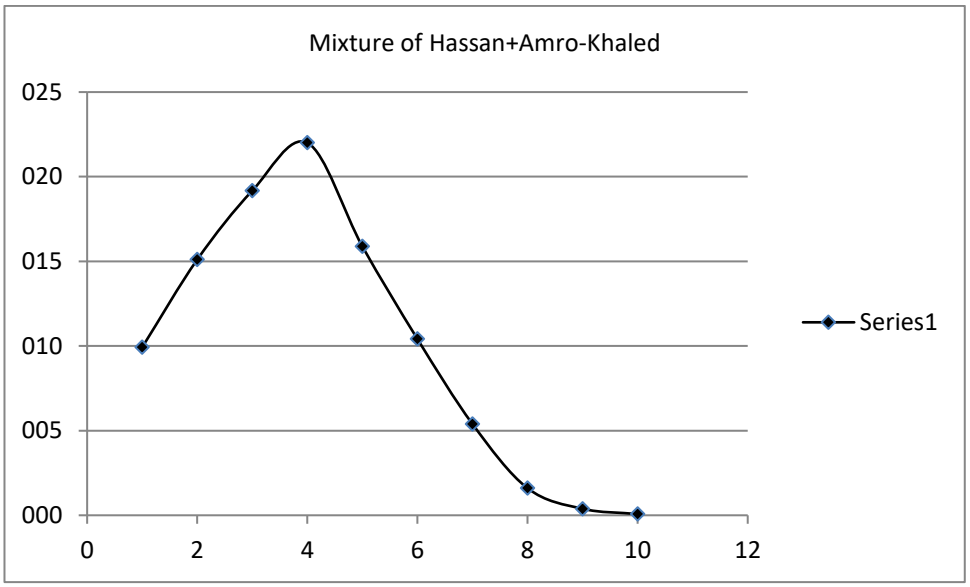


Figure 14.4.a: WLF of a simulated mixture between the texts of two different authors: Dr Hassan and Dr Amro-Khaled.

In the second experiment, we simulated the text mixture of the two authors: Dr Abd-AlKafy and Dr Amro-Khaled. See figure 14.4.b.

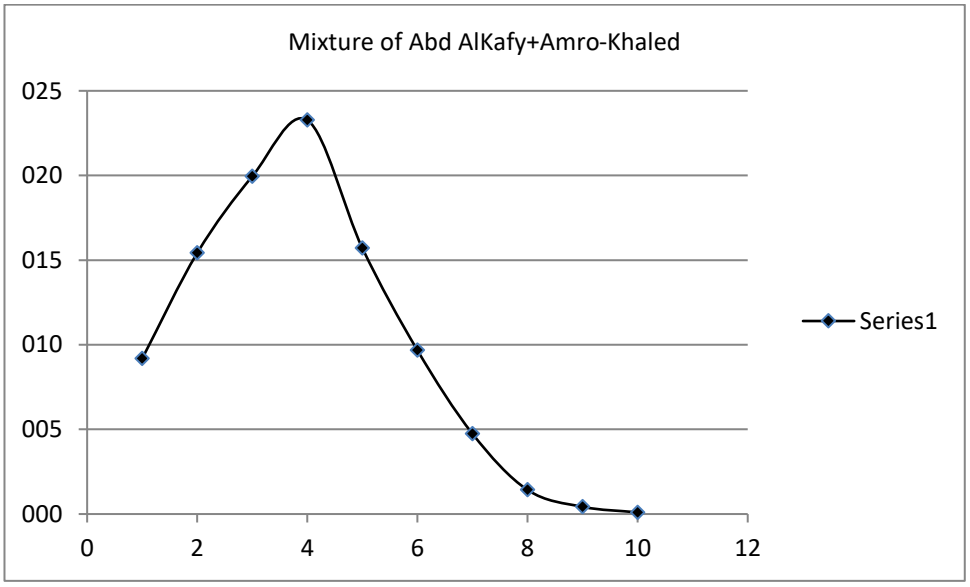


Figure 14.4.b: WLF of a simulated mixture between the texts of two different authors: Dr Abd-AlKafy and Dr Amro-Khaled.

In the third experiment, we simulated the text mixture of the two authors: Dr Al-Ghazali and Dr Al-Qaradawi. See figure 14.4.c.

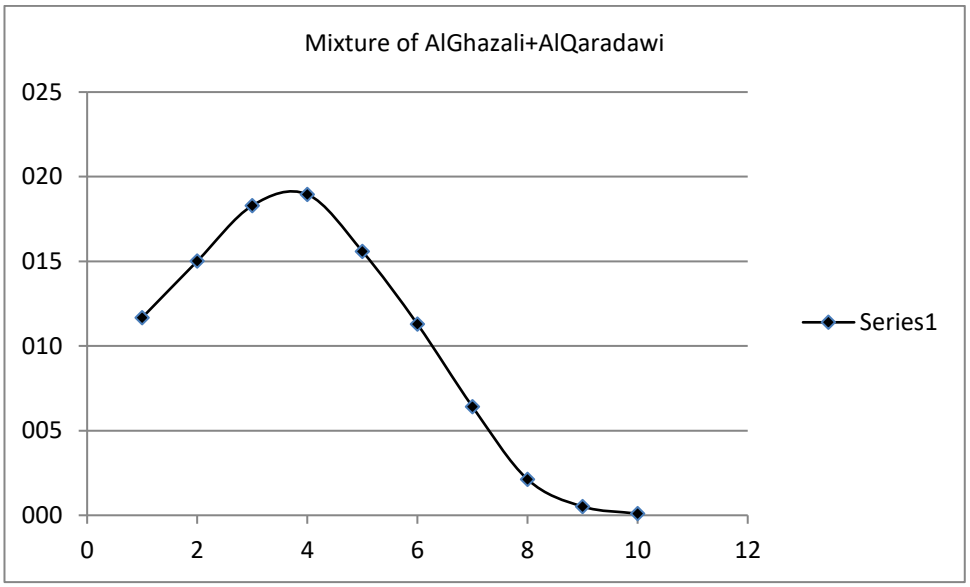


Figure 14.4.c: WLF of a simulated mixture between the texts of two different authors: Dr Al-Ghazali and Dr Al-Qaradawi

In the fourth experiment, we simulated the text mixture of the two authors: Dr Al-Arifi and the Prophet (Pbuh). See figure 14.4.d.

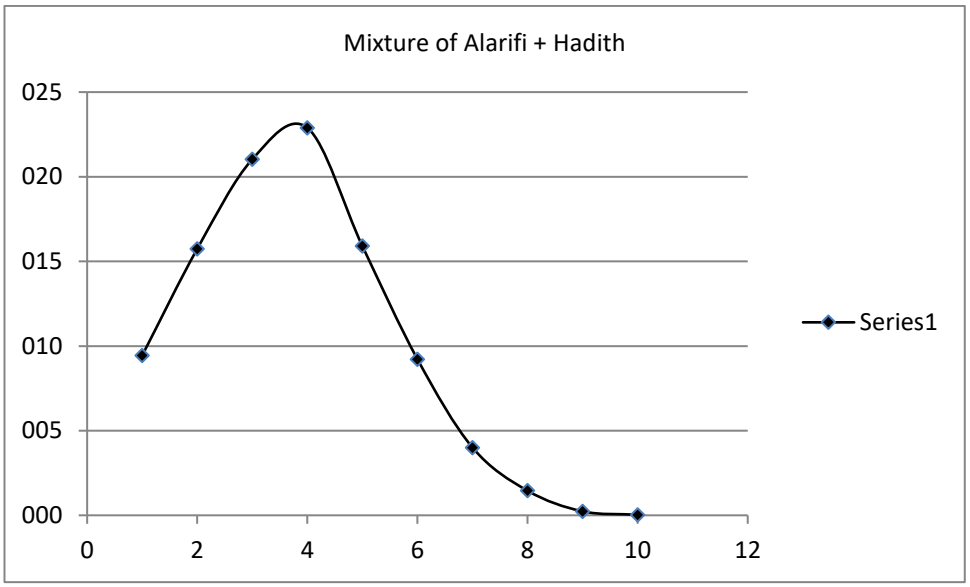


Figure 14.4.d: WLF of a simulated mixture between the texts of two different authors: Dr Al-Arifi and the Prophet (Pbuh).

In the fifth experiment, we simulated the text mixture of 7 different authors: Dr Hassan, Dr Alarifi, Dr Alkarny, Dr Abdelkafy, Dr Alghazali, Dr Alquaradawi and Dr AmroKhaled. See figure 14.4.e.

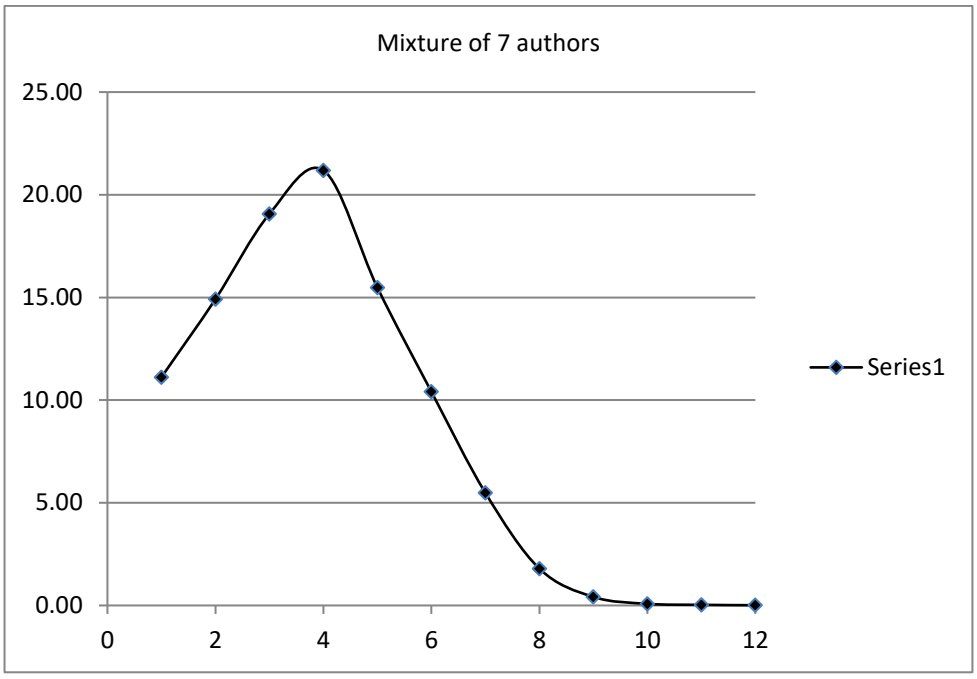


Figure 14.4.e: WLF of a simulated mixture between the texts of 7 different authors: Dr Hassan, Dr Alarifi, Dr Alkarny, Dr Abdelkafy, Dr Alghazali, Dr Alquaradawi and Dr Amro-Khaled.

As we can see in all the simulated texts/authors mixture, we got always a Gaussian (on at least one side), and we do not observe any division of the curve into two or multiple Gaussian as one could expect.

This fact shows that the hypothesis of multiple styles in the holy Quran is excluded, since that phenomenon has not been noticed in the previous simulations.

So, what could be then the reason of that strange unexplained form?

In the opinion of the author, there is no classic explanation for that fact, except that the Quran should be the work of the Creator who made his holy scripture above the classic rules of statistics and mathematics.

14.3.3. Hadith model interpolated with Gaussian fitting

From the previous results, showing that the Hadith should respect a certain Gaussianity and Interpolability, we performed a computation of a Gaussian curve in a form given by equation 14.1, and optimized it to get the lowest error possible (i.e. optimal coefficients for the best fitting).

$$f(x) = a1*\exp(-((x-b1)/c1)^2) \qquad (14.1)$$

The obtained results are given below:

Parameters

a1 = 23.61
b1 = 3.512
c1 = 2.497

Goodness of fit:

SSE: 2.263
R-square: 0.9969
RMSE: 0.5686

The resulting fitted curve is represented in figure 14.5.

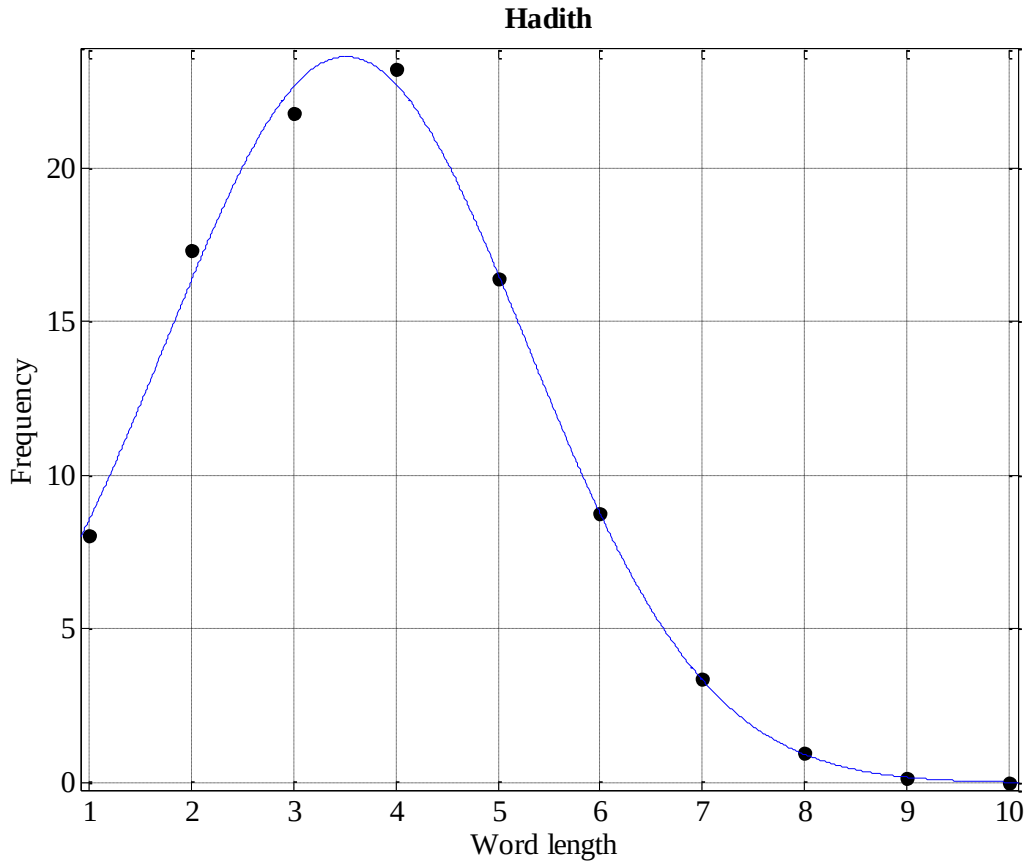


Figure 14.5: The word length frequency of the Hadith , obtained with a Gaussian fitting: we notice that the Gaussian curve fits the data with a quite good precision.

Observation

From the previous results, we notice that the Hadith and all the cited books obey to the law of Gaussianity and Interpolability, except the holy Quran, which does not respect any of these classical laws. That is, we do not understand why this particular exception is noticed for the holy scripture. Again, the only interpretation, one can give, is that the concerned book should have a mysterious origin.

14.4 Investigation on Numbers citation frequency

In this second investigation, we consider the citation of numbers in the text, such as one, two, three, etc. However, those numbers are sorted from the most frequent to the least one. For concreteness, if the numbers 1, 2 and 3 have the following frequencies 10%, 15% and 12% respectively, then they will be sorted into the following sequence: 2 (*1st number*), 3 (*second number*) and 1 (*3rd number*). That scheme makes the representation curve monotone (*decreasing*) and easier to interpolate.

On the other hand, only numbers that are cited at least more than 5 times are considered, for a purpose of consistency. Consequently and in practice, only the 6 or 7 most frequent numbers are kept in the graphical representation.

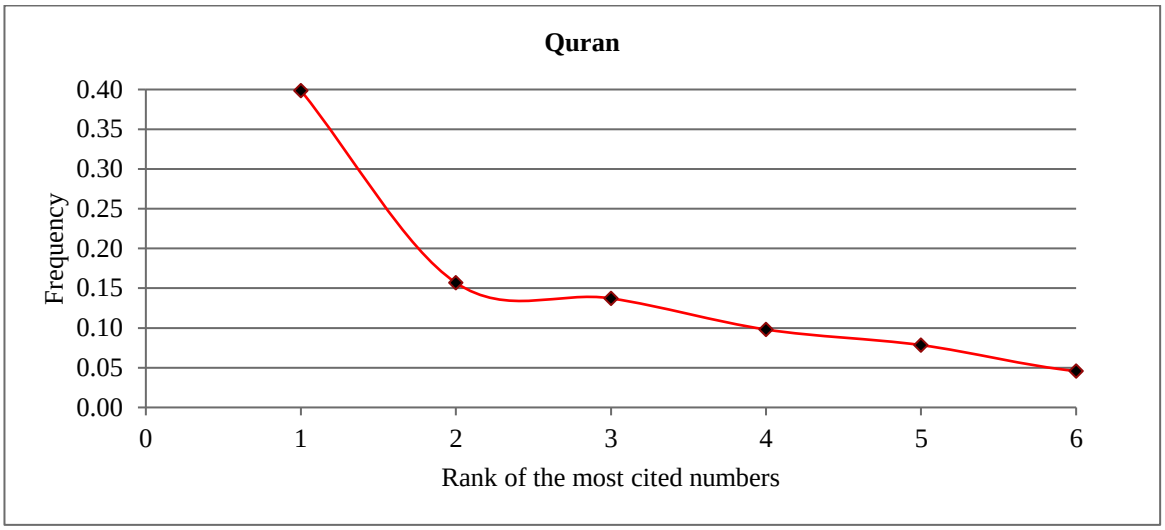


Figure 14.6: Number citation in the holy Quran (*sorted from the most frequent to the least frequent*). The curve is obtained by Bezier interpolation.

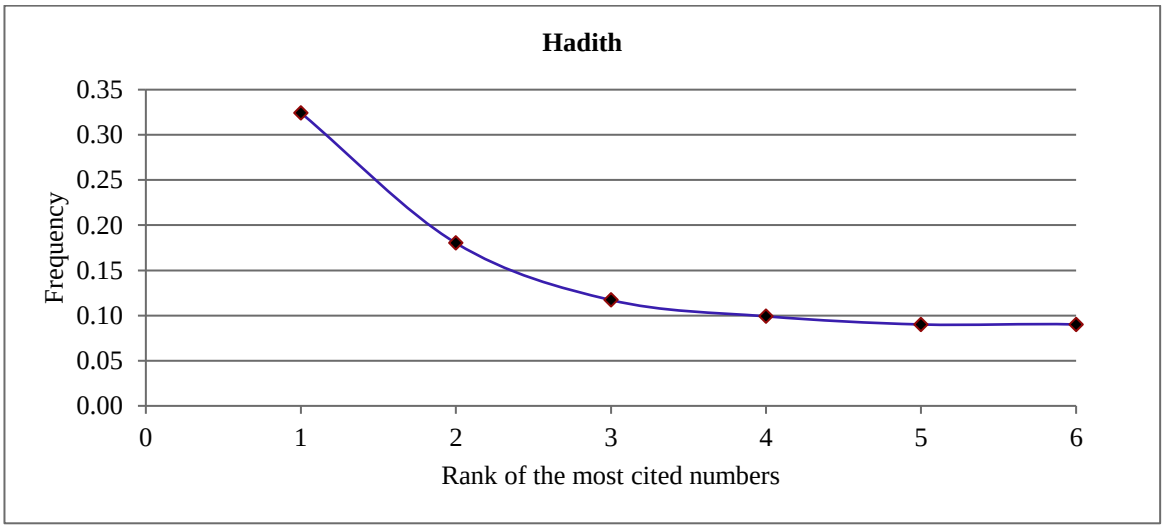


Figure 14.7: Number citation in the Hadith (*sorted from the most frequent to the least frequent*). The curve is obtained by Bezier interpolation.

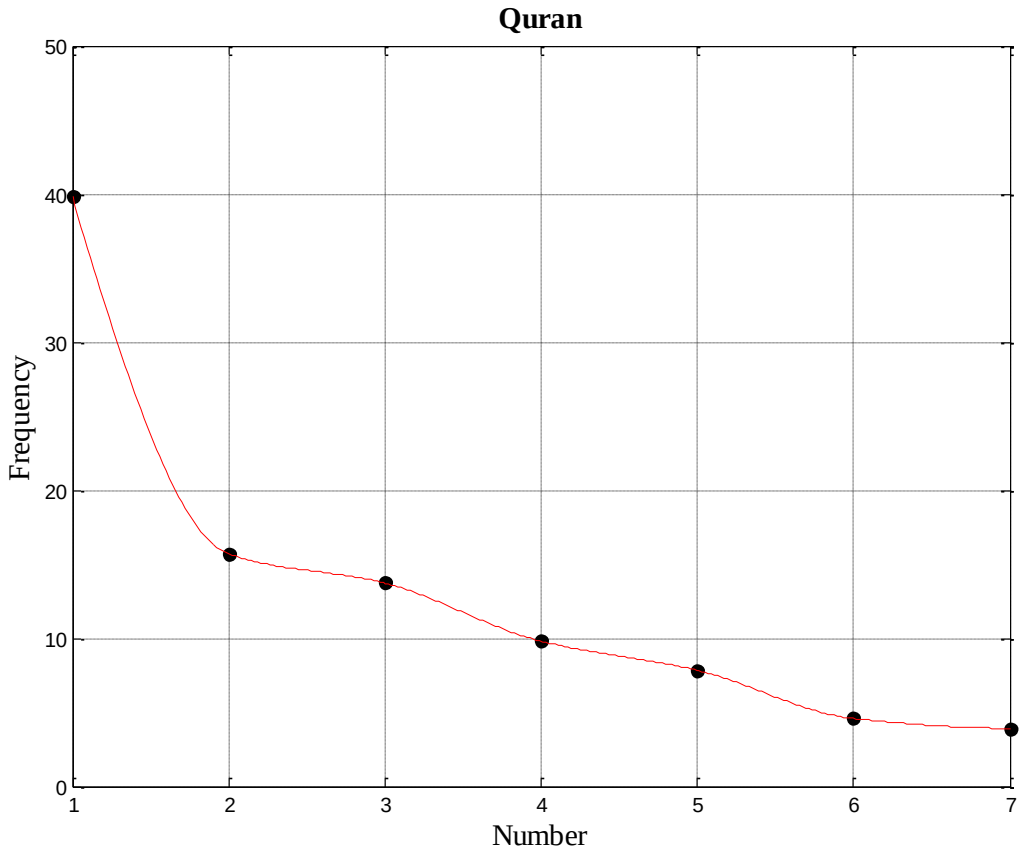


Figure 14.8: Quran number frequency: there is no Gaussianity and no Interpolability either. Once again, the curve doesn't resemble to any known mathematical or physical shape and the curve shape is strange, mathematically speaking. The curve is obtained by PCHIP interpolation.

By observing figure 14.8, we remark that the curve presents different slopes at each segment located between two successive points. The general curve is quite strange and unfamiliar, mathematically speaking. Once again, there is no Gaussianity and no Interpolability either. Furthermore, the curve does not resemble to any known mathematical or physical shape. Also, we notice that there exists a pseudo-horizontal segment between the 2nd and 3rd numbers and between the 6th and 7th ones.

Hadith model interpolated with Exponential fitting f(x) for the Number frequency

As in the previous investigation, and due to the fact that the Hadith curve appears to respect a certain Gaussianity and Interpolability (*i.e. visually*), we performed a computation of an exponential curve in a form given by equation 14.2, and optimized it to get the lowest error possible.

$$f(x) = a*\exp(b*x) + c*\exp(d*x) \quad (14.2)$$

The obtained results are given below:

Parameters

a = 61.85

b = -0.9579
c = 8.81
d = -0.007993

Goodness of fit:

SSE: 0.472
R-square: 0.999
RMSE: 0.3967

The resulting fitted curve is represented in figure 14.9.

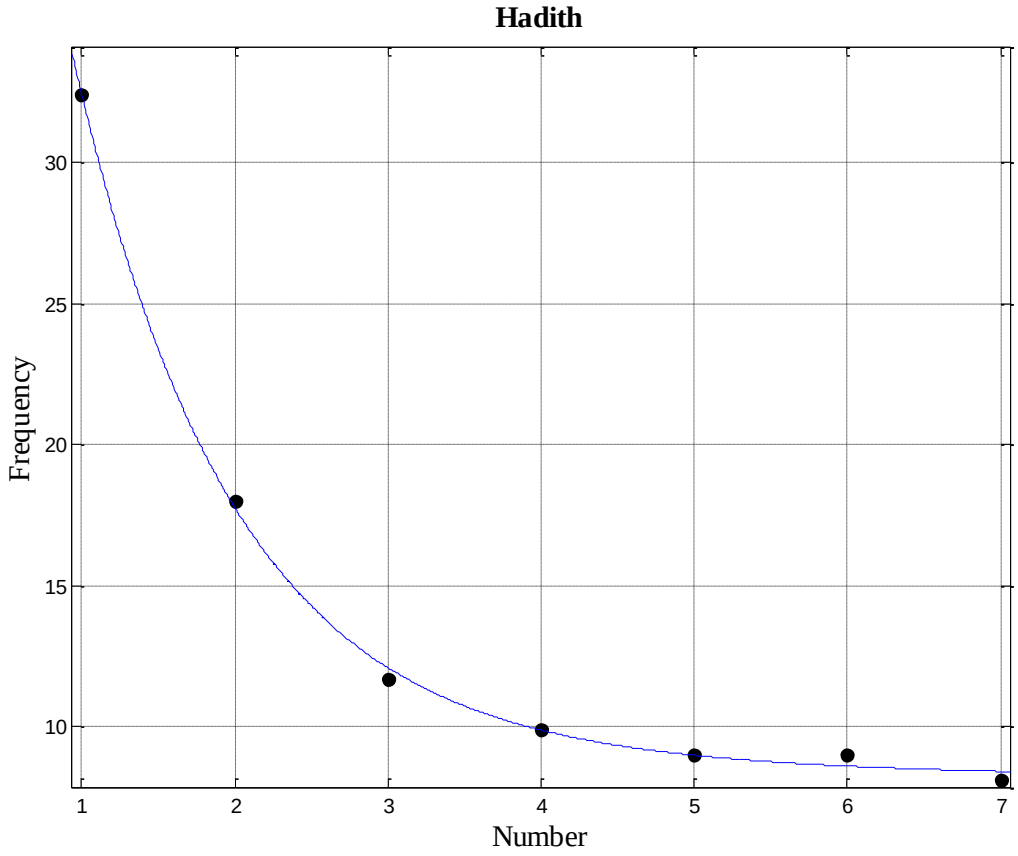


Figure 14.9: Hadith number frequency: Once again, the overall curve seems to follow a partial Gaussianity shape and the Interpolability is possible for every point, by an exponential polynomial, as we can see in the corresponding fitting equation. The curve is obtained with exponential fitting.

Observation

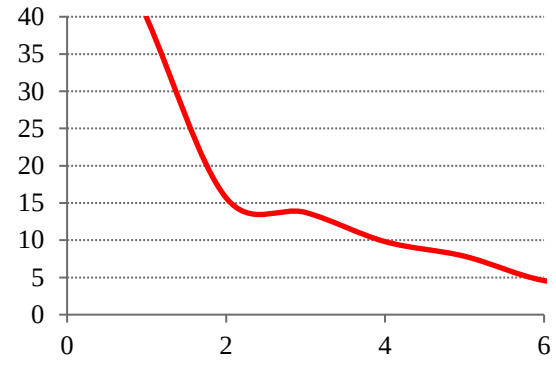
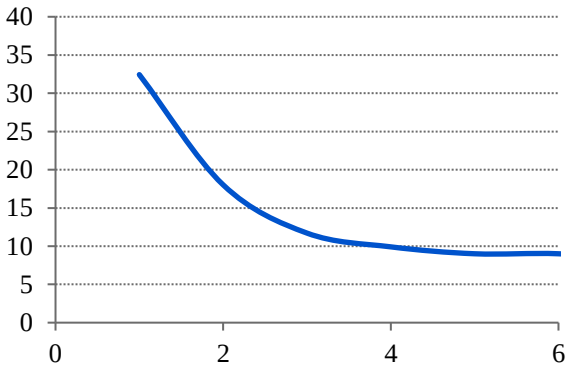
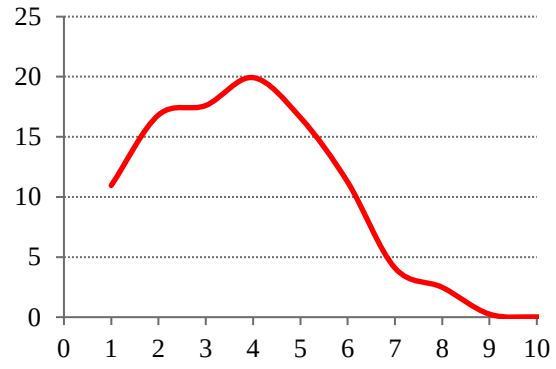
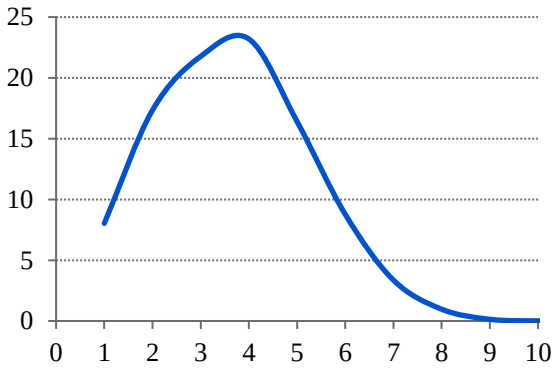
As in the previous investigation, we notice again that the Hadith obeys to the law of Guassianity and Interpolability by presenting a nice exponential shape for the sorted number citation. Contrariwise, the Quran's number citation presents a complex curve with no Guassianity or Interpolability either. This fact, once again, proposes that the holy scripture should have a mysterious origin.

14.5 Conclusion and Discussion

An investigation of Gaussianity and Interpolability has been conducted on the Holy Quran, in order to see if it respects, as the other human books, the physical properties of Gaussianity and Interpolability.

A summary of the different curves is given in the following table (table 14.2).

Table 14.2. Comparison between the Quran and Hadith curves (Bezier interpolation is used in both books).

	Holy Quran	Hadith
Sorted Number citation frequency		
Word length frequency		

As a comparison with other books, we positively verified that the Hadith book does obey to the Gaussianity and Interpolability rules for both Word length frequency and Number citation. Similarly, 6 other books written by different human authors have been analyzed and experimented in the same manner. Once again, those 6 different books appear to obey perfectly to the Gaussianity and Interpolability rules for the word length frequency.

On the other hand, and contrariwise, we strangely noticed that the holy Quran does not respect those rules for the word length frequency and for the number citation frequency either.

Theoretically, Gaussianity is a rule to which obey every physical phenomenon respecting the “Large Numbers” condition (*at least to the knowledge of the author*). However in the case of the holy Quran, neither the Gaussianity nor the continuity of the curve evolution (second derivative) is respected. This fact proposes that the holy Quran could not be a human invention but probably the work of a Superior Non-Human Intelligence who is beyond the prescribed rules and who does not respect any of the well known physical properties. Consequently, we may deduce two important facts:

- Firstly, the two investigated books: Quran and Hadith are quite different in terms of textual structure statistics, which leads to the conclusion that the two corresponding Authors should be different;
- Secondly and more strangely, we do not see any possible human origin for the holy Scripture. Hence, the hypothesis of a Divine origin, for the holy Quran, is widely supported by the result of this investigation.

Finally, the present paper is a pure Statistical/ Computational-Linguistic investigation regardless of the religious aspect of the studied books. It tries only to bring a new scientific discovery, which was hidden and unknown before, to the scientific community.

15 Ninth Series of Experiments: A Mysterious Numerical Structure in the Quran making it different from other Human Books

In this investigation, a mixed linguistic-statistical-numerical analysis is performed on the text of the holy Quran in order to look for any possible presence of hidden numerical structure. In our case, we focused our interest on number seven, which seems to have a long historical and religious presence in the holy book. This task could be seen as a pure statistical analysis using some simplified linguistic rules and numerical properties to check the possibility of a Divine origin (i.e. beyond the human capacities) of the book.

The dataset consists of the 2 forms of the Quran: once with diacritics and another time without diacritics. The first text is considered as our reference book since it consists in the authentic Quran that is preserved by the Muslim religious community and respecting the Saudi Quran edition (narrated by Hafs). The second one is similar to the first book, except that we have removed the diacritics to speed up the word search process.

Several simplified experiments of computational linguistics, which are based on the multiplicity by seven, are performed and commented on the Holy Scripture. Results are outstanding and surprising: in fact, the seven-based structure has been discovered in many parts of the holy Quran. This result shows that the possibility of a human origin for such a fascinating structure is quite impossible. In fact, the advanced scientific structure of the Quran antedates the period of the Prophet by many centuries.

15.1 Motivation based on the citation of number 7 in the holy Quran

The history of numerical structure began in the 70s with some discoveries on number 19. Later on, very few scientists were interested in this domain (ie.e “Ejaz Adadi”) (Yakub, 2008) and recently (2002-2008) several numerical evidences based on number 7 were reported by Kaheel (Kaheel, 2015). These last results concerned the manipulation and concatenation of some numbers related to the holy Quran, which in many cases, produces a multiple of 7.

By reading the available documentation on the matter we notice that there exist a sort of hidden numerical organisation inside the Holy Scripture, which could be seen as a watermarking (Cox, 2007). For instance, let us look at the following watermarked English text: « *So I do let him accept a great space!! Yes its* ».

In this example, we do not care about the semantic meaning of the text, it is not our objective, but if we decompose this text into segments of 7 characters each, we will get the following table (table 15.1). Now, again if we keep only the letters located at the 1st diagonal of the table (in red), we will get the following word: « Secret », which was really embedded in the precedent text but in a hidden form.

Table 15.1: Example of text watermarking.

	S	o	-	I	-	d	o
	l	e	t	-	h	i	m
	a	c	c	e	p	t	-
	a	-	g	r	e	a	t
	s	p	a	c	e	!	!
	y	e	s	-	i	t	s
Watermark	*S*	*e*	*c*	*r*	*e*	*t*	

Such techniques of watermarking (Cox, 2007) are relatively recent and have been proposed during the last century only.

Now, if such a watermarking does exist in the holy scripture, it is more likely that this signature should be more sophisticated and much more powerful, since it is supposed to belong to a super-intelligent Author (Allah).

This reason was so motivating that it prompted us to conduct a thorough mixed linguistic-statistical-numerical analysis on the holy Quran and especially on number 7.

15.2 Citation of number 7 in the holy Quran

The reader of the holy book can easily notice that the number seven is very often used and in many circumstances too, as we can see in the following cases:

1. The number of heavens (as described in the Quran 65:12) is 7 [Holy Quran]
2. Number 7 is the first number stated in the Quran (Kaheel, 2015)
3. Number 7 is the most repeated number after number 1 (Kaheel, 2015) and (H. Sayoud, 2012). See figure 15.1.

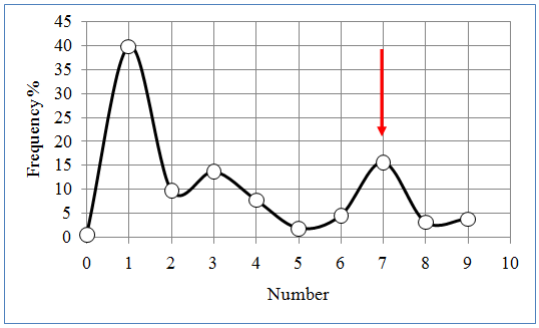


Figure 15.1: Number citation frequency (*Statistics made by H. Sayoud in 2012*) (H. Sayoud, 2012)

4. Al-Fatihah (the opening of the Quran) is composed of 21 alphabets (=3x7): a multiple of 7 (Kaheel, 2015)
5. Al-Fatihah (the opening of the Quran) is composed of 7 verses (Kaheel, 2015)
6. As described in the Quran, the Hell possess 7 doors (ref. 15:44) [Holy Quran]
7. The number of alphabets used as Initials (disconnected letters) is 14 (=2x7): a multiple of 7 (Kaheel, 2015)
8. The numbers of the Quran's alphabets is 28 (= 7×4): a multiple of 7.
9. The Prophet (PBUH) lived for 63 years (=7x9): a multiple of 7 (Kaheel, 2015)
10. During the pilgrimage, Muslims turn around the Kaaba 7 times [Al-Bukhari Hadith 854]
11. Also, during the pilgrimage, Muslims walk between the two hills of Safa and Marwa 7 times [Al-Bukhari Hadith 854]

12. Also, during the pilgrimage, as stated in the Quran, for special circumstances in the pilgrimage, it is required to fast 7 days when returning back [Holy Quran]

13. Once again, during the pilgrimage, the pilgrims throw 7 pebbles at the 'Jamrat Al-Aqabah' [Al-Bukhari Hadith 854]

14. In the verse 12:43, the Quran describes the history of Youssof (Joseph) and the famous dream about the 7 cows [Holy Quran]

15. In the Islamic religion, children are required to begin their prayer from the age of 7 [Al-Bukhari Hadith 854]

16. In the verse 69:7, the Quran describes the history of Punishment of Aad during 7 nights [Holy Quran]

17. In the verse 2:261, the Quran gives an example for the charity recompense: "*The example of those who spend their monies in the cause of Allah is that of a grain that produces seven* (7) *spikes...*" [Holy Quran]

18. In the verse 31:27 , the Quran speaks about the unlimited/numerous words of Allah and cite as example the « 7 seas »" [Holy Quran]

19. The verse 15 :87 speaks about the 7 repeated verses : « *And We have certainly given you, (O Muhammad), seven of the often repeated (verses) and the great Qur'an* » [Holy Quran]

All these examples show that it should exist an enigma about number seven, but what could be the secret? This question prompted us to analyse the holy book and try to see if we could retrieve other statistical facts related to that number.

15.3 New Statistical Evidences based on Number Seven

Trying to respond to the previous question, we made several experiments of statistical word analysis on the Holy Scripture (Certified Saudi Version of the holy Quran, narrated by Hafs). So, basically, we first selected the most important words in term of occurrence frequency (most employed words), and then completed that list by some potential keywords that could have a relationship with any possible numerical evidence, such as the word Adad (meaning "NUMBER") for example.

Results were interesting and sometimes surprising, as we can see in the following discovered cases:

1. One of the most frequent words in the holy Quran is the name of God « ٱللَّه ». This Divine name has 3 morphological forms in the holy Quran: **a)** « ٱللَّهِ » spelled « Allahee », **b)** « ٱللَّهُ» spelled «Allahu» and **c)** « ٱللَّهَ » spelled « Allaha ». We present the occurrences of those 3 words as follows:

1.a The name of God « ٱللَّهِ » spelled « Allahee » has a total occurrence number of 1239=7x3x59 (multiple of 7) if we include all the 112 bismala(s) introducing the different chapters. Again, if we do not count those 112 bismala(s), the new count becomes 1127=7x7x23, which is still multiple of 7 (double multiplicity by 7: 7x7). -S- (*Ref. Statistics made in April 2015*).

1.b The name of God « ٱللَّهُ» spelled «Allahu» has a total occurrence number of 980=7x7x20, which is a multiple of 7 (double multiplicity by 7: 7x7), by either keeping the 112 bismallah or not. -S- (*Ref. Statistics made in April 2015*).

1.c The third name of God « اللّٰهَ » spelled « Allaha » has a total occurrence number of 592=37x16 (multiple of 37). In this case number 37 seems to be particularly interesting since the expression "لا اله الا الله \هو \ انا …" , meaning "*There is no God but Allah*", and which represents the first pillar of the Islam, has also an occurrence number of exactly 37 in the holy Quran. -S- (*Ref. Statistics made in April 2015*).

2. Again, concerning the words that are very frequently used, we strangely noticed that the total occurrence number of the coordination conjunction « أَو », meaning « OR », in the holy Quran, is 280=7x40 (*Ref. Statistics made in April 2016*).

3. In the holy Quran, it is clearly stated that there are 7 heavens. Amazingly, the term « Heavens » « السَّمَاوَات /سَمَوَات » is cited 189 times =7*27=3*3*3*7 (*Ref. Statistics made in April 2016*).

4. We strangely noticed that the total occurrence number of the word/verb « عدد », meaning « Number », in the holy Quran, is exactly 7. This word is extremely interesting since it is the most important keyword in such numerical analysis. In fact, this was one of my first discoveries making me so surprised that I pushed farther my analysis to other potential keywords in the holy book. So, how could be this structure based on number 7 a pure hazard while the token "NUMBER" itself has been repeated exactly 7 times in the Quran? -S- (*Ref. Statistics made in April 2015*).

5. We strangely noticed that the total occurrence number of the word « كلمات», meaning « Words», in the holy Quran, is exactly 14 =2x7 (multiple of 7). This word is also particularly interesting since it is one of the most important keywords in our investigation. In fact, it seems like an enigma where the central point is focused on the repetition of words, and we strangely notice that the token "WORDS" itself is repeated by a number that is multiple of 7. -S- (*Ref. Statistics made in April 2015*).

6. We strangely noticed that the total occurrence number of the verb « أَحْصَى», meaning « To Count/ Make Statistics », in the holy Quran, is exactly 7. As previously, this word is semantically pertinent and represents a potential keyword in statistical analysis. Do we understand that the verb "MAKE STATISTICS" itself has been repeated exactly 7 times? -S- (*Ref. Statistics made in April 2015*).

7. We strangely noticed that the total occurrence number of the word « القُرْأَن», meaning « The Quran. », in the holy Quran, is exactly 49= 7x7 (double multiplicity by 7). This word represents the central keyword of our investigation, since it represents the dataset (the book) where all the numerical structures are discovered. Moreover, the repetition of "THE QURAN" is not only a multiple of 7, but even more: it is exactly equal to 7x7, namely a strong relationship with number 7. Could this be really a coincidence? On the other hand, we can note that the occurrence of "قُرْأَن" without the prefixed article "the" is 69 =3x23. – Furthermore, note that the Quran had been sent down for a period of 23 years: a strange coincidence again. -S- (*Ref. Statistics made in April 2015*).

8. We strangely noticed that the total occurrence number of the expression «تلك ايات الكتاب » , meaning « Those are the signs/evidences of the book », which make reference to the mysterious disconnected letters (sort of acronyms) [Sayoud 2013] in the holy Quran, is exactly 7. As stated in its translation (those are the signs/evidences of the book), it is likely that this expression involves a hidden secret or proof in the holy Quran with regards to the disconnected letters. In fact, why is this expression put just after the mysterious disconnected letters? And what sort of sign or evidence should it provide? And finally, why is it repeated seven times? Sincerely, one cannot say that this is a pure hazard. -S- (*Ref. Statistics made in April 2015*).

9. We strangely noticed that the total occurrence number of the *triplet-of-characters* « سبع», representing "the connected successive letters composing number 7 in Arabic (i.e. the word *seven* without diacritics), in the holy Quran", is exactly 28=7x4. It is also equal to the number of Arabic alphabets. Once again this specific keyword, representing the root of the word seven, seems to respect the 7-based structure with a repetition of 28 times, which is not only a multiple of 7, but also equal to the number of the Arabic alphabets composing the holy book. -S- (*Ref. Statistics made in April 2015*).

10. We strangely noticed that the total occurrence number of number seven «سبعا/ سبع/ سبعة », representing the number 7 used as Noun (not as adjective), such as: "*All the Seven are present"*, is exactly 7. Herein, number seven is used seven times, but why only when it is employed as noun? Maybe it is an indication on its profound entity/symbol and not in its function. -S- (*Ref. Statistics made in April 2015*).

11. We strangely noticed that the total occurrence number of number seven « سبع » used without prefix and without suffix, is exactly 14= 7x2. This is the simplest way to estimate the seven's occurrence number by any computer, without looking for any additional prefix or suffix. -S- (*Ref. Statistics made in April 2015*).

12. We strangely noticed that the total occurrence number of the word « غيب», meaning « Hidden-secrets/ the Unseen/ Knowledge-on-the-future», in the holy Quran, is exactly 49 =7x7 (double multiplicity by 7). This word, which refers to unseen things or hidden secrets, is closely related to the subject of our main investigation and should represent a serious keyword to investigate. -S- (*Ref. Statistics made in April 2015*).

13. We strangely noticed that the total occurrence number of the word « العلم», meaning « The Science », in the holy Quran, is exactly 28 (=4x7), a multiple of 7. Again, if we count the occurrence of this word with or without the article "the", we will find exactly 105 (=3x5x7), which is still a multiple of 7. It is clear that all fields of numerical analysis or Mathematics represent a small field of the wide domain called "Sciences". So, we do expect to retrieve a great secret in this particular significant word and probably much more scientific evidences in the holy book. -S- (*Ref. Statistics made in April 2015*).

14. We strangely noticed that the total occurrence number of the adjective/name (of God) « عليم», meaning « Knowing everything (God)», in the holy Quran, is exactly 161 (=23x7) -S- Note also that the Quran had been sent down for a period of 23 years (*Ref. Statistics made in April 2015*).

15. We strangely noticed that the total occurrence number of the composed word « أولو الألباب », meaning « Those who possess intellect/ who reflect (*on the origin of creation/ on the matters of the Quran*)», in the holy book, is exactly 7. It should probably exist a subtle relationship between this composed word and the mathematical structure, which we are looking for: hence, the seven repetitions of this composed word could not be in vain. -S- (*Ref. Statistics made by H. Sayoud in April 2015*).

16. We strangely noticed that the total occurrence number of the expression/adjective « لا ريب فيه», meaning « There is no doubt in it (Quran/ Judgement day)», in the holy Quran, is exactly 14 =2x7 (multiple of 7). In the opinion of the author, this expression is one of the most important keywords that could possess a specific mathematical code in the holy Quran, since it clearly states that "THERE-IS-NO-DOUBT-IN-IT". In fact, we easily retrieve that the total occurrence number of that expression is a multiple of 7, confirming the previous opinion. Moreover, we strangely noticed that the number of alphabetical characters used in this expression is exactly 7 alphabets (ب ي ا ر ل

ه ف), which represents another surprising fact. Again, in the opinion of the author, this expression has a profound significance and semantically states that there is no doubt in the Divine origin of the holy scripture. On the other hand, it statistically shows an extreme intelligence behind those words. -S- (*Ref. Statistics made in April 2015*).

17. We strangely noticed that the total occurrence number of the word « أَيَة» without suffixes, meaning "Verse/ Sign/ Proof", is exactly 84 =7x12 (multiple of 7). On the other hand, by looking for any possible link between the occurrences of this word, meaning "VERSE" and the number of verses in the holy book, we can easily discover that the total number of verses in the holy Quran, by keeping the bismallas (i.e. 6236+112), is equal to 6348 =12x23x23. Hence, it strangely appears that number 12 is a common factor between both equalities, but we do not know what is the real reason for that coincidence. We also notice, in the number of verses equation (12x23x23), that 23 could represent the Quran revelation period, which is really 23 years indeed, and then by multiplying it to 12 (e.g. twelve months), we also get the total number of months concerning the revelation period, namely: 12x23 months. Sincerely, we cannot argue that this is a pure hazard.-S- (*Ref. Statistics made in April 2015*).

18. We strangely noticed that the total occurrence number of the word « تَنْزِيل» when related to the holy book, meaning « Revelation (of the holy book)», is 14=7x2. Note that it is employed 14 times in relation with the holy book (meaning: Revelation), and once (another time) with regards to the Angels with another meaning (i.e. Coming down). Moreover, all the occurrences of this word (with the two meanings) are present in exactly 14 (7x2) different chapters: 17, 20, 25, 26, 32, 36, 39, 40, 41, 45, 46, 56, 69 and chapter 76. Strangely again, the number of concerned chapters is also a multiple of 7. -S- (*Ref. Statistics made in April 2015*).

19. We strangely noticed that the total occurrence number of the word « الفرقان /فرقان », meaning « Quran/ Criterion/ Salvation», in the holy Quran, is exactly 7. This word, which usually refers to the "Quran" meaning, has a great importance in the statistical analysis of the Quran words. And what is found, herein, proposes the probable existence of a strong numerical structure in the holy book. -S- (*Ref. Statistics made in April 2015*).

20. We strangely noticed that the total occurrence number of the word « عَدْل / العَدْل», meaning « Justice/Equality. », in the holy Quran, is exactly 14= 7x2 -S- (*Ref. Statistics made in April 2015*).

21. We strangely noticed that the total occurrence number of the expression/adjective « على كل شيء قدير», meaning « (God) is able to do all things », in the holy Quran, is exactly 35 (=5x7) -S- (*Ref. Statistics made in April 2015*).

22. We strangely noticed that the total occurrence number of the double-word God's Name « الغفور الرحيم », meaning « The Forgiving, the Merciful. », in the holy Quran, is exactly 7. -S- (*Ref. Statistics made in April 2015*).

23. We strangely noticed that the total occurrence number of the double-word adjective « غفور رحيم », meaning « (is) Forgiving, Merciful. », in the holy Quran, is exactly 49 =7x7 (double multiplicity by 7). -S- (*Ref. Statistics made in April 2015*).

24. We strangely noticed that the total occurrence number of the double-word God's Name « رب العالمين », meaning « the Lord of the Worlds », in the holy Quran, is exactly 42 (=7x6). -S- (*Ref. Statistics made in April 2015*).

25. We strangely noticed that the total occurrence number of the double-word God's Name « التواب الرحيم », meaning « The Receiver of repentance, the Merciful. », in the holy Quran, is exactly 7. -S- (*Ref. Statistics made in April 2015*).

26. We strangely noticed that the total occurrence number of the double-word God's Name « العزيز الحكيم », meaning « the Almighty, the Wise », in the holy Quran, is exactly 42 (=7x6). -S- (*Ref. Statistics made in April 2015*).

27. We strangely noticed that the total occurrence number of the double-word adjective « شديد العقاب », meaning « (Allah is) Severe in punishment », in the holy Quran, is exactly 14 (=7x2). -S- (*Ref. Statistics made in April 2015*).

28. We strangely noticed that the total occurrence number of the double-word « بغير حساب », meaning « (Allah gives of His Bounty) Without limit. », in the holy Quran, is exactly 7. -S- (*Ref. Statistics made in April 2015*).

29. We strangely noticed that the total occurrence number of the double-word related to God: « الفضل العظيم », meaning «(God is possessor of) the Sublime Grace », in the holy Quran, is exactly 7. -S- (*Ref. Statistics made in April 2015*).

30. We strangely noticed that the total occurrence number of the double-word « يرجع الأمر /ترجع الأمور », meaning « All affairs/ Decisions are returned back (to Allah) », in the holy Quran, is exactly 7. -S- (*Ref. Statistics made in April 2015*).

31. We strangely noticed that the total occurrence number of the word «المؤمنين», meaning « The Believers», in the holy Quran, is exactly 105 (=7x15) -S- (*Ref. Statistics made in April 2015*).

32. We strangely noticed that the total occurrence number of the word «مؤمنون/ المؤمنون», meaning « The Believers », in the holy Quran, is exactly 35 (=7x5) -S- (*Ref. Statistics made in April 2015*).

33. We strangely noticed that the total occurrence number of the word «مؤمنة/ مؤمنات», meaning « Female Believer(s) », in the holy Quran, is exactly 28 (=7x4) -S- (*Ref. Statistics made in April 2015*).

34. We strangely noticed that the total occurrence number of the word «رُوح/ الرُوح», meaning « the Spirit », in the holy Quran, is exactly 21 (=7x3) -S- (*Ref. Statistics made in April 2015*).

35. We strangely noticed that the occurrence number of the expression «حَسْبُنا /حَسْبُه /حَسْبِي الله», meaning « God is sufficient for (protecting) us/me/him », in the holy Quran, is exactly 7. -S- (*Ref. Statistics made in April 2015*).

36. We strangely noticed that the total number of the 3rd Arabic pronoun « هُوَ/هِيَ », meaning «She/ He », is respectively 65 (13x5) + 481 (13x37) = 546 =7x78 (multiple of 7 too). -*Note that the pronoun "it" doesn't exist in Arabic (it is replaced by He or She)*-. Another strange relationship between those two pronouns is the number 13 that is present in the 2 equalities. -S- (*Ref. Statistics made in April 2015*).

37. We strangely noticed that the total occurrence number of the double-word God's adjective « وَاسِع عَلِيم », meaning « (Allah is) all-Pervading, all-Knowing », in the holy Quran, is exactly 7. -S- (*Ref. Statistics made in April 2015*).

38. A particular interest is given to the word «الحمد», which means « All Praise (*is due to Allah*) ». This one is the first word in the whole text of the holy Quran after the Bismala and is often used as the

first word in several chapters too, which makes it quite important. Moreover, Muslims repeat it in their prayers (Fatiha chapter) at least 17 times per day and the word is exclusively reserved to God. In terms of statistics, it is cited 28 times: 28=7x4 (multiple of 7). -S- (*Ref. Statistics made in April 2015*).

39. We strangely noticed that the total occurrence number of the word «بُشْرَى », meaning « good tidings/ good news », in the holy Quran, is exactly=14=2x7. -S- (*Ref. Statistics made in April 2015*).

40. We strangely noticed that the total occurrence number of the word « صَفًّا », meaning "rank" , is exactly 7. -S- (*Ref. Statistics made in April 2015*).

41. We strangely noticed that the total occurrence number of the word « سَوْفَ », meaning "will/ shall (i.e. *future auxiliary*)", is exactly 42 =7x6 (multiple of 7). -S- (*Ref. Statistics made in April 2015*).

42. We strangely noticed that the total occurrence number of the verb « افتراه », meaning "(they pretend that) He invented it (the Quran)" , is exactly 7 (*Ref. Statistics made in April 2016*).

43. We strangely noticed that the total occurrence number of the word « الْهُدَى /هُدَى », meaning "Right guidance", is exactly 42 =7x6 (multiple of 7) (*Ref. Statistics made in April 2016*).

44. We strangely noticed that the total occurrence number of the word « ذِكْرى », meaning "Reminder", is exactly 21 =7x3 (multiple of 7) (*Ref. Statistics made in April 2016*).

15.4 Discussion

From this investigation and according to the numerous results showing the presence of a 7-based multiplicity in many parts of the holy Quran, we can state that it should exist a sort of signature or watermarking based on number 7 and maybe other numbers too (eg. 19, 37, …), probably to preserve or prove the authenticity of the holy book. Consequently, the possibility that a human being could have invented such a complex scripture appears to be impossible. Again, the probability that the holy book could be falsified or altered is consequently extremely weak.

Concerning the author style difference between the Quran and Hadith, this new result shows that the Quran style that appears to embed various numerical signatures, is really different from the Hadith style, which does not contain any such signature.

In the verse: « إِنَّا نَحْنُ نَزَّلْنَا الذِّكْرَ وَإِنَّا لَهُ لَحَافِظُونَ» (15:9), it is clearly stated that the Holy Scripture is/will be protected and preserved by His Creator (Ibnu-Kathir, s. d.). So, it is noticeable that the new numerical structure, which has been discovered in this investigation, could not be a hazard at all. The above results make the hypothesis advanced by those who think that the Quran could be a human invention illogical. How could a man (the Prophet), who was illiterate, develop a so sophisticated mathematical structure, for preserving his scripture, fourteen centuries ago?

16 Tenth Series of Experiments: Authorship Attribution based on the Interrogative Form

In this section, we tackle the problem of author discrimination between the two religious books, by segmenting the text documents into several segments. The originality of this research work lies in the use of a new set of linguistic features based on 26 interrogative features and a special fusion. The different experiments are performed by using a statistical measure and 8 Machine-Learning classifiers, namely: SVM, MLP, Naïve-Bayes, Bayes Network, Simple-Logistic Regression, Voted Perceptron, J48-Decision and Random-Forest.

Acknowledgements: This research work had been made in cooperation with my PhD student Hassina Hadjadj.

16.1 Text Segmentation

A text segmentation is applied in order to construct individual documents with the same size. In fact, when comparing two books with different sizes, it is difficult to know if a specific part of the book is similar to another one or not. That is why a smart segmentation has been proposed and applied to the different books.
The sizes of the segments are more or less in the same range: we have 29 different text segments for the Quran and 8 different text segments for the Hadith, with approximately the same size. Hence, we get 37 different text segments of about 2900 words each.
The segmented dataset is decomposed into 2 parts: training part and testing part, and since the two books have different sizes (29 texts for the Quran and 8 texts for the Hadith), a logical rule has been established: 4 text segments are used for the training of the Hadith book and 8 text segments are employed for the training of the Quran book. And the remaining text segments are used for the testing step.
The choice of the training dataset size is defined by a special parameter we called big/small, which gives a qualitative estimation on the size of the book.

16.2 Proposed Interrogative Features

In our investigation, a mixture of 26 interrogative features is proposed. All those features are original and used for the first time in stylometry (when writing this article).

Why Interrogative features?
An important step consists in retrieving distinctive features that exhibit the writing style and represent a certain authorship individually.
Arabic is one of the world's greatest languages. Its graceful script, magnificent style and rich vocabulary give to the language a unique character and specificity. However, when we read the two books, we notice that the Interrogative form is widely used within the books. The interrogative form is one of the specificities that characterizes the Arabic language and has a great variety and richness in the literary style.
During the preparation of this research work, there were no published works in stylometry using interrogative features, especially in Arabic, despite the importance of the interrogative style in all languages of the world. This fact encourages the idea of establishing a research in Arabic authorship discrimination using those interrogative features.

Questioning features in Arabic
Each language uses morphemes and specific interrogative structures, and the interpretation of the meaning depends on the social and cultural status of the speakers. Thus, the question differs in its interpretation and purpose concerning the communication situation. Furthermore, the question mark in Arabic has some properties that distinguish it from the English language and other languages.
The Arabic language, like English, has two types of questions: direct questioning and indirect questioning. The latter does not end with a question mark and has different properties from those documented in the English language (Mohamed-Cherif, 2007) (Naghich, 2012). The Arabic interrogation is total or partial depending on

its scope, and according to its value, it can be seen as in the case of the English language: rhetoric, inquiries, disguised order, etc.
Interrogative modality in Arabic has some characteristics that depend not only on the methods of the interrogation, i.e. intonation, interrogative morpheme and phrasal structure, but may also vary in terms of goals and sometimes the way of speaking.

Types of questions in Arabic

Direct and indirect inquiry

The Arabic language has two types of questions: the first is direct and it ends with a question mark, the second is called indirect and ends with a full-stop: it is a question where the verb introducer is semantically interrogative. Arabic, in questioning, is generally defined by two structures: one is the enquiry that is to say the interrogative element that appears on the original sentence or "Ism Al-istifham", "harf Al-istifham" (see below) and the other one is ensured by some specific verbs.

ألا ترى أنه غريب ؟
Do not you see he is a foreigner?

أتسأل إذ كنت ترى أنه غريب
I wonder if you see that he is a foreigner

Question 1 represents a direct query. However, it can be ensured by using some specific verbs such as the verb "TasâAla تسأل " (meaning: I wonder if) to become an indirect question, such as in question 2. The verb "TasâAla تسأل" can be prefixed with the morphological tag bearing the features of the type of the subject in the unfinished appearance (a: I), (ta: you), (there: he, she, they) and (na: us). The indirect question is introduced by a verb like "TasâAla تسأل ", for instance.

The total and partial query

Depending on the scope of the question, the question is answered by an affirmation or a reversal of the content of the statement. In this case, it is called total interrogation, and is introduced by the interrogative element, namely "Harf Al-istifham?" ("Hal هل" and "Hamza aa أ", which are both equivalent to the auxiliary verb do in English) as shown in the following example below. Both sentences have the same meaning and the expected answer is yes or no.

هل أتى أحمد؟
Did Ahmed come?

أدخل أحمد ؟
Has Ahmed entered?

The total indirect and polling can be introduced by a verb in interrogative semantic such as "tasâAla تسأل" or interrogative proposal such as "uridu an aårifa: أريد أن أعرف " as shown in the following examples :

أريد أن أعرف هل أتى أحمد
I want to know if Ahmed came.

أتسأل أ دخل أحمد
I wonder if Ahmed entered.

The partial query is introduced by an interrogative pronoun called "Ism Al-istifhâm" such as (ما: what, من: who, كيف: how, أين: where). It can be, like the total query, direct or indirect. In the case of a partial direct question, the interrogative pronoun is at the initial position of the interrogative sentence. In the case of a total

indirect question, the interrogative pronoun is replaced by a subordinating conjunction that introduces the sentence, as shown in the following examples:

أين ذهبت ؟
Where was you gone?

أتسأل أين ذهبت
I wonder where you went

The semantic features of interrogative prepositions

The interrogative modality in English may be denoted through the ascending intonation or the presence of an interrogative element to the initial sentence. Another method is the indirect question, which is marked by the subordinating conjunction (if) or (that). However, in Arabic, there are several other conjunctions, for example: inn إن (if), idâ إذا (if), ma idâ ما إذا (in case of).
Interrogative morphemes are used to check the trueness of a statement or question in the interrogative sentence.

Pronouns

- هل / hal: meaning "does-it", or "is-it" : this is an interrogative pronoun used to ask questions about the trueness of a fact, for example:
  - Did he come ? هل جاء ؟
- أ /Aa: meaning "does-it", or "is-it": this is an interrogative pronoun covering the trueness of a fact. It has the same meaning as "hal", for example:
  - Did he come ? أجاء ؟

- ماذا/ mâdâ: meaning "what": this is used in questions for the determination or specification of an object/ act, for example:
  - What are you doing? ماذا تفعل ؟

- ما/ ma: meaning "what": this is a variant of "ماذا"/ mâdâ, for example:
  - What made you unable to come today? ما دهاك أن لا تأتي اليوم ؟
  - What do you have? ما لديك ؟

- أي/ Ayyu: meaning "what/ which": it is used to ask the identity of someone or something.
  - Which guest came? أي ضيف جاء ؟

- من/ man: meaning "who": the interrogative pronoun may be concerned with the subject or a determinative pronoun.
  - Who says that? من يقول هذا ؟

Adverb

When the question is submitted by an interrogative adverb, it becomes an adverbial phrase.

- لماذا/ Limada: meaning "why" (or with variations " لما lima"): it is used to request the cause of an event. There is also the variation" بماذا/ bimâdâ" meaning "with what", which is used to build an adverbial question, example:
  - Why do you come now? لماذا جئت الآن ؟
  - With what you write? بماذا تكتب ؟
- كيف / kayfa: meaning "how": it is used to question about the way or state of something.

- How did you get here? كيف وصلت إلى هنا ؟
- How is your new car? كيف هي سيارتك الجديدة ؟
- How are you? كيف حالك ؟

- كم / kam: meaning “how many/much”: this interrogative word is used to inquire about the duration or quantity, example:
  - How many days you were absent? كم من يوم غبت ؟
- متى/ mata: meaning “when”: it is basically used for interrogative questions about the time.
  - When will you come back? متى ستعود ؟
- أم / am: meaning “or/ isn’t it”: it is used to specify which alternative is true and which one is false.

The particle "Am" is used concurrently with the interrogative particle "Aa", which is sometimes implied. This particle is used when the speaker knows that the fact is true, and wonders who/which associated person/thing is concerned with, example:

Who is with you: Zayd or Amr? أعندكم زيد أم عمر ؟ / عندكم زيد أم عمر ؟

Interrogative Features selection
According to the importance of the interrogative style in our investigation, a set of 26 interrogative features was selected and extracted and collected from the two religious books.
This set contains four types of interrogative features as follows:

- Interrogative words/verbs with the particle Hamza (أ / aa): this type of interrogation is commonly used in old Arabic and this particle could represent the question mark “?” or an auxiliary verb in the interrogative form. In our approach, we employed 22 interrogative elements of this type.
- Wondering verb “يسألونك”, meaning “they ask you” in English: this type represents an indirect questioning.
- The starting particle أم meaning “or/ isn’t it” in English. When it is employed at the beginning of a sentence, it refers to a form of question.
- The interrogative expression “من ذا” meaning “who?” in English: this interrogative expression is interesting since it gives a feeling of a strong personality when speaking.
- The interrogative expression “ما لكم” meaning “what is the matter with you? / why?” in English. This interrogative expression is also interesting for its strong style when speaking.

16.3 Authorship Discrimination Approach

In our method, several steps are performed, as shown in Figure 16.1, namely: data pre-processing, feature extraction, classification and author discrimination decision, while the data set is organized into training and testing.

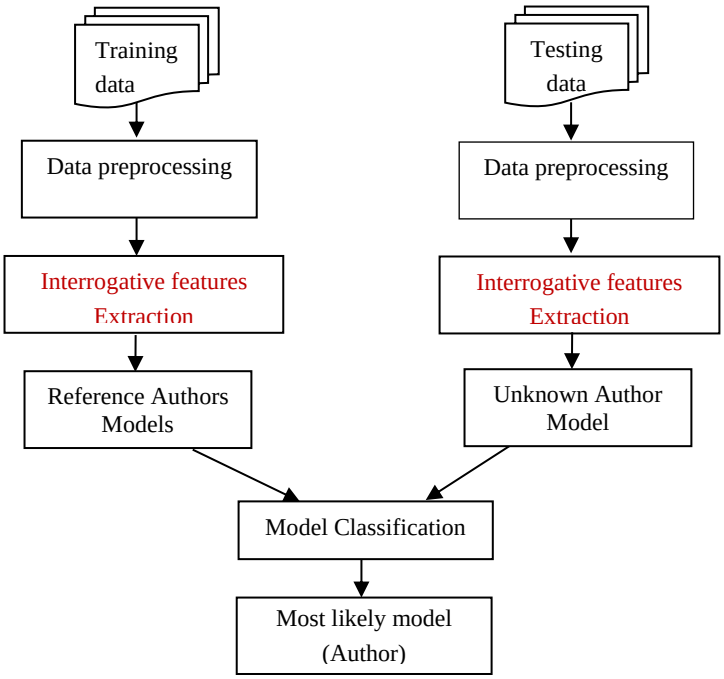


Figure16.1. Authorship discrimination method

16.4 Feature extraction and LFF fusion

This is an important step in author identification. In our case, we recall that the originality of this investigation consists in the use of the proposed interrogative features.
Furthermore, we have defined the Logarithmic Feature Fusion (LFF) as the logarithmic sum of the normalised features frequencies (see equation 16.1). A graphical representation of the LFFs for the 37 text segments is displayed in figure 16.2.

We can notice that all the Quran LFFs are positive while the Hadith LFFs are negative.

The Logarithmic Feature Fusion (LFF) is given as follows:

$$Fusion = \log\left(\sum_i feature_i / \max\left(feature_i\right)\right) \qquad (16.1)$$

whith i representing the feature index (i=1,2,...26).

Figure 16.2 shows a sharp difference between the Quran segments, which present positive logarithmic values, and the Hadith segments, for which the logarithmic values are negative.

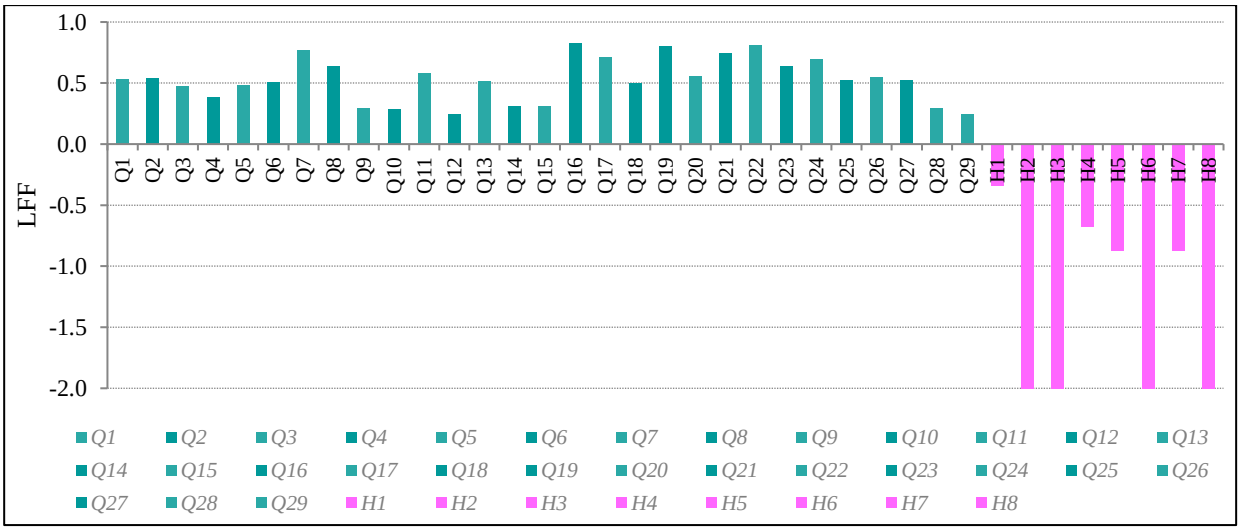


Figure.16.2. LFF values of the Quran text segments (in dark green) and Hadith text segments (in pink)

16.5 Classification methods

Description of the classifiers

The different experiments of authorship attribution are made by using several classifiers, namely:

- Euclidean Distance
- Multi-layer-Perceptron
- Naïve-Bayes classifier
- Bayesian-Network
- Support Vector Machines
- Voted Perceptron
- J48-Decision-Trees
- Random-Forest
- Simple-Logistic-Regression

Performance evaluation

The performance given by the classifier is measured in terms of accuracy of classification, which is calculated as the ratio of the number of correct attributions over the total number of testing segments.

$$Accuracy = \frac{Number\ of\ correctly\ classified\ segments}{Total\ number\ of\ tested\ examples} \qquad (16.2)$$

16.6 Experimental results and analysis

This investigation performs a segmental analysis on the two religious books: Quran and Bukhari Hadith, for the task of authorship discrimination. The analysis of the different text segments is performed

by using 27 interrogative features (i.e. 26 interrogative features + their fusion) and by employing 9 different classifiers.

The dataset is composed of 37 different text segments of about 2900 words each, consisting of 8 Hadith segments {H1, H2, …H8} and 29 Quran segments {Q1, Q2, …Q29}. During the experiments, 4 segments of the Hadith and 8 segments of the Quran are used for the training and the remaining segments are used for the testing. Therefore, there are 25 different segments to classify according to 2 referential Authors (Quran Author or Hadith Author).

Table 16.1 shows the A.A. accuracy on the two religious books. From this table, we notice that the proposed feature set is powerful for discriminating the two books styles and we can notice that all the Quran segments are attributed to the referential "Quran Author" and all the Hadith segments are attributed to the referential "Hadith Author". That is, the 25 different text segments are classified into 2 main classes: "Quran class" and "Hadith class", with a performance accuracy of 100%. Hence, we can say that the new proposed features have efficiently separated the two books according to their interrogative styles.

So, the proposed feature set appears to be effective in authorship attribution, especially with the Arabic language. Furthermore, since only interrogative features are employed, this analysis represents an evaluation of the interrogative style in each book.

Experimentally speaking, we noticed that the interrogative styles in the Quran and Bukhari Hadith are completely different (an accuracy of 100% is achieved with all the used classifiers).

TABLE 16.1. Authorship Attribution accuracy on the 2 religious books by using the new proposed features

Classification Algorithm	Accuracy	Attribution Error
Euclidean Distance	*100%*	*0%*
SMO-SVM	*100%*	*0%*
MLP	*100%*	*0%*
Naïve-Bayes	*100%*	*0%*
Bayes-Network	*100%*	*0%*
Simple-Logistic-Regression	*100%*	*0%*
Voted Perceptron	*100%*	*0%*
J48-decision Tree	*100%*	*0%*
Random-Forest	*100%*	*0%*

***Notes*:** It has been noticed that the particle hamza is very commonly used in the Quran book; whereas, in the Hadith, this particle is rarely used. Again, we noticed that the following interrogative expressions «أغير, أفلا, أأنت, أأنتم, أتريدون ... » are more commonly used in the Quran; but almost never used in the Hadith. Moreover, we also noticed that the Author of the Quran used indirect questioning form quite frequently, whereas in the Hadith it is not used at all, which shows that the two authors possess two distinct interrogative styles that are completely different.

According to the structure and the morphology of the interrogative style in the Arabic language, it is very difficult to write two books in the same topic with so many differences in the style, and we do not see any other explanation except the fact that the Quran and Hadith should have two different authors.

16.7 Discussion

This research work has addressed the problem of author classification by segmenting the concerned documents into several texts segments of the same size. The originality of this research work lies in the use of a new set of features based on 26 interrogative features exclusively and a special fusion; and it is the first time that those features are used alone in stylometry.

That is, different classification algorithms were used to build an automatic classification model to discriminate between the authors: Centroid-Euclidean distance, Support Vector Machines, Multi-Layer-Perceptron, Naïve-bayes, Bayes-Network, Simple-Logistic-Regression, Voted-Perceptron, J48-Decision-tree algorithm and Random-Forest.

A specific application of this research work concerned the problem of author discrimination between two old Arabic religious books: The Quran and Hadith to check whether the two books could have been written by the same author or not.

Basically, the different experiments of authorship attribution have led to the following important points:

- Concerning the proposed interrogative features, they appear to be interesting for performing author identification tasks in Arabic;
- Concerning our application of author discrimination between the two religious books, the experimental results have clearly revealed that the interrogative styles structures of the Quran and Hadith are very different.

Now, despite this result, if we assume that the two books belong to the same Author, then an important arising question would be: Could an author possess two interrogative styles completely different for the same topic?

If so, what could be the reason to change his questioning style? And in what proportions would it be possible to do that variability?

The only possible response to all those questions leads to the fact that there could be no valid reason to use two dissimilar interrogative styles completely different by the same author.

Actually, we do not see any other explanation except the fact that the two studied books should have two different Authors.

17 Eleventh Series of Experiments: Investigation on the Quran/Hadith Authorship Using Visual Analytics Approaches

An important raising question, in the stylometric study of the Quran, is: Was this religious book written by the Prophet? An interesting scientific way, to answer that question, is the use of authorship attribution techniques. However the use of conventional features and classifiers has some disadvantages such as the automatic authorship decision, which usually gives us a speechless authorship classification without (often) any way to measure or interpret the consistency of the results.
In this research work, we present a visual analytics based investigation for the task of authorship discrimination between the holy Quran and Hadith.
Seven types of features are combined and normalized by PCA (Principal Components Analysis) reduction and seven visual analytics based clustering methods are employed and commented, namely: Hierarchical Clustering, Fuzzy C-mean Clustering, K-mean Clustering, Sammon Mapping, Principal Component Analysis, Gaussian Mixtures Models and Self Organizing Maps.
Results are quite interesting and the disposition of the visual clusters provide valuable information on the Authors' styles.

17.1 Stylometric features

Seven types of features have been proposed for this task of AA, as one can see in the following subsections.

17.1.1. Author's Pronoun Based Stylistics

In the Quran, we notice that the author's pronouns He (هو) and We (إنا - نحن) are the most used ones for representing Allah (God) in the holy Quran. See some examples below:

- هو الذي خلق لكم ما في الأرض جميعا ثم استوى إلى السماء فسواهن سبع سموات وهو بكل شيء عليم (He)
*"**He** it is who has created for you all that is on earth, and has applied **His** design to the heavens and fashioned them into seven heavens; and **He** alone has full knowledge of everything"*

- وإذ واعدنا موسى أربعين ليلة ثم اتخذتم العجل من بعده وأنتم ظالمون (We)
*"And (remember) when **We** appointed for Mûsa (Moses) forty nights, and (in his absence) you took the calf (for worship), and you were Zâlimûn (polytheists and wrong-doers)"*

However, in some few cases, the pronoun I (أنا - إني) is also used, but in special circumstances. See some examples below:

- فاذكروني أذكركم واشكروا لي ولا تكفرون (I / Me)
*"So remember **Me**, and **I** shall remember you; and be grateful unto **Me**, and do not deny **Me** "*

- وإذا سألك عبادي عني فإني قريب أجيب دعوة الداع إذا دعان فليستجيبوا لي وليؤمنوا بي لعلهم يرشدون (I / Me)
*"And when **My** servants ask you, (O Muhammad) ,concerning **Me** – indeed **I** am near. **I** respond to the invocation of the supplicant when he calls upon **Me**. So let them respond to **Me** (by obedience) and believe in **Me** so that they may be (rightly) guided"*

Perhaps, in those last examples, the pronoun I is used for representing the closeness of Allah (God) to his believers, or perhaps, the pronoun I is used for meaning that Allah (God) is ONE, but when speaking about God statements, Allah usually uses the pronoun "**We**".

Consequently, in the Quran, we notice that Allah (God) uses the 3 pronouns: **Me** / **I (**أنا-إني), **We (**نحن-إنا**)** and **He (**هو**)** for representing His Excellency Allah.

Differently, in the Hadith, the Prophet Muhammad uses exclusively the pronoun: **Me** / **I (**أنا-إني) for representing himself.

17.1.2. On the use of "أبا" (father of) for naming people

In the Arabic language, it is usual to call a person using the name of his oldest child (often the son). That is, if somebody has a son called Youssof for instance, then it is possible to call him *Aba*-Youssof, which can be translated in English into *Father-of*-Youssof. This fact was often noticed in the Hadith, when the Prophet speaks to his companions.

Nowadays, although this mean of appellation has become rare, it is still widely employed in some countries of the Arabic gulf region and middle-east.

On the other hand, namely in the Quran, the style of appellation is quite different, since the persons are directly called by their own names (first-names), such as in the following verse:

(78) وداوود وسليمان إذ يحكمان في الحرث إذ نفشت فيه غنم القوم وكنا لحكمهم شاهدين

Which could be translated into: And [mention] David and Solomon, when they judged concerning the field - when the sheep of a people overran it [at night], and We were witness to their judgement (78).

17.1.3. Frequency of some discriminative words

As first feature, we have proposed to use 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 Quran book; whereas, in the Hadith, these words are rarely used.

17.1.4. On 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 (H. Sayoud, 2012). The description of this parameter can be found in section 7.3

17.1.5. Word length frequency

The fifth feature is the word length frequency. Herein, the word length is the number of letters composing that word and 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.

17.1.6. Frequency of the coordination conjunction «و» (meaning AND in English)

An interesting type of features corresponds to the coordination conjunctions, which are widely used in the two investigated books.

In reality, we limited our investigation to one of the most interesting conjunction, which seems to be very discriminative between the two books: it is the conjunction " و" , which corresponds to the coordination conjunction AND (*in English*) and which is widely used in Arabic.

17.1.7. Frequency of the coordination conjunction Waw «و» at the beginning of sentence

Similarly to the previous section, herein we are still interested in the frequency of the coordination conjunction " و". However, in this case we only keep the conjunctions that are localized at the beginning of sentences, such as in the following sentence: "And now, what should we do?"

In fact, we noticed that several verses in the Quran begin with that coordination conjunction, which is relatively rare in Arabic. Even in English, it is extremely rare to find a phrase beginning by "And", because as its appellation says (i.e. coordination), it connects/coordinates two sentences, two verbs or two names and consequently it may not be localized at the beginning of a sentence, except in few rare cases.

17.2 Visual Analytics based Clustering methods

By definition, the term clustering corresponds to the fact of grouping some things together; which can be physical objects, numerical data, concepts or any sort of elements.

In pattern recognition, cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (*ie. cluster*) are more similar to each other than to those in other groups (Norusis, 2008). This task is commonly used in data mining, statistical data analysis, machine learning and information retrieval.

On the other hand, visual Analytics (Ellis et al., 2010), which is a combination of several fields (*ie. computer science, information visualization and graphic design*) is often used in cluster analysis to make the analyst's judgment easier to develop and more objective.

That is, the combination of those two research fields can lead to a strong and efficient analysis tool for handling some classification tasks that could be extremely difficult to perform with conventional analytic tools.

Furthermore, a great advantage of clustering over conventional classification tools is its non-supervised property (for several clustering techniques).

Consequently, it appears that the association of visual analytics with clustering analysis may be interesting for solving some stylometric problems, for which we do not possess any training possibility or information to make a supervised classification task. So, it should be extremely motivating to apply them in our main task of authorship discrimination (ie. Quran vs Hadith).

Concerning the methods using the association of visual analytics with clustering analysis, there exist several approaches that have been proposed during the last five decades, such as: K-mean Clustering, Hierarchical Clustering, Sammon Mapping, Self Organizing Maps, Gaussian Mixtures Models, Fuzzy C-mean Clustering, Principal Component Analysis, etc.

In this survey, we propose to use all those seven methods separately in order to find out the possible clusters related to the different investigated texts.

17.2.1 Hierarchical clustering

Our first clustering method is based on the hierarchical clustering.

Definition

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters (Greenacre, 2014). In general, there are two types:

-Agglomerative clustering: This is a "bottom up" approach, where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

-Divisive clustering: This is a "top down" approach, where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In our case, we used the first clustering type with a Manhattan distance measure, which is defined below:

If we assume that X and Y represent two vectors, then Manhattan distance (between those 2 vectors) is given by the following equation:

$$d(X,Y) = \sum_i |x_i - y_i| \quad (1)$$

The resulting linkage of the different documents is called “Dendrogram”. It represents the different possible clusters in a graphical way. By observing the dendrogram, it will be possible to estimate the actual number of clusters and the corresponding documents for each cluster, since all similar documents should be linked together with a consistent linkage.

Results of the Hierarchical clustering

The hierarchical clustering has yield to the following dendrogram, where we can observe two separate clusters, one cluster in red in the right and another one in blue in the left. The corresponding result shows clearly that the two investigated documents (i.e. Quran and Hadith) have two different styles.

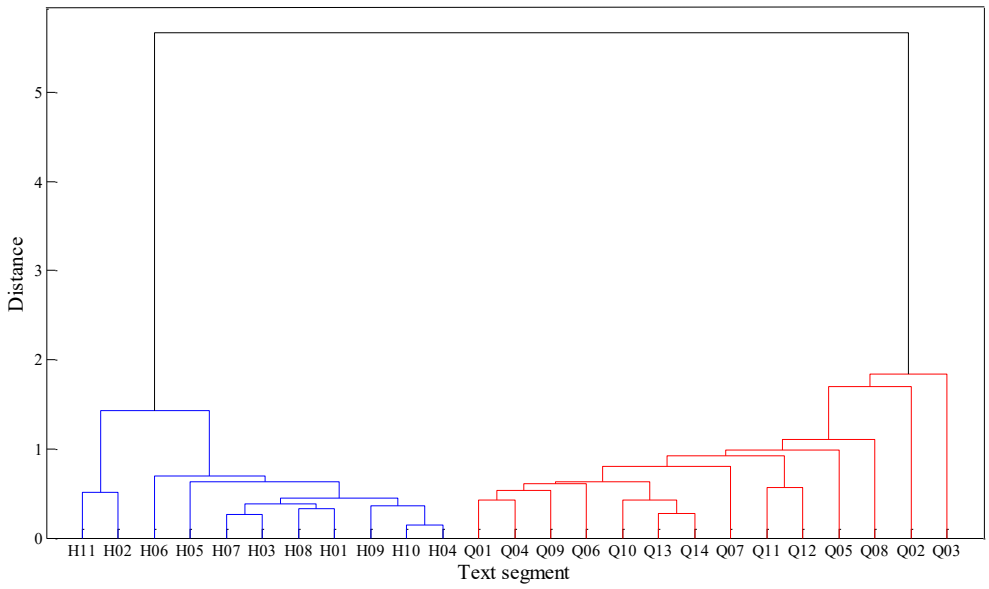


Figure 17.1: Results of the Hierarchical Clustering

17.2.2 C-Means clustering

Definition

Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" (all or nothing) but "fuzzy" in the same sense as fuzzy logic (Suganya, 2012).

In fuzzy clustering, every point has a degree of belonging to clusters, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be *in the cluster* to a lesser degree than points in the center of cluster.

That is, any point x has a set of coefficients giving the degree of being in the k^{th} cluster $w_k(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$C_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m} \quad (2)$$

In 3D or 2D dimensions, Fuzzy C-mean can provide a graphical representation of the different samples and the corresponding clusters to which they should belong. This representation allows separating the different samples with regards to their similarities automatically and in a visual manner.

Results of the Fuzzy C-Means clustering

The Fuzzy C-Means clustering has provided the following 3D representation, where we can observe two separate clusters, one cluster in red in the right and another one in blue in the left. This result shows that the two investigated documents (i.e. Quran and Hadith) should have two different styles.

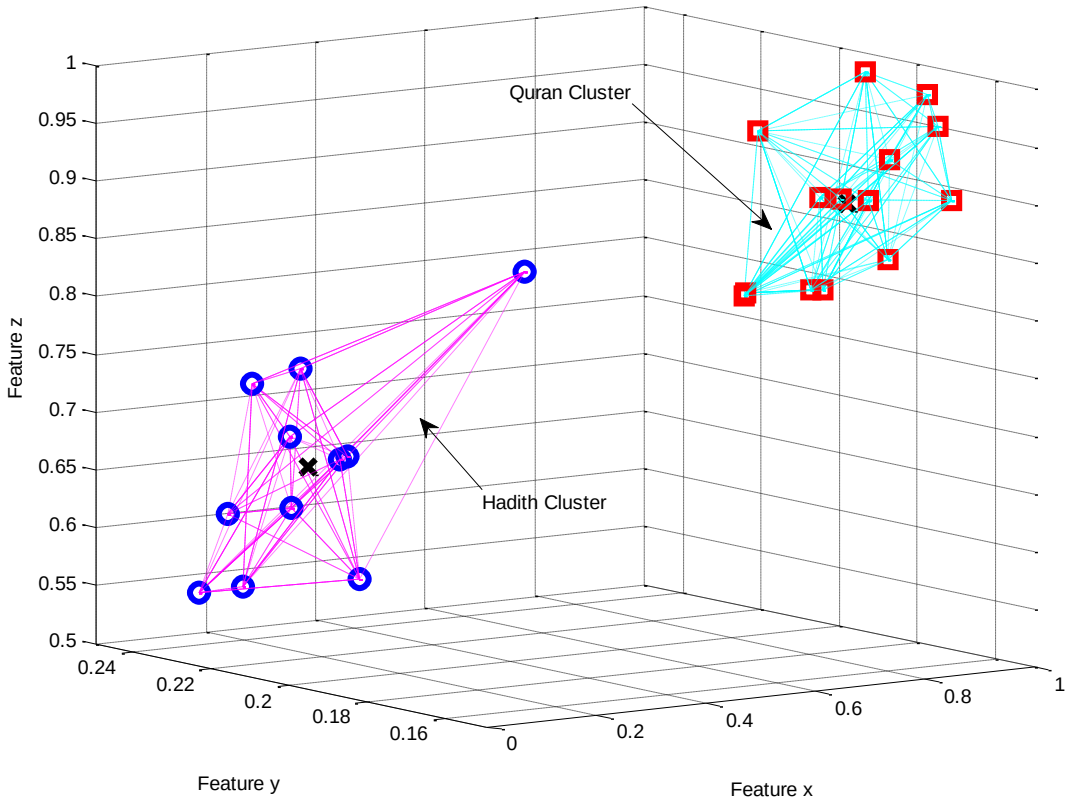


Figure 17.2: Fuzzy FCM Clustering

17.2.3 K-Means clustering

Definition

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis (Kardi, 2007).

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

So, given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS). In other words, its main objective is to find the following value:

$$argmin_S \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \qquad (3)$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

As in Fuzzy C-mean, the K-means can provide a graphical representation of the different samples and the corresponding clusters to which they should belong. This representation allows separating the different samples with regards to their similarities in an automatic way.

Results of the K-Means clustering

The K-means based clustering led to the following 3D representation, where we can easily notice that the different text segments have been grouped into two main clusters: a Quran cluster in the right and Hadith cluster in the left of the 3D representation. This sharp separation suggests that the two types of texts should have two different authors.

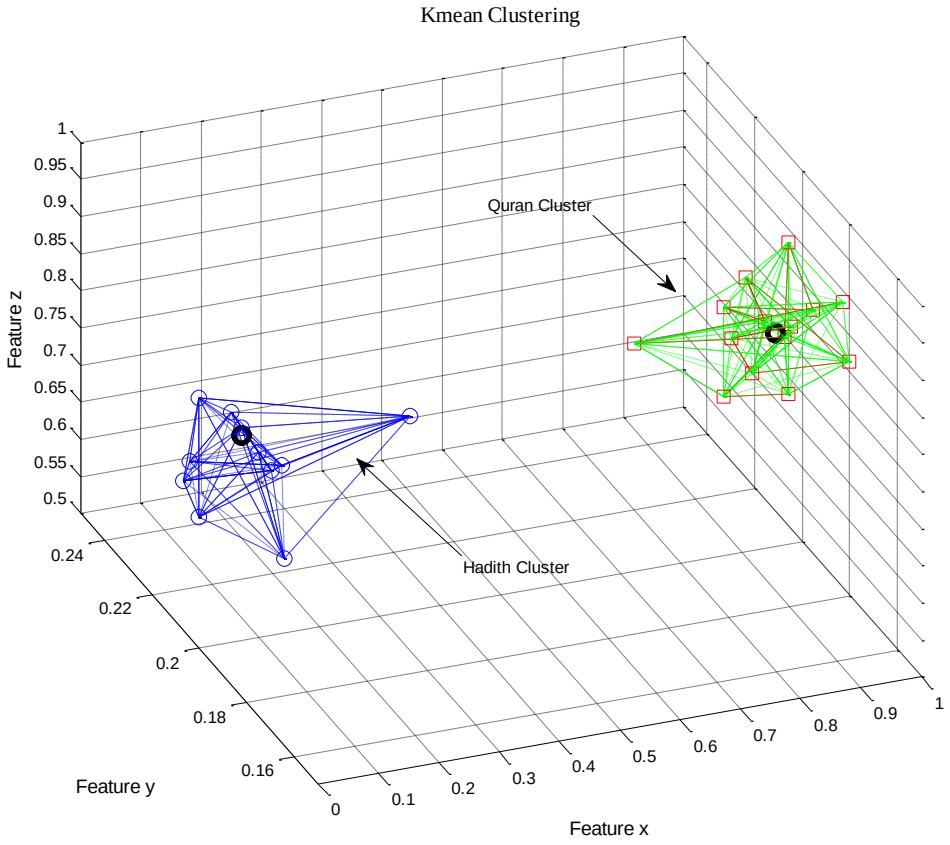


Figure 17.3: K-means Clustering.

17.2.4 Sammon Mapping

Definition

Sammon mapping or Sammon projection is an algorithm that maps a high-dimensional space to a space of lower dimensionality by trying to preserve the structure of inter-point distances in high-dimensional space in the lower-dimension projection (Kim, 2003). It is particularly suited for use in exploratory data

analysis. The method was proposed by John W. Sammon in 1969. It is considered a non-linear approach since the mapping cannot be represented as a linear combination of the original variables such as in the principal component analysis.

That is, by denoting the distance between the i^{th} and j^{th} elements in the original space by d^*_{ij}, and the distance between their projections by d_{ij}. Sammon's mapping aims to minimize the following error function, which is often called Sammon's error:

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}} \quad (4)$$

By choosing a 3 or 2 dimensional space, the Sammon-based graphical representation is quite interesting, since it makes a sharp separation of the different elements by bringing closer all the similar ones.

Results of the Sammon mapping

A 3 dimensional reduction has been employed using Sammon representation. The 3 retained features are denoted by alpha, beta and gamma. Thereafter, the 25 text samples are represented according to those 3D axes/features. The resulting visual representation shows 2 main clusters: one in the right grouping all the Hadith segments and another one in the left grouping all the Quran segments. Furthermore those two sets of texts are covered by an interpolated surface between samples of a same type in order to better show the different resulting clusters. In other words, those surfaces are not provided by Sammon mapping but by our personal algorithm for a visual confort only.

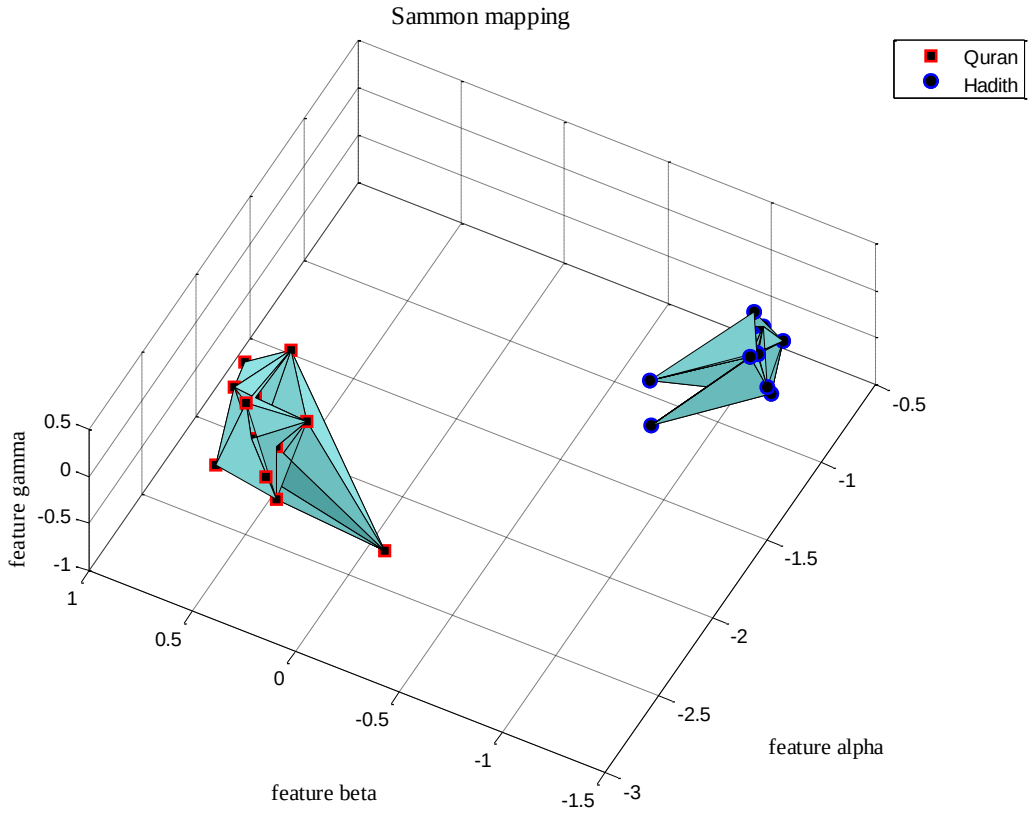


Figure 17.4: Sammon Mapping and intra-elements surface interpolation

17.2.5 Principal Components Analysis

Definition

Principal component analysis (PCA) can be considered as one of the most interesting results of applied linear algebra. PCA is used abundantly in all forms of analysis - from neuroscience to computer graphics, because it is a simple and non-parametric method of extracting relevant information from confusing data sets. With minimal additional effort PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that often underlie it (Shlens, 2003).

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (i.e. the first principal component), the second greatest variance on the second coordinate, and so on.

Consider a data matrix, X, where the sample mean of each column has been shifted to zero and where each of the *n* rows represents a different repetition of the experiment, and each of the *p* columns gives a particular kind of datum (e.g. the results from a particular sensor). Mathematically, the transformation is defined by a set of *p*-dimensional vectors of weights or *loadings* $W_{(k)}=(w_1,\ldots,w_p)_{(k)}$ that map each row vector $X_{(i)}$ of **X** to a new vector of principal component *scores* $t_{(i)}=(t_1,\ldots,t_p)_{(i)}$, given by

$$t_{k(i)}= X_{(i)} \cdot W_{(k)} \qquad (5)$$

in such a way that the individual variables of **t** considered over the data set successively inherit the maximum possible variance from **x**, with each loading vector **w** constrained to be a unit vector.

PCA is quite interesting in complex data analysis, when the most important features are not known in advance. And by reducing the dimensionality to a lower more consistent one, the data analysis becomes usually easier and more pertinent.

Results of the PCA analysis

A PCA representation of the data, using the 3 most important eigenvectors, is given in the following figure. The Quran texts are symbolized by red circles, while the Hadith texts are symbolized by blue crosses. In that figure, we can notice that all the Quran documents are grouped together in the right side, while all the Hadith ones are grouped in the left side. Once again the stylistic discrimination can be easily noticeable in the 3D representation.

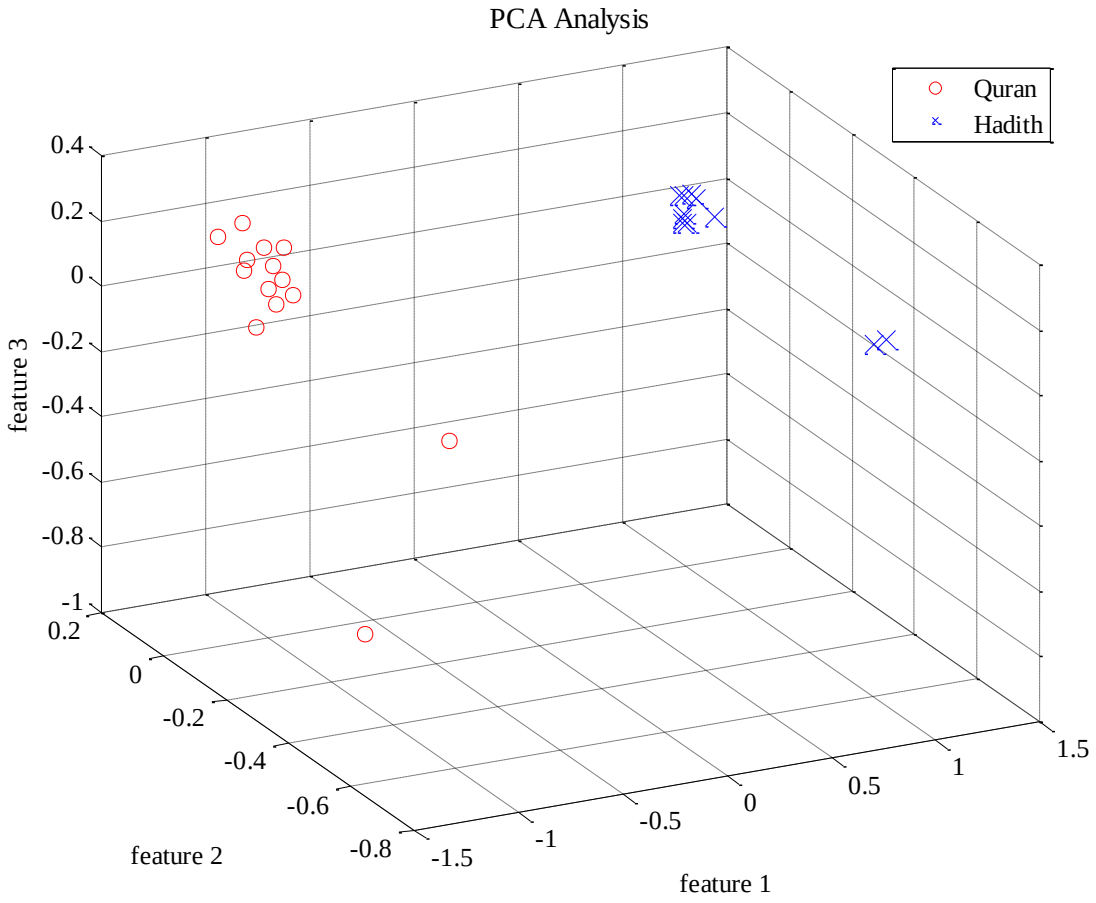


Figure 17.5: PCA representation of the data using the 3 most important eigenvectors. Quran texts are represented by red circles while Hadith texts are represented by blue crosses.

17.2.6 Gaussian Mixture Model based clustering

Definition

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modelling of a wide variety of random phenomena (McLachlan, 2003). Because of their usefulness as an extremely flexible method of modelling, finite mixture models have continued to receive increasing attention over the years, both from a practical and theoretical point of view. For multivariate data of a continuous nature, attention has focused on the use of multi-variate normal components because of their wide applicability and computational convenience. They can be easily fitted iteratively by maximum likelihood via the expectation maximization algorithm.

With a normal mixture model-based approach, it is assumed that the data to be clustered are from a mixture of an initially specified number g of multivariate normal densities in some unknown proportions pi_1, …, pi_g, That is, each data point is taken to be at realization of the mixture probability density function.

$$f(u;\Psi) = \sum_{i=1}^{g} \pi_i \phi(y;\ \mu_i, \Sigma_i) \qquad (6)$$

where $\phi(y;\ \mu_i, \Sigma_i)$ denotes the p-variate normal density probability function with mean μ_i, and covariance Σ_i.

Here the vector Ψ of unknown parameters consists of the mixing proportions π_i, the elements of the component means μ_i and the distinct elements of the component- covariance matrices Σ_i.

Once the mixture model has been fitted, a probabilistic clustering of the data into g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data. An outright assignment of the data into g clusters is achieved by assigning each data point to the component to which it has the highest estimated posterior probability of belonging (McLachlan, 2001).

Results of the GMM based clustering

The GMM based clustering is performed after PCA reduction into the 2 most important components. That is, two types of visualizations are provided: a 2D representation (with those two components) and a 3D representation including the probability density function as third component (see figures 17.6.a and 17.6.b).

In both figures, we notice that the different text samples have been clustered into 2 main groups: Quran cluster, at the bottom left side, gathering all the Quran texts and a Hadith cluster at top right, gathering all Hadith texts.

In the 2D representation, the Gaussian mixtures are represented by different ellipsoids surrounding the two clusters, while in the 3D representation, the Gaussians are more visible since they are represented in form of 3D Gaussians surrounding the different clusters.

While, the first representation is sharper, the two representations are similar in terms of clustering information: so, we easily notice that all Quran texts are closely clustered together and all Hadith ones are closely grouped together too. This fact confirms, once again, that the two writing styles of the 2 books are probably different.

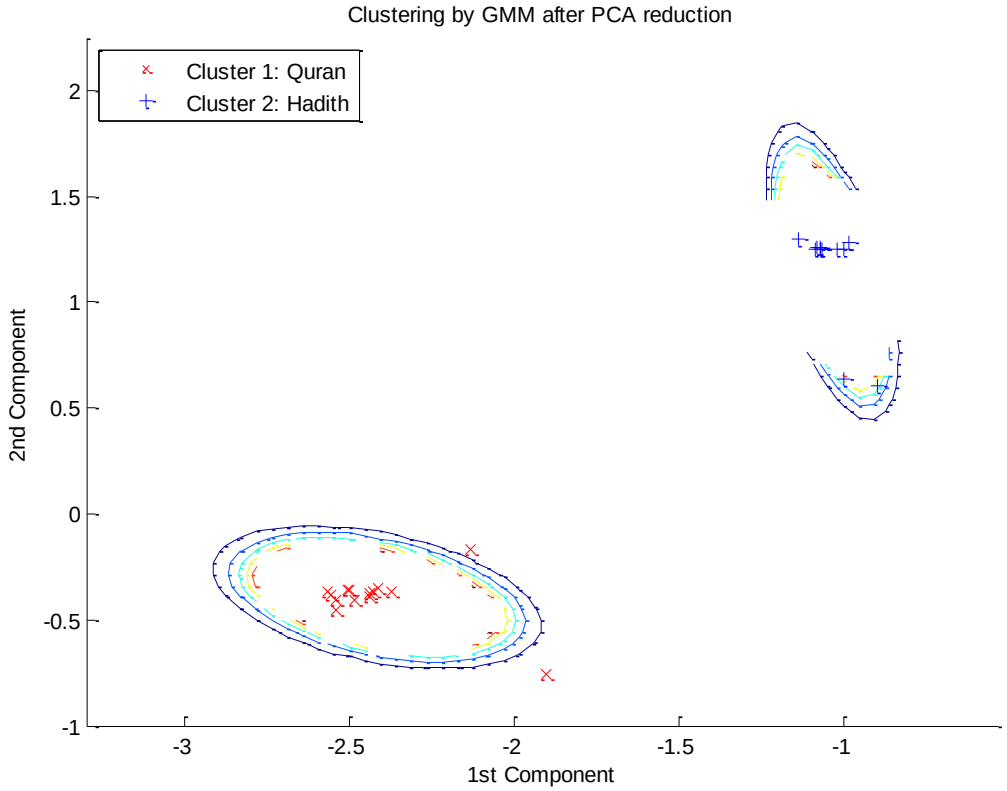


Figure 17.6.a: GMM clustering in 2D representation using two components.

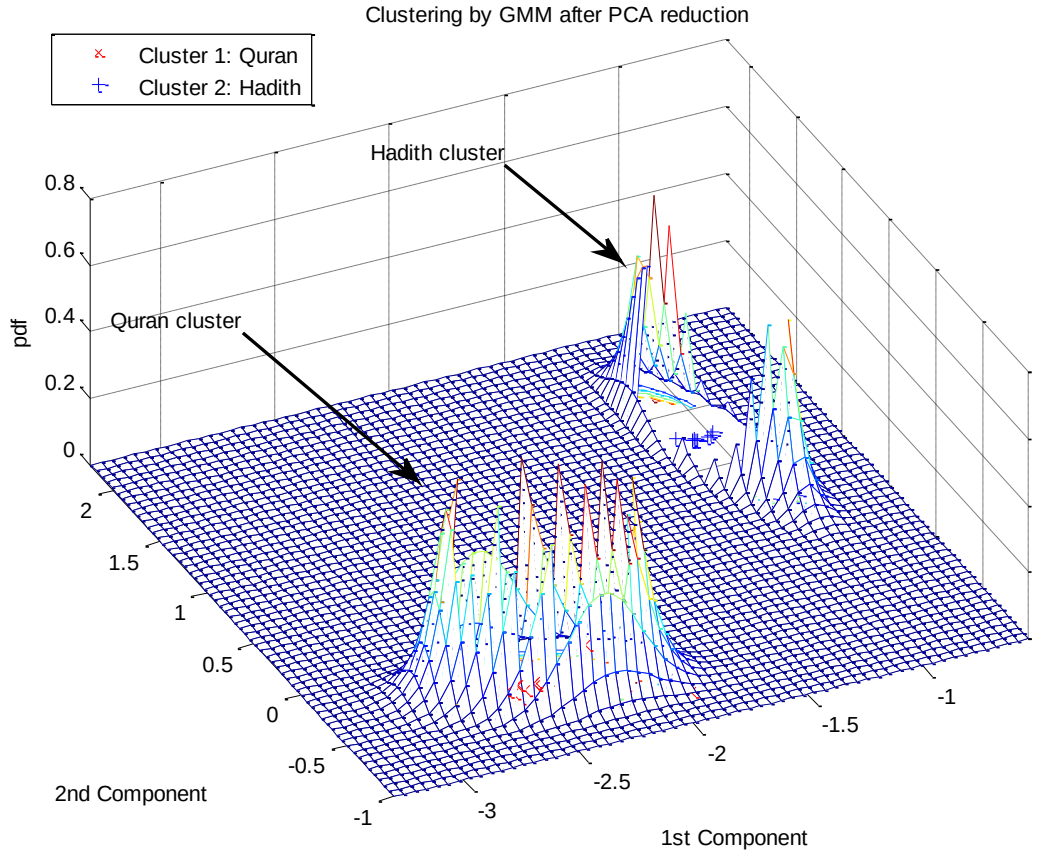


Figure 17.6.b: GMM clustering in 3D representation. The 3rd dimension represents the probability density function.

17.2.7 Self-Organizing Map based clustering

Definition

A Self-organizing Map is a data visualization technique developed by Teuvo Kohonen in the early 1980's (Kohonen, 1990) (Tambouratzis et al., 2003).

SOMs map multidimensional data onto lower dimensional subspaces where geometric relationships between points indicate their similarity. SOMs generate subspaces with an unsupervised learning neural network trained with a competitive learning algorithm.

The SOM learning tries to make the different parts of the network respond similarly to certain input patterns. This is partly motivated by how visual, auditory or other sensory information is handled in separate parts of the cerebral cortex in the human brain.

The weights of the neurons are initialized either to small random values or sampled evenly from the subspace spanned by the two largest principal component eigenvectors. The network must be fed a large number of example vectors that represent, as close as possible, the kinds of vectors expected during mapping. The examples are usually administered several times as iterations.

The training utilizes competitive learning. When a training example is fed to the network, its Euclidean distance to all weight vectors is computed. The neuron whose weight vector is most similar to the input is called the best matching unit. The weights of the best matching unit and neurons close to it in the SOM lattice are adjusted towards the input vector. The magnitude of the change decreases with time and with distance from the best matching unit. The update formula for a neuron v with weight vector $\mathbf{W}_v(s)$ is

$$W_v(s+1) = W_v(s) + \Theta(u,v,s)\,\alpha(s)(D(t) - W_v(s)) \qquad (7)$$

where “s” is the step index, “t” an index into the training sample, “u” is the index of the BMU for D(t), α(s) is a monotonically decreasing learning coefficient and D(t) is the input vector; Θ(u, v, s) is the neighborhood function which gives the distance between the neuron u and the neuron v in step s. Depending on the implementations, t can scan the training data set systematically (t = 0, 1, 2...T-1, then repeat, T being the training sample's size), be randomly drawn from the data set, or implement some other sampling method such as jackknifing.

The main advantage of using a SOM is that the data is easily interpreted and understood. The reduction of dimensionality and grid clustering makes it easy to observe similarities in the data. The major disadvantage of a SOM is that it requires necessary and sufficient data in order to develop meaningful clusters. Moreover, lack of data will usually add randomness to the groupings.

Results of the SOMs clustering

According to the figures obtained with SOM clustering, we can easily notice that there are mainly 2 clusters. Hence, in the 3D figure below (figure 17.7.a), representing the Distance matrix (inter-distances), We can see that there are 2 distinct dark regions (representing low inter-distances). Those 2 dark regions involve the presence of 2 distinct clusters, since every black area represents a cluster (in this case).

In the 2D figure below (figure 17.7.b), a Self-Organizing Map (SOM) using 3 PCA components has been performed. Here, the U-matrix is shown on the left, and an empty grid named 'Labels' is shown on the right.

In the left figure (U-matrix), we note 2 main clusters in white (the light colors represent clusters). The black (or dark) cells represent boundaries between clusters, unlike in the previous figure. Hence, one big cluster is visible at the right bottom and another big one at the left top.

In the middle figure, the different cells have been labeled (with regards to the book origin) by using 2 colors (red for the Quran and green for the Hadith), showing what texts belong to each cluster. Furthermore the distribution of the data set has been added to the map by using the corresponding hit histograms.

In fact an important tool in data analysis using SOM is called hit histogram. It is formed by taking a data set, finding the BMU (Best Matching Unit) of each data sample from the map, and increasing a counter in a map unit each time it is the BMU (Best Matching Unit). The hit histogram shows the distribution of the data set on the map. Here, the hit histogram for the whole data set is calculated and visualized on the U-matrix.

Once again, we notice that the Quran samples in red are well grouped together and separated from the Hadith samples in green, by a sharp horizontal black (dark) line.

Consequently, we can see that the SOM clustering leads to the same conclusion as previously, which is: the two books should have two different authors (or at least two different writing styles).

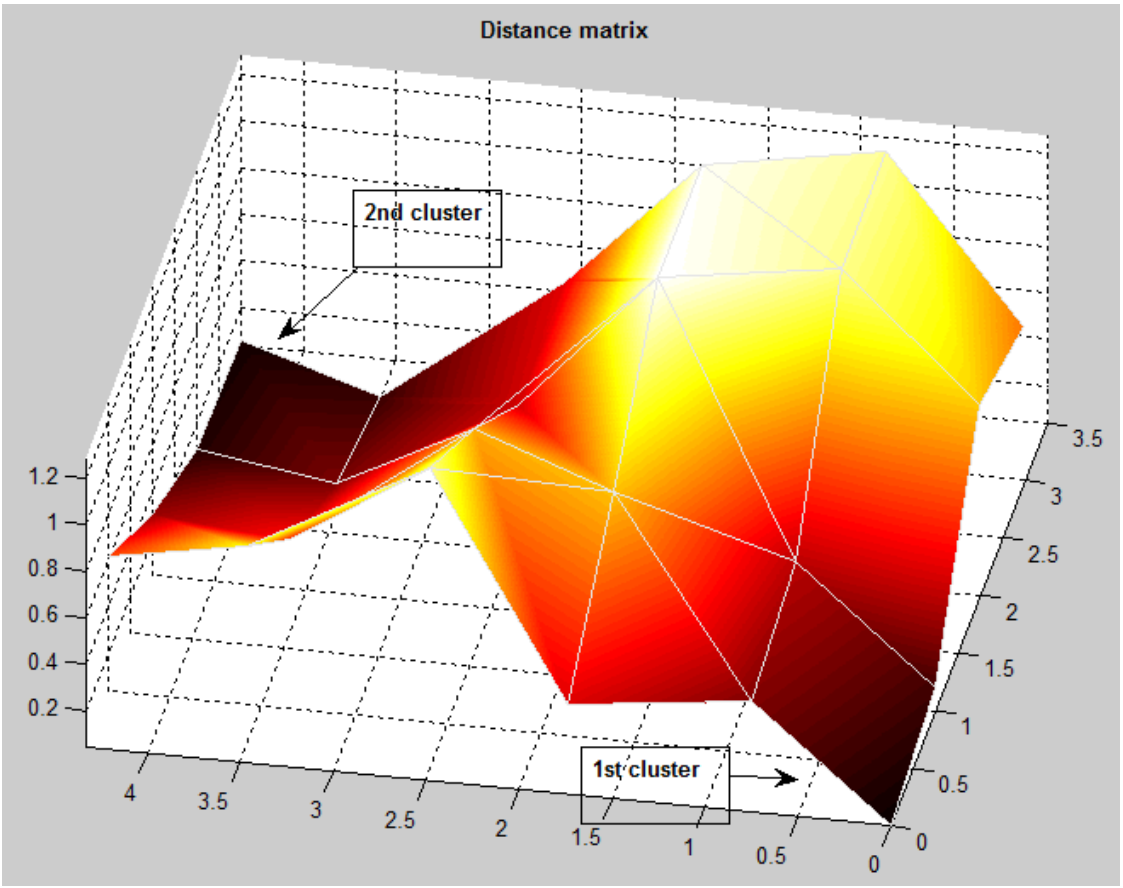


Figure 17.7.a: 3D representation of the Distance matrix (inter-distances). We can see that there are 2 distinct dark regions (representing low inter-distances). Those 2 dark regions involve the presence of 2 distinct clusters. In other words every black area represents a cluster.

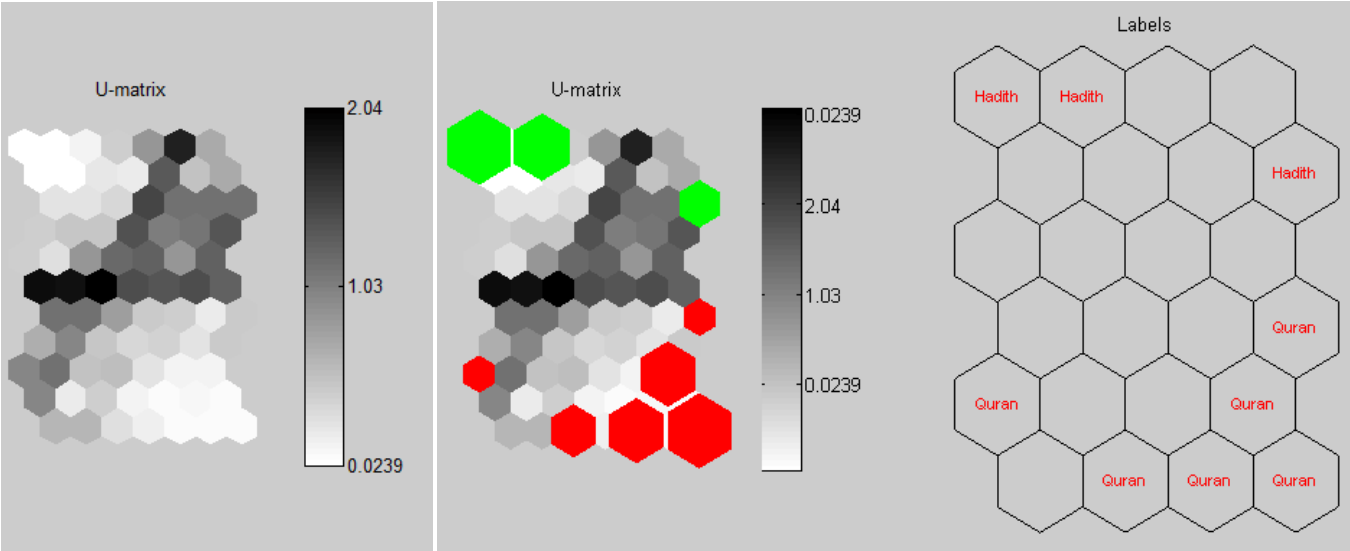


Figure 17.7.b: 2D Self-Organizing Map (SOM) using 3 PCA components. In the left figure, we can see 2 main clusters in white. In this representation, the light colors represent clusters. Hence, one cluster is visible at the right bottom and another one at the left top. In the middle figure, the different cells have been labeled by using 2 colors (green for the Hadith and red for the Quran), showing what are the cells belonging to each cluster. In the right, we only have the labels of the different SOM cells.

17.3 Results summarization

In this investigation, seven visual analytics based clustering approaches have been employed to make a visual authorship clustering of 25 religious text segments.

The different approaches are as follows: Hierarchical Clustering (HIC), Fuzzy C-mean Clustering (FCM), K-mean Clustering (KMC), Sammon Mapping (SAM), Principal Component Analysis (PCA), Gaussian Mixtures Models (GMM) and Self Organizing Maps (SOM).

- In the first approach, namely HIC, the resulting dendrogram has shown two separated clusters: the Quran cluster in the right and Hadith cluster in the left (see figure 17.1). We can also see that there is no intersection between the different clusters and that the final linkage is extremely weak since the corresponding distance is relatively very large. This result shows that there are two different writing styles and then probably two different authors too: Quran Author and Hadith Author.

- In the second approach (i.e. FCM), which is an automatic clustering technique, the resulting 3D representation shows two main clusters: Quran cluster located at the right top area and Hadith cluster located at the left bottom area of the 3D representation (see figure 17.2). Although the Quran cluster is more condensed, the two sets of text segments have been automatically organized into 2 sharp clusters (with different symbol markers), showing that there are probably two main authors: Quran Author and Hadith Author and that the two authors are different.

- In the third approach (i.e. KMC), which is also an automatic clustering technique, the three-dimensional K-mean representation reveals two main clusters too: Quran cluster in the right and Hadith cluster in the left (see figure 17.3). Those two clusters are distant and well separated one from the other. Consequently and as previously, it is clear that the two authors (Quran Author and Hadith author) should be different.

- The fourth approach, called SAM, is not a real clustering method but only a mapping technique. However, the resulting mapping has provided an interesting 3D representation of the texts

samples, showing two separated writing styles and consequently two different authors. This result is easily observable in figure 17.4, where all Quran texts have been concentrated in the left region and all Hadith text have been concentrated in the right side.

- The fifth approach (i.e. PCA) is not a clustering method either, but it is only used for feature reduction and low dimensional mapping. The resulting 3D representation, in figure 17.5, shows the position of the text samples thanks to their 3 first PCA coordinates. We can observe that the Quran texts are located in the left, whereas the Hadith ones are located in the right and that the two text groups are quite separated. We also notice that Hadith samples are more condensed than Quran ones. The PCA representation suggests that the two books have two different author styles.

- The sixth approach, namely GMM, is a clustering technique based on mixture models. In our case it is used with only 2 components (2D representation). As we can see in figures 17.6.a and 17.6.b, two main GMM-based clusters have been obtained: Quran cluster at the left bottom (grouping all Quran segments) and Hadith cluster at the right top (grouping all Hadith segments). Consequently, and thanks to this 2D representation, the two books appear to belong to two different authors, or at least two different writing styles.

- The seventh approach (i.e. SOM) is a self organizing neural network, which makes a 2D representation of the different possible clusters in an interesting way, since it gives a rich amount of information regarding the clusters and their consistency.

  The resulting SOM mapping (figures 17.7.a and 17.7.b) shows a dark horizontal region separating the different SOM cells into two main regions: Top and Bottom. The bottom area contains Quran segments, while the top area contains Hadith segments. Furthermore, one can observe that Hadith is represented by a big sub-cluster in the left and a small one in the right (top area), whereas Quran is represented by a big sub-cluster in the right and a small one in the left (bottom area). However, in the overall, the two main clusters are well separated one from the other: Quran samples in the bottom and Hadith samples in the top, which implies that there are 2 different author styles: one author style common to all Quran texts and another author style common to all Hadith texts.

17.4 Discussion

The principal purpose of this investigation is to conduct some experiments of authorship discrimination concerning two religious books: the holy Quran and Hadith, in a visual analytics way. And as described in the beginning of this manuscript, seven different features are used to make a stylometric comparison between the two books: Author Related Pronouns (ARP), Father Based Surname (FBS), Discriminative Words (DisW), COST value, Word Length Frequency (WLF), Coordination Conjunction (CC) and Starting Coordination conjunction (SCC).

On the other hand, the task of comparison is ensured by seven clustering approaches based on visual analytics techniques. We recall those seven clustering approaches: Hierarchical Clustering (HIC), Fuzzy C-mean Clustering (FCM), K-mean Clustering (KMC), Sammon Mapping (SAM), Principal Component Analysis (PCA), Gaussian Mixtures Models (GMM) and Self Organizing Maps (SOM).

Furthermore, every clustering method is performed alone and the resulting clusters are commented regardless of the other classifiers results. We also recall that every book has been segmented into 25 several text segments (i.e. 14 for the Quran and 11 for the Hadith) and that there is no prior information on how could be the general configuration of the clusters (i.e. resulting clustering).

That is, knowing that there are two sets of texts: {Q1, Q2, …Q14} and {H1, H2, …H11}, which are extracted from the two different books: Quran and Hadith respectively, it is quite evident to get interesting information from the number of obtained clusters and the text segments contained within each cluster. For instance:

a) If we get only 1 cluster, this means that probably the different texts are written by the same author (i.e. one author);

b) Also, if we get several clusters, but some Quran texts are grouped with some Hadith ones in a same cluster, this means that some Quran texts were probably written by the Hadith author;

c) However, if two clusters appear in the graphical clustering area and all the Quran texts are grouped in one cluster and all the Hadith texts are grouped in another distinct cluster, this will implies that the two books (Quran and Hadith) are written by two different authors (two different styles).

That is, by exploring the results section and by observing all the clusters and the texts disposition in those clusters, we easily see that all the obtained results correspond to the third case (i.e. case c). In other words, all the clustering methods led to two distinct clusters: one cluster containing the Quran texts and another distinct cluster containing the Hadith texts, in a visual/graphical way.

The visual analytics approach has revealed a lot of information; since it does not only show the distinction between the author styles, but also shed light on how consistent was that distinction. And that consistency can be visually estimated thanks to the 3D or 2D separation distance between text samples. So, the new visualization diagram of the clustering techniques is quite interesting since it allows seeing how much a particular text sample is far from one author style or near to it.

Finally and statistically speaking, it appears that the two investigated books (Quran and Hadith) possess two different writing styles and should belong to two different authors, which means that the Quran could not be an invention of the Prophet.

18 Twelfth Series of Experiments: Authorship Discrimination based on Word Transition Probability

After several studies on the intrinsic characteristics of the authors' styles, especially in Arabic, we noticed a thorough link between the author's style and the use of specific word bigrams. Moreover, some bigrams were found to have some forward and backward probabilities that are specific to only one author (in a closed set).
That is, in this study, we propose the use of a new set of features based on the Forward Probability (FWP) and Backward Probability (BWP). This set of features and their normalized form (deeply described in this paper) is proposed and employed for the first time to the knowledge of the author.
So, it could be interesting to try employing this new set of features in a task of Author discrimination between the two religious books.

18.1 Probability Computation Procedure

In the following, we describe the required steps to compute the Forward and Backward probabilities FWP and BWP of the word bigrams.

Let us recall that, in our study, the term Word Bigram represents a couple of successive words as follows: [*Word1 Word2*], where *Word1* denotes the prefix (1st word) and *Word2* denotes the suffix (2nd word).

Now, let us take an example:

Suppose we are interested in the following Arabic bigram (السميع العليم), which is written from the right to the left and which is referred to by [*Word1 Word2*]. See figure 18.1.

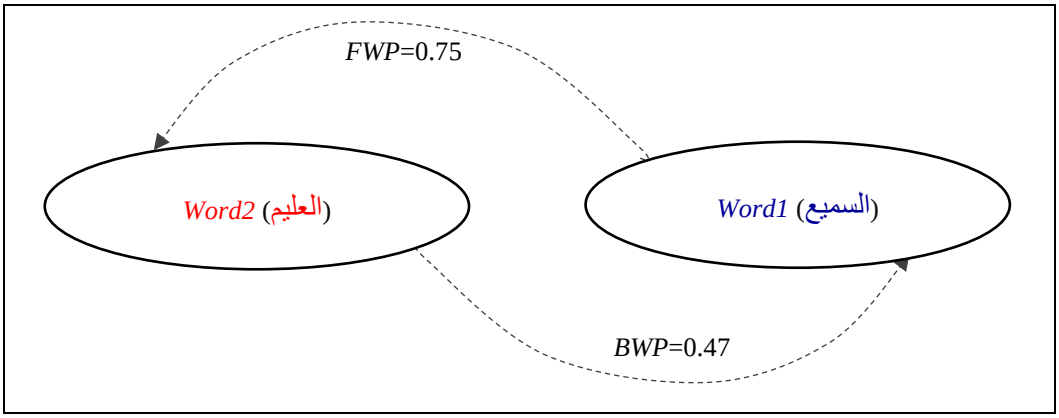


Figure 18.1: Graphical description of the FWP and BWP probabilities

The different steps of the procedure are:

- Compute the probability of *Word1* (e.g., السميع).
- Compute the probability of *Word2* (e.g., العليم).
- Compute the probability of the bigram [*Word1 Word2*] (e.g., السميع العليم).

- Finally, compute the forward and backward probabilities *FWP* and *BWP* using the equations 18.1 and 18.2.

In this context, we will denote the bigram probability by p(W_i), the prefix probability by p(P_i) and the suffix probability by p(S_i), so that $W_i = [P_i, S_i]$.

After computation of the different occurrences related to the previous example (see figure 18.1), we get the *Forward Probability* (denoted by F*WP*):

$$FWP = \frac{\mathrm{p}([P_j,S_j])}{\mathrm{p}(P_j)} \text{ (equal to 0.75 in this example)} \quad (18.1)$$

and, the *Backward Probability* (denoted by *BWP*):

$$BWP = \frac{\mathrm{p}([P_j,S_j])}{\mathrm{p}(S_j)} \text{ (equal to 0.47 in this example)} \quad (18.2)$$

18.2 Probability Normalization Procedure

In the context of transition probability, the sum of the *Forward Probability* and *Backward Probability* is not equal to one. In order to make the sum of the different probabilities equal to one, for a purpose of normalization, we have proposed the normalized forward and backward probabilities as follows:

$$NFWP = \frac{FWP}{FWP+BWP} \quad (18.3)$$

and

$$NBWP = \frac{BWP}{FWP+BWP} \quad (18.4)$$

where *NFWP* denotes the normalized *Forward Probability* and *NBWP* denotes the normalized *Backward Probability*.

In this context, $NFWP + NBWP = 1$ (18.5)

18.3 Selection of the bigrams

18.3.1 Case of limited set of bigrams

The number of the used word bigrams is very large, that is why we limited our selectin to only 20 bigrams. The selected word bigrams, which are investigated in this research work, are given below (*prefix and suffix are put in vertical disposition*).

Table 18.1 Used bigrams.

رسول الله	سبيل الله	الحمد لله	الا الله	الاسماء الحسنى	الْعَرْشِ الْعَظِيمِ	الحياة الدنيا	عزيز حكيم	لسَّمِيع الْعَلِيم	المسجد الحرام	غفور رحيم	كُنْ فَيَكُونُ
السماوات والارض	المشرق والمغرب	العزيز الحكيم	الْعَزِيزِ الْعَلِيمِ	شديد العقاب	الليل والنهار	ربك العظيم	موسى وهارون				

18.3.2 Case of unlimited set of bigrams

If one considers all the existing bigrams or at least those that appear at least twice in the text, it will be more interesting to do the classification by using Machine Learning tools, such as Linear Regression, Support Vector Machines, Neural Networks, etc. Though it is still possible to employ simple statistical distances such as Manhattan distance, Cosine distance or Spearman distance, efficiently.

18.4 Experimental Results

18.4.1 Experiments with limited bigrams

The results of the experiments, made with limited bigrams, are displayed in figure 18.2 for the Quran, and in figure 18.3 for the Hadith. A cumulative histogram representation is used in both figures, where the black color represents the forward probability and the orange color represents the backward probability. Note that the light gray color has been used to represent a null probability.

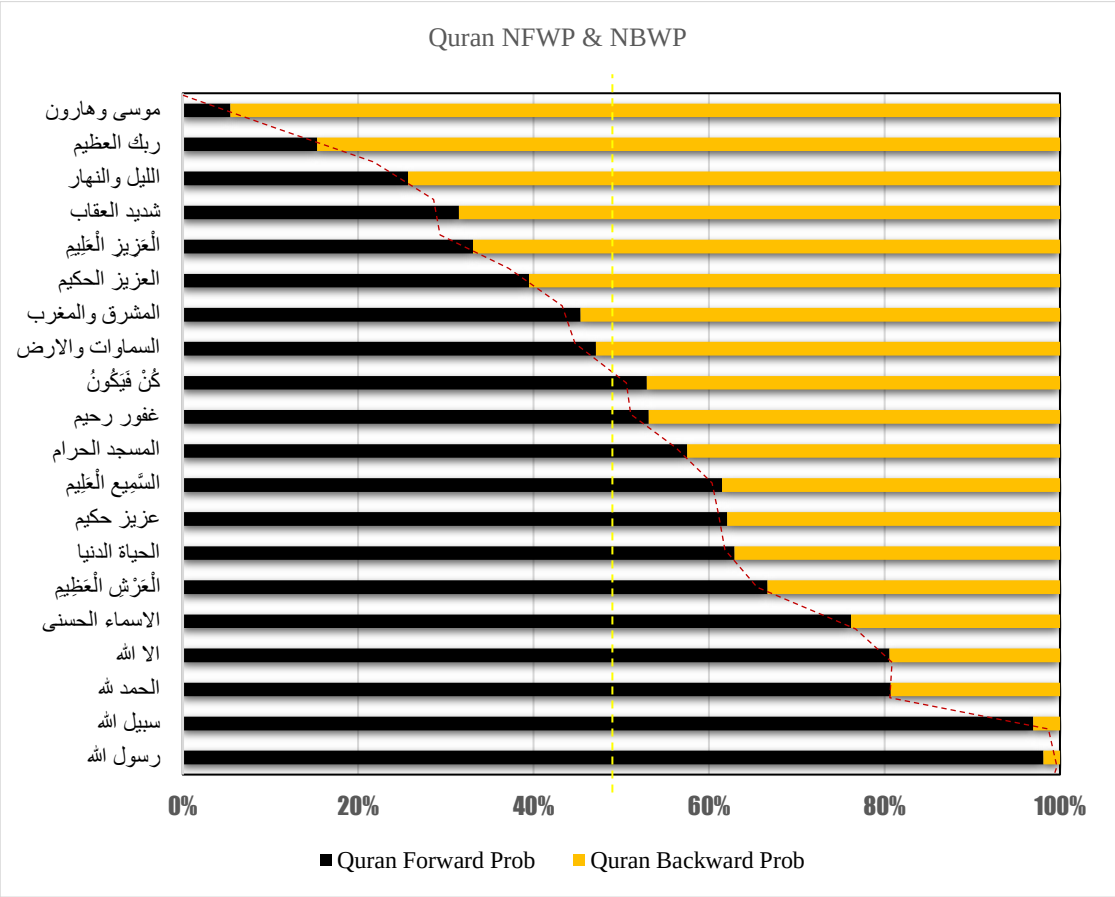


Figure 18.2: Graphical representation of the NFWP and NBWP probabilities in the Quran

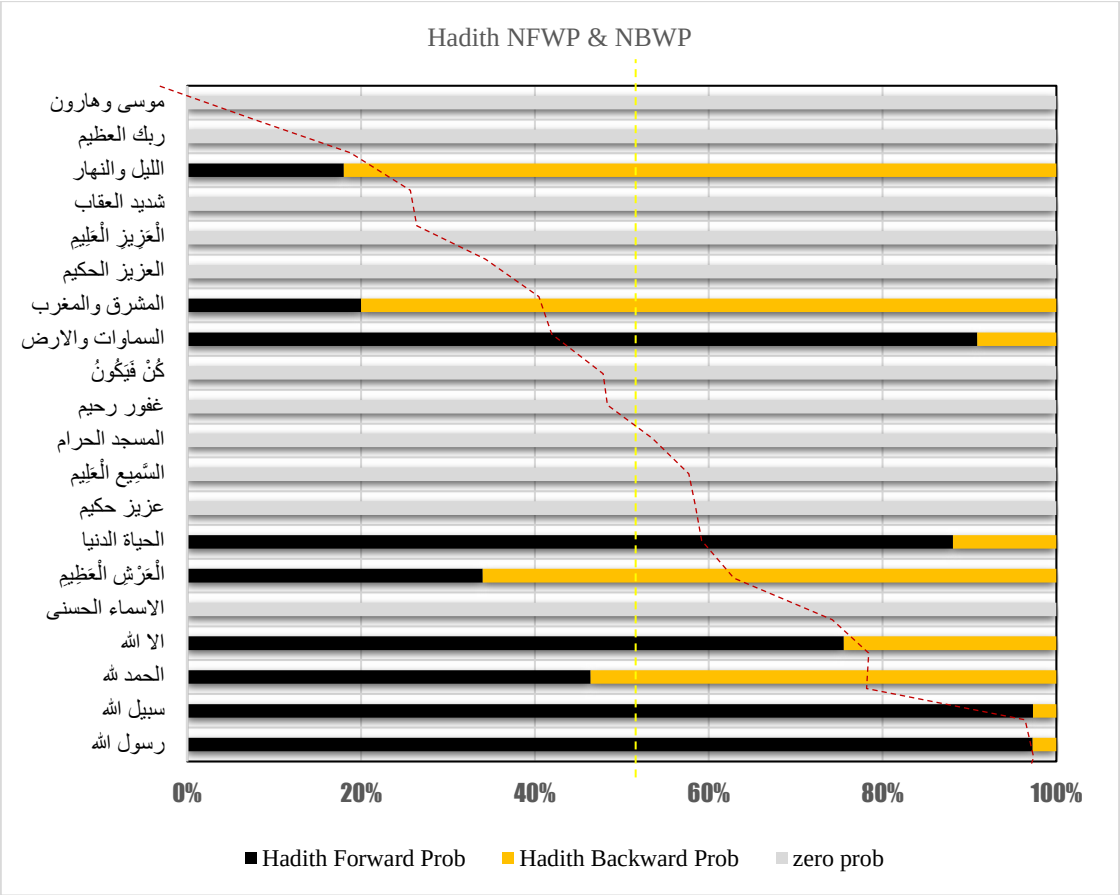


Figure 18.3: Graphical representation of the NFWP and NBWP probabilities in the Hadith. For the case of zero probabilities, we plotted a light gray bar charts. The red dashed line represents the shape of the intersection line between the NFWP and NBWP bars in the Quran.

According to figure 18.2, there are 4 bigrams from the Quran that have balanced normalized probabilities (i.e. NFWP ≅ NBWP), namely: غفور رحيم, كُنْ فَيَكُونُ, السماوات والأرض, المشرق والمغرب

While for the Hadith (see figure 18.3), only 1 bigram has a balanced normalized probability, namely: الحمد لله

As for the intersection line, which represents a borderline curve between the NFWP and NBWP bars (dashed red curve) in figure 18.3, we notice that the Quran's borderline is completely different from the Hadith's one, except for the 1st and 2nd bigrams (i.e. رسول الله, سبيل الله) and maybe also the 4th and 18th bigrams. Hence, among 20 analyzed bigrams, only 4 over 20 bigrams have almost similar NFWP and NBWP (i.e. 20% of the whole bigrams), while 16 among 20 bigrams, namely 80% of the whole bigrams, present normalized probabilities that are completely different between the 2 books.

18.4.2 Experiments with unlimited bigrams

In these experiments, the 37 text segments extracted from both the holy Quran and Hadith go through an LOO cross-validation technique, by using the centroid cosine distance for classification.
The obtained accuracy for this classification is 100% for each validation, and the medium cross-validation accuracy is 100% too, leading to a total separation between the two books and then the two Authors' styles.

18.5 Discussion and Conclusion

In this research work, a new approach of Authorship Attribution has been proposed. The new idea brought by this approach focuses on the probabilistic transitions within a word bigram.

Hence the different statistical parameters, namely: FWP, BWP, NFWP and NBWP, were computed and employed accordingly as features to perform the identification task and get an attribution decision.

The task of authorship discrimination between the two religious books was ensured, once by investigating the transition likelihoods of 20 different word bigrams, and another time by investigating the transition likelihoods of all the word bigrams present in the text.

The excellent results of both experiments show that this type of probabilistic transition feature is interesting by unveiling very specific values for every author. Furthermore, unlike classical features that may be topic dependent or common to several authors, the new proposed probabilistic transition features are not topic dependent.

Furthermore, a graphical assessment based on visual-analytics has been made thanks to the normalized parameters NFWP and NBWP (figures 18.2 and 18.3), where we visually noticed that the bar charts displayed for the 2 books are completely different (i.e. visual analytical assessment), and where 80% of the word bigrams presented different normalized probabilities.

Finally, according to the obtained results (i.e., 100% of correct discrimination), it appears that the two investigated books should possess two different Author's styles, which confirms once again the previous results stating that they should probably belong to two different Authors. Concerning the new proposed approach, we can say that it could be quite interesting in author identification provided that the documents size is large enough.

19 Thirteenth Series of Experiments: Authorship Discrimination based on Deep Learning Technology

In this new research work, we conduct an analysis of authorship discrimination between the Holy Quran and Hadith. As usually, the primary objective is to check whether the Quran and Hadith could have been authored by the same Author or not, while the second objective is to leverage the new Artificial Intelligence based approach (i.e., Deep Learning), to explore this religious enigma.
The global textual corpus, composed of the Holy Quran and a confident part of the Bukhari Hadith, is divided into segments of 500 words each.
The original aspect of this new work is the use of a deep neural network model based on Long Short-Term Memory (LSTM), in comparison to two Machine Learning based classifiers: Support Vector Machine (SVM) and Multilayer Perceptron (MLP).
The different results of authorship classification have shown that the Quran and Hadith were most likely authored by different Authors, as proved by the high classification accuracy: 100% with the LSTM, 99% with SVM and 99% with MLP.
Interestingly, the results of this study are in total concordance with the previous research works, reaffirming once again, that the holy Quran could not have been composed or invented by the Prophet, and which consequently supports the authenticity of the holy Book.

19.1 Dataset

The global corpus is derived from the complete Quran and a certified selection from the Bukhari Hadith. The Quran, in its entirety, contains over 87,341 tokens, while the selected subset of the Bukhari Hadith contains more than 23,068 tokens. The global dataset possesses a high degree of consistency, with the Quran averaging 315 A4 pages and the Hadith part averaging 87 pages.
Since the sizes are not very close between the two books, we have divided the texts into equal-sized segments. For this research work, we segmented the books into segments of 500 words each (figure 19.1). This segmentation produced a total of 222 text segments, distributed into 175 for the Quran and 47 for the Hadith.

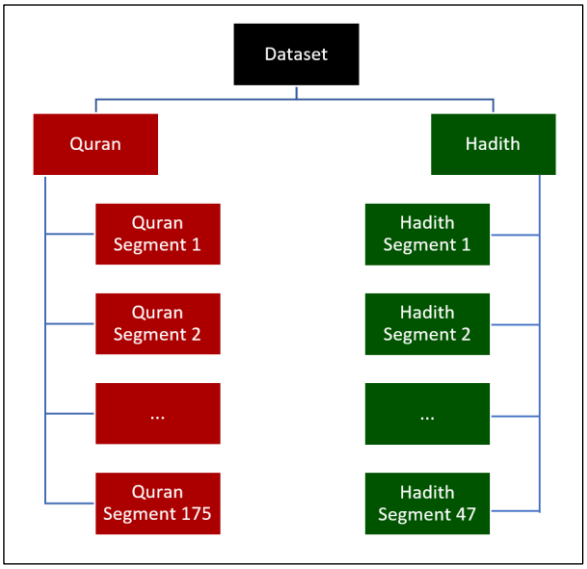


Figure 19.1. Book segmentation into text segments of 500 words each: 175 for the Quran and 47 for the Hadith.

To ensure the consistency of the results, a 5-fold cross-validation technique is employed. Every fold contains 222/5 ≈ 44 documents, while every training fold contains 222/5*4 = 177.6 5222×4 ≈ 178 documents. This technique of cross-validation is used to ensure the integrity of the results.

19.2 Proposed Model based on LSTM

The proposed Deep Learning model is a Sequential model with five layers, as depicted in figure 19.2. It is composed of the following layers.

- **Embedding Layer** (Wang, S., 2023), with an input dimension of 5000 and an output dimension of 64, handling input sequences of limited maximum length. This layer is used for word embedding where each word is represented as a 64-dimensional vector.
- **LSTM Layer** (Staudemeyer, R. C. 2023), with 64 units for processing temporal dependencies. The LSTM (Long Short-Term Memory) is a type of recurrent neural network that is effective in sequence prediction problems.
- **Dense Layer**, with 32 neurons using a ReLU activation for non-linear transformation.
- **Dropout Layer**, which helps to prevent overfitting by randomly setting the outgoing edges of hidden units to 0 at each update during training time, in order to mitigate overfitting.
- **Output Dense Layer**, with 2 neurons using Softmax activation for binary classification probabilities.

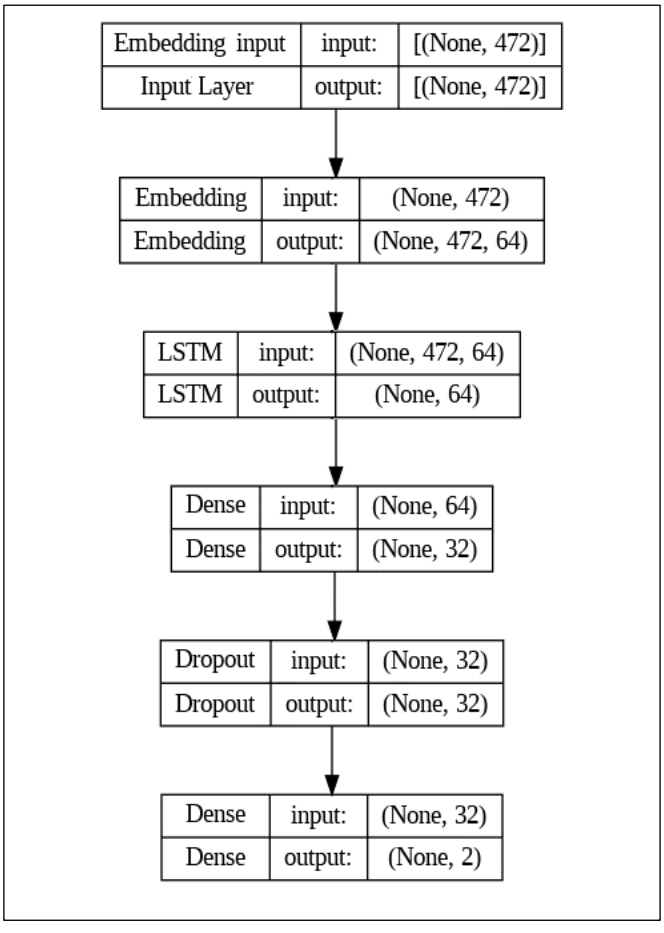


Figure 19.2. The proposed Deep Learning model with five layers.

19.3 Experimental Results

In this experiment, we used a deep neural network model based on Long Short-Term Memory (LSTM), without any specific feature extraction and by employing a 5 folds cross validation (figure 19.3).

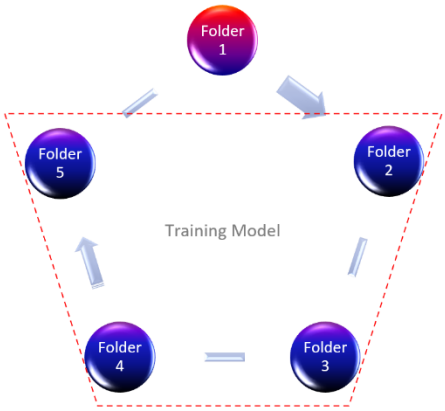


Figure 19.3. Set of the 5 folders representing the global dataset, with a 5 folds cross validation rotation.

Remarkably, the LSTM model achieved very high accuracy in all five folds of the cross-validation (see table 19.1), resulting in an average accuracy of 100%.

Table 19.1 Experiments of Authorship Classification based on LSTM with 5 folds cross validation, without any feature specification.

Fold	**Accuracy**
1st Fold	100%
2nd Fold	100%
3rd Fold	100%
4th Fold	100%
5th Fold	100%
Average 5-fold cross-validation	**100%**

A second experiment using a Support Vector Machine (SVM) in the same dataset, with character trigrams as features, provided an average accuracy of 99% across a 5-fold cross-validation. Similarly, a third experiment using a Multilayer Perceptron (MLP) in the same dataset, with character trigrams as features, also provided an average accuracy of 99% across a 5-fold cross-validation.

These results have shown that all the models are efficient in authorship discrimination between the Holy Quran and Hadith, while the LSTM high accuracy indicates its superior ability to capture the stylistic nuances of the books. Once again, these new results support the hypothesis that the Quran and Hadith should come from different authors.

19.4 Conclusion

In a challenging continuation of our previous work, for which the main objective was to check whether the Quran and Hadith could have been authored by the same Author or not, in this new investigation, we proposed and applied a deep neural network model based on Long Short-Term Memory in a task of automatic text classification per author style. This deep architecture was applied on quite small text segments of 500 words each, comparatively to Support Vector Machines and Multi-Layer Perceptron.

The Deep neural network (i.e., LSTM) has shown two advantages: firstly, it presented better performances than conventional machine learning classifiers; and secondly it did not require any feature at its input.

Thus, based on the high classification accuracy achieved by all the models, it can be concluded that the Quran and Hadith are very likely to have been authored by two different Authors. The LSTM model, in particular, achieved a very high accuracy of 100%, suggesting a distinct difference in the linguistic features and stylistic characteristics between the two books, even with small text segments.

This conclusion is well supported by the alignment of these results with our previous research works in this field. So, once again, the current findings support the authenticity and originality of the holy text, by refuting the skeptical assertion that the Prophet could have fabricated or dictated the holy Quran.

20 Book Analysis based on Embedded Scientific Knowledge

In the present section, we will cite some new scientific discoveries that were described in the holy Quran 14 centuries ago.

20.1 Scientists talking about the scientific aspect of the Quran

Here are some famous scientists who gave their appreciation on the authenticity of the holy book.

Prof. Jeffrey Lang, Department of Mathematics, Kansas University, USA

… We are merely comparing Qur'anic statements that deal with the physical universe and certain scientific notions. Often there appears to be profound similarities. But, more notably, as Bucaille observes, the Qur'an is distinguished from all other works of antiquity that describe or attempt to explain the workings of nature in that it avoids mistaken concepts.
…Although we may argue about the true meaning of any of these passages, the topics discussed thus far in this chapter lead to the definite impression that the author of the Qur'an anticipated an evolution in the mentality of man, culminating in an age when reason and science would be viewed as the final criterion of truth. (*Ref. J. Lang, Book: Struggling to Surrender, page 37*).

Prof. Arthur Alison, Department of Electrical and Electronic Engineering in the University of London, UK

"…During the conference on Medical Inimitability in the Qur'an, I could realize that the difference was great. Then I was convinced that Islam is the most proper religion that befits my inborn nature and conduct. In the heart of my hearts I had felt that there is a God controlling the Universe. He is the Creator"
"…Therefore, when I studied Islam, I found that it did not conflict with reason and science. I believe that is the revealed religion from the one and only God." (*Ref. First Islamic International Conference on the Medical Inimitability in the Quran, Cairo 1985)*

Prof. Zaghloul El-Naggar, Professor of Geology, Head, Committee on Scientific Nations in the GQPS, Cairo, Egypt.

"…However, it is strange to notice that every signal/sign present in the holy Quran, concerning the universe or one of its components, come in a scientific way and with an extreme precision, which proves that this holy book cannot be a human invention." (*Ref. Zaghloul El-Naggar, http://www.assabile.com/zaghloul-el-naggar-316/series/wa-yatafakkarun-316*).

Dr. Gary Miller, Canadian Assist. Professor in Mathematics and former Christian theologian, Toronto/Canada and KFUPM University Saudi Arabia.

“A truly scientific approach to the Qur’an is possible because the Qur’an offers something that is not offered by other religious scriptures, in particular, and other religions, in general. It is what scientists demand. Today there are many people who have ideas and theories about how the universe works. These people are all over the place, but the scientific community does not even bother to listen to them” (*Ref. www.irfi.org/articles/articles_551_600/scientific_approach_to_the_qur.htm*).

Prof. Gerald G. Goeringer, Georgetown University, Washington, DC, USA

" No such distinct and complete record of human development, such as classification, terminology, and description, existed previously...this description antedates by many centuries the recording of … traditional scientific literature." (*Ref. http://scienceislam.com/scientists_quran.php*).

Prof. William W. Hay, University of Colorado, Boulder, Colorado, USA

" I find it very interesting that this sort of information is in the ancient scriptures of the Holy Quran...I would think it must be [from] the divine being." (*Ref. http://scienceislam.com/scientists_quran.php*).

Prof. T. V. N. Persaud, University of Manitoba, Winnipeg, Manitoba, Canada

" Muhammad .. could not read, didn’t know to write. - You have someone illiterate making

profound pronouncements - amazingly accurate about scientific nature... this is a divine inspiration or revelation.." (*Ref. http://scienceislam.com/ scientists_quran.php*).

Prof. E. Marshall Johnson, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

" The Quran describes not only the development of external form, but emphasizes also the internal stages, the stages inside the embryo, of its creation and development, emphasizing major events recognized by contemporary science...I see nothing in conflict that ...divine intervention was involved." (*Ref. http://scienceislam.com/scientists_quran.php*).

Prof. Alfred Kroner, Institute of Geosciences, Johannes Gutenberg University, Mainz, Germany

" Thinking about many of these questions and thinking where Muhammad came from, he was after all a Bedouin. I think it is almost impossible that he could have known about things like the common origin of the universe,...Someone 1400 years ago could not know the heavens and the earth had the same origin.." (*Ref. http://scienceislam.com/scientists_quran.php*).

Prof. Keith Moore, University of Toronto, Ontario, Canada

"It is clear to me that these statements must have come to Muhammad from God or Allah, because most of this knowledge was not discovered until many centuries later.." (*Ref. http://scienceislam.com/scientists_quran.php*).

Prof. Joe Simpson, Baylor College of Medicine, Houston, Texas, USA

".. It follows, I think, that not only there is no conflict between genetics and religion but, in fact, religion can guide science by adding revelation to some of the traditional scientific approaches, that there exist statements in the Quran shown centuries later to be valid, which support knowledge in the Quran having been derived from God." (*Ref. http://scienceislam.com/ scientists_quran.php*).

Prof. Yoshihide Kozai, Tokyo University, Hongo, Tokyo, Japan

"I am very much impressed by finding true astronomical facts in Quran... So, by reading [the] Quran and by answering to the questions, I think I can find my future way for investigation of the universe." (*Ref. http://scienceislam.com/ scientists_quran.php*).

Prof. Tejatat Tejasen, Chiang Mai University, Chiang Mai, Thailand

" During the last three years, I became interested in the Quran.... From my study and what I have learned from this conference, I believe that everything that has been recorded in the Quran fourteen hundred years ago must be the truth, that can be proved by the scientific means. " (*Ref. http://scienceislam.com/scientists_quran.php*).

Dr Maurice Bucaille, French medical doctor, member of the French Society of Egyptology, France

"The above observation makes the hypothesis advanced by those who see Muhammad as the author of the Qur'an quite untenable. How could a man, from being illiterate, become the most important author, in terms of literary merit, in the whole of Arabic literature? How could he then pronounce truths of a scientific nature that no other human being could possibly have developed at the time, and all this without once making the slightest error in his pronouncements in the subject?
The ideas in this study are developed from a purely scientific point of view. They lead to the conclusion that it is inconceivable for a human being living in the seventh century A.D. to have made statements in the Qur'an on a great variety of subjects that do not belong to his period and for them to be in keeping with what was to be known only centuries later. For me there can be no human explanation to the Qur'an" (*Ref. Maurice Bucaille, Book: The Bible, The Qur'an & Science, page 91*).

20.2 Number of Months and Days in the Quran: "An Enigma"

In the Quran, it is stated that the number of months (per year) is 12, as quoted in the verse [9-36]:

"إن عدة الشهور عند الله اثنا عشر شهرا في كتاب الله يوم خلق السماوات والأرض"

Translation: ["The number of months in the sight of Allah is twelve (in a year)- so ordained by Him the day He created the heavens and the earth"].

But strangely, by counting the occurrence number of the word "**Month**" in the Quran text, we find exactly **12 occurrences** (cited 12 times in the Quran) – see figure 19.1.

Again, if we count the number of "**Day**" in the Quran text we find exactly **365 occurrences** (cited 365 times in the Quran), which is equal to the real number of days per year. Really, it is a fascinating enigma – see figure 19.2.

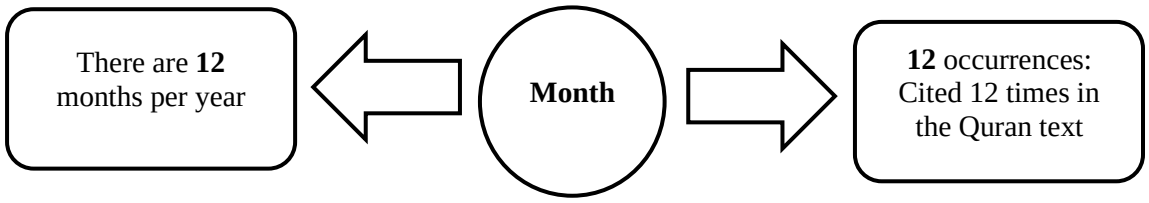


Figure 19.1 Strange concordance between the occurrences of the word Month and the real number of months per year.

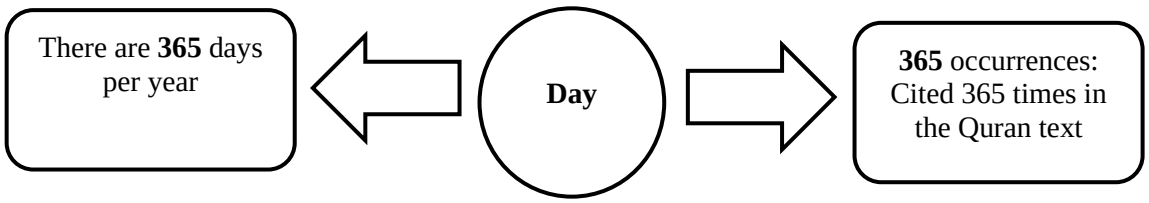


Figure 19.2 Strange concordance between the occurrences of the word Day and the real number of days per year.

Really, by observing these two strange coincidences, we do not find any word to say except the fact that the Divinity origin of the holy Quran is incontestable and evident.

20.3 Earth Rotation in the Quran

It is well-known that the first person who discovered that the earth rotates around its axis was Copernicus(16th century). Prior to that, scholars thought that the earth was immobile. However, when we read the holy Quran, we see that it clearly speaks about the earth movement. Hence, in the Ant verse (verse number 27), Allah says:

وَتَرَى الْجِبَالَ تَحْسَبُهَاجَامِدَةً وَهِيَ تَمُرُّ مَرَّ السَّحَابِ صُنْعَ اللَّهِ الَّذِي أَتْقَنَ كُلَّ شَيْءٍ إِنَّهُ خَبِيرٌ بِمَا تَفْعَلُونَ (النمل)

-Translation: "*You see the mountains and think that they are immobile and fixed in place while they are in movement like the clouds. This is the work of God who perfected all things, and verily He is aware of what you do.*" (27:88).

This verse speaks about the mountains movement although we all imagine them to be still and fixed at first impression. It also informs us that they are moving like the clouds (see, below, the NASA image showing the movement of the mountains/earth and a Real photo showing the movement of the clouds). Consequently, since the scientific community was mistaken before the 16th century, it is clear that this very ancient information represents a miracle that shows the divine origin of the holy book.

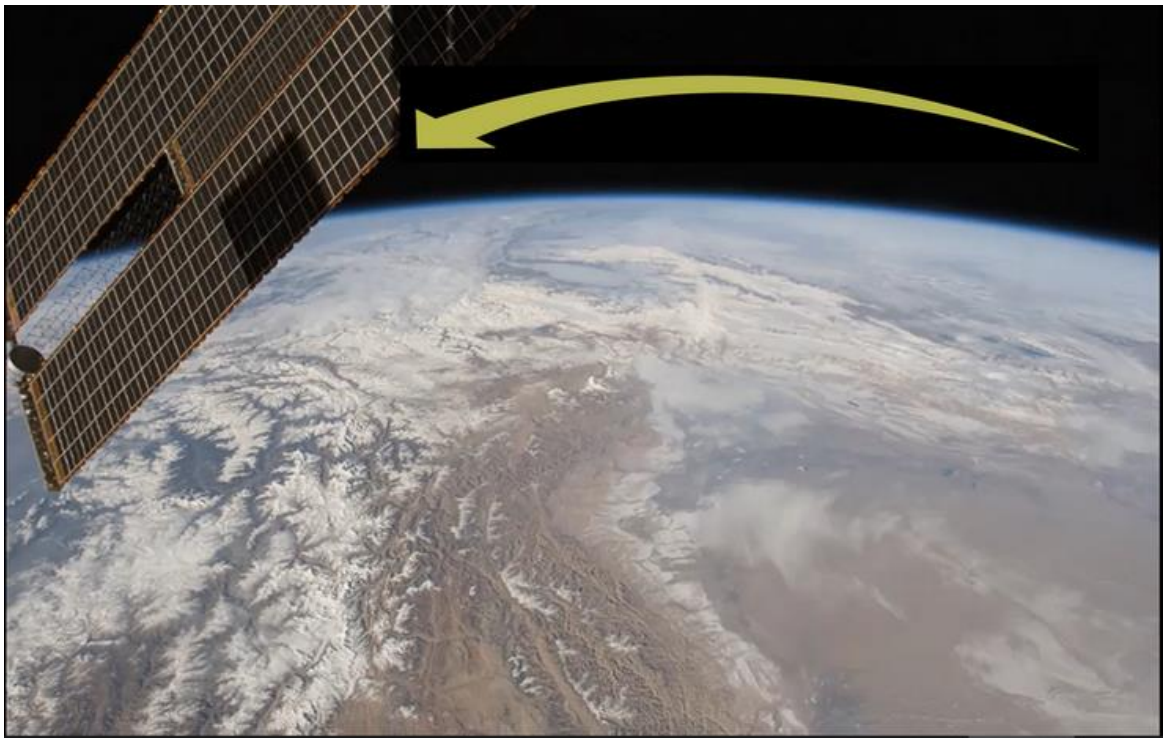


Figure 19.1 Mountains movement seen by satellite (NASA courtesy)

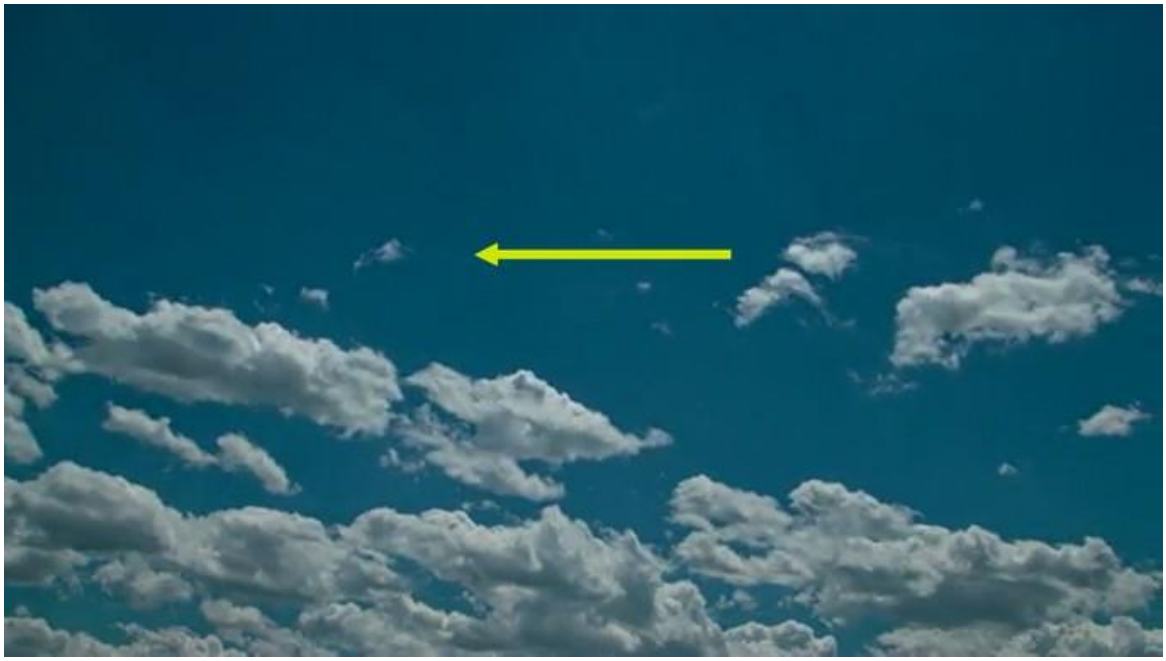


Figure 19.2 Clouds movement

20.4 Expansion of the Universe in the Quran

The discovery of Edwin Hubble stating that the Universe is expanding at enormous speed was revolutionary. He noted that galaxies were all moving away from us, each at a speed proportional to its distance from us. Today, several NASA spacecrafts continue Hubble's work of measuring the expansion of the Universe. So, the Universe expansion theory is proved scientifically: it is called BigBang (see figure 19.3), but the problem is that there are unknown forces at work we do not understand at all.

Some fascinating evidences on the expansion of the Universe are present in the holy Quran: see the following verse [51:47]:

والسماء بنيناها بأيد وإنا لموسعون

-Translation: "And it is We who have built the universe with (Our creative) power; and, verily, it is We who are steadily expanding it ".

In fact, this verse clearly states that the Universe is continuously expanding, which is in total confirmation and concordance with that new discovery. That is, how could the Prophet know that scientific knowledge 14 centuries ago?

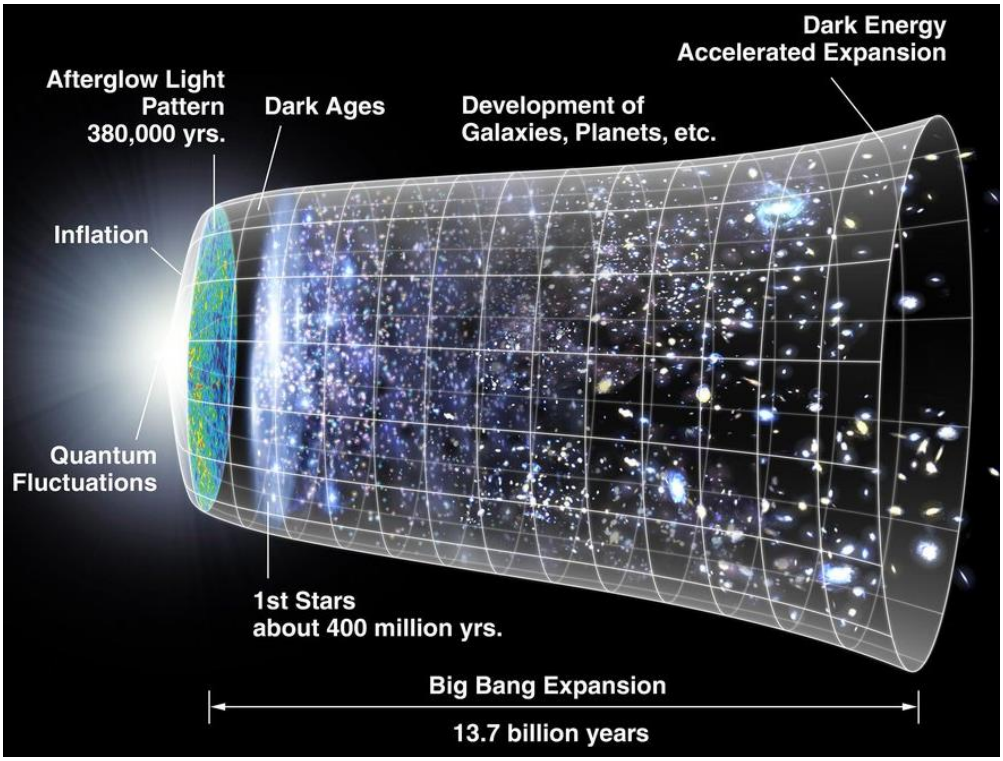


Figure 19.3 The Big Bang expansion – NASA courtesy.

20.5 A Scientific Evidence on the Sun Movement in the Holy Quran

Even though the scientific community thought that the Sun was the center of the Universe and that it was immobile until the 17th century, it has been clearly stated in the holy Quran, 14 centuries before, that it does really move in the Universe by respecting a precise orbit that was well defined (by the Creator); and in perfect concordance with the latest discoveries.

Hence, we can retrieve the related information in the following verse [36:38]:

(وَالشَّمْسُ تَجْرِي لِمُسْتَقَرٍّ لَّهَا ذَٰلِكَ تَقْدِيرُ الْعَزِيزِ الْعَلِيمِ)[36:38]

This verse can be translated into: *"And the Sun moves for the fixed/stable course (orbit) assigned for it. That is the decree/design of the Almighty, the All-Knowing" [36:38].*

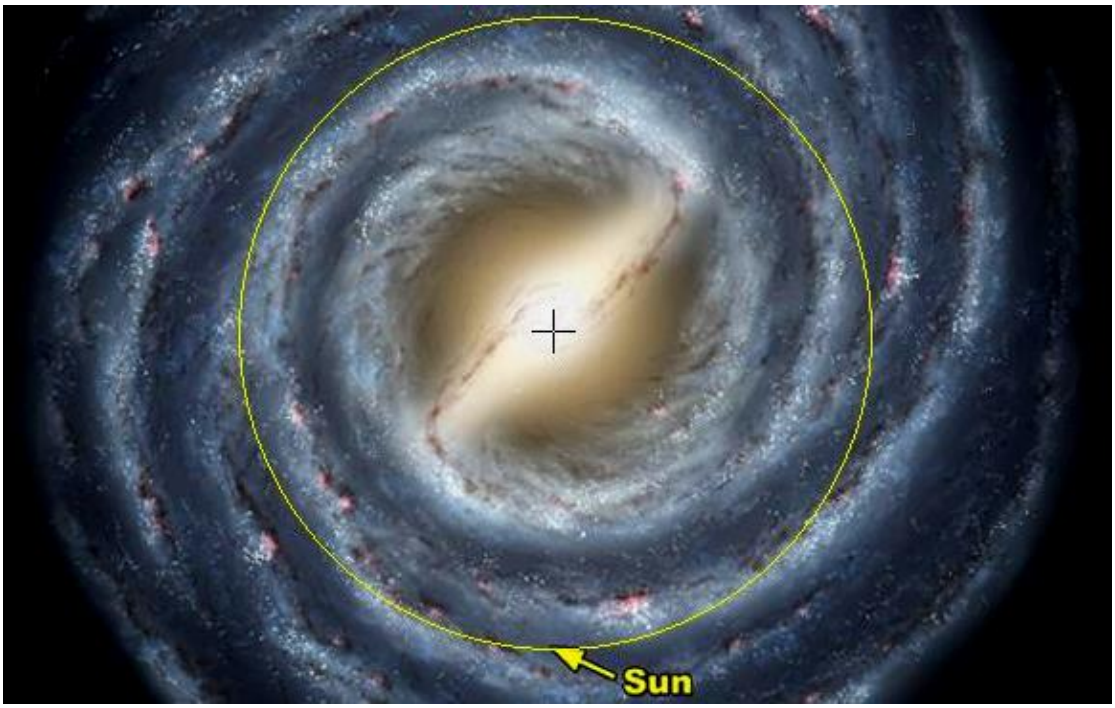


Figure 19.4: Image of the Milky Way. One can see the orbit of the Sun in yellow. NASA courtesy.

Before the 18th century the Sun has always been considered as the center of the Universe, and then considered to be at rest (i.e. fixed).

Thus, we can see the heliocentric model of Nicolaus Copernicus in his publication (On the Revolutions of the Celestial Spheres) in 1543, which was inspired from the assumptions of Aristarchus. His hypotheses are that the fixed stars and the Sun remain unmoved, and that the Earth revolves about the Sun in the circumference of a circle, the Sun lying in the middle of the orbit.

Thereafter, Johannes Kepler accepted that theory and published his first two laws about planetary motion in 1609. In the 17th century, the general idea that the sun was the center of the Universe and immobile was widely accepted.

Later on, Newton proposed a new heliocentric view of the solar system developed in a modern way (17~18th century). For Newton, it was not precisely the centre of the Sun that could be considered at rest, but rather the common centre of gravity of the solar system.

Recently (20~21th century), scientists have discovered that the Sun does orbit around the center of the Milky Way galaxy in a period of about 250 million years, and with a speed of 251 km/s. Moreover, the Milky Way itself is moving through the Universe within the local group of galaxies (cluster), so that the estimated global speed of the solar system is about 600km/s (Reference: http://antwrp.gsfc.nasa.gov/apod/ap050508.html) according to the latest scientific results of NASA.

However, 14 centuries before**,** in the holy Quran, it has been clearly stated that the Sun does really move in the Space by respecting a well-defined orbit. So, how could the Prophet give such precise scientific information in a period of time known by its limited scientific knowledge and instruments, beside the fact that the Prophet was illiterate?

20.6 About the Embryo description in the Quran

Dr. Keith Moore, Professor of Anatomy and Chairman of the Department, Faculty of Medicine, at the University of Toronto, wrote an article entitled "*Highlights of Human Embryology in the Koran and the Hadith*" in 1982.

Dr Moore pointed out that when he studied certain statements in the Quran on this subject, "*I was amazed at the scientific accuracy of these statements which were made in the 7th century CE.*"

In the Quran (39:6) it is well stated that God created us in the wombs of our mothers in stages. Dr. Moore says: "*The realization that the embryo develops in stages in the uterus was not discussed or illustrated until the 15th century* [https://www.muslimink.com].

Figure 19.5: The embryo stages. GIMS courtesy.

The staging of human embryos was not proposed until the 1940's, and the stages used nowadays were not adopted worldwide until a few years ago."

Furthermore, he comments: "*The idea that development results from a genetic plan contained in the chromosomes of the zygote was not discovered until the end of the 19th century. The verse from the Quran (80:18) clearly implies that the nutfah (i.e. the initial drop of fluid) contains the plan or blueprint for the future characteristics and features of the developing human being.*"

The following verse of the Quran (23:12-16) show that there is a gap or lag between two of the early stages of growth.

Strangely, Dr. Moore says: "*It is well-established that there is a lag or delay in the development of the embryo during the implantation. The agreement between the lag or gap in development mentioned in the Qur'an and the slow rate of change occurring during the second and third weeks is amazing. These details of human development were not described until about 40 years ago.*"

He then concludes by saying that the agreement he has found in the Quranic statements "*may help to close the gap between science and religion which has existed for so many years.*"

Once again, this scientific discovery shows that the Quran must be from God.

20.7 Description of the Pharaoh's death and the preservation of his body in the Quran

Figure 19.6 Pharaoh's body preserved after his death (Cairo Museum).

This is the body of Firawn (Rameses II), believed to be the Pharaoh in the time of Prophet Musa (Moses). His mummy is preserved and is currently on display in the Egyptian Museum, Cairo.

The following verses of the holy Quran say (10:90-92):

- “We helped the children of Israel cross the sea safely. The Pharaoh and his army pursued the children of Israel with wickedness and hate until the Pharaoh was drowned. As he was drowning the Pharaoh said, "I declare that there is no God but the One in whom the children of Israel believe and I have submitted to the Word of God" (Quran 10:90).
- (God replied), "Now you declare belief in Me! but before this you were a disobedient rebel (Quran 10:91).
- “We will save your body on this day so that you may become evidence (of Our existence) for the coming generations; many people are unaware of such evidence." (Quran 10:92).

When the Quran was transmitted to the humanity by the Prophet, the bodies of all the Pharaohs, who are probably related to the Exodus, were in their tombs of the Necropolis of Thebes, on the opposite side of the Nile from Luxor. At the time however, absolutely nothing was known of this fact, and it was not until the end of the 19th century that they were discovered (*Ref. The Bible, The Qur'an and Science, by Maurice Bucaille*).

As the Quran states in the verse 10:92 (*We will save your body on this day so that you may become evidence - of Our existence- for the coming generations*), the body of the Pharaoh of the Exodus was rescued, and visitors may see him in the Royal Mummies Room-of the Egyptian Museum.

Note that nothing was known at the time of the revelation of the Quran about the mummy of Rameses II. This fact and the different discoveries reported by Dr Maurice Bucaille confirm that the evidence given by the holy Quran (on the preservation of the Pharaoh’s mummy) presents a clear proof on its truth and its Divine origin.

21 About the "Last Prophet" Meaning and Prediction of a Prophet called Muhammad in the Ancient Religious Books

According to the holy Quran, it is well known that Muhammad is considered as the last Prophet sent by Allah (God). In this study we explore and debate that fact fourteen centuries later, by trying to find out what could be the consequences and conclusions to deduce. Moreover, we try to explore some ancient prophecies about the apparition of the Prophet Muhammad, which are reported in some ancient religious books such as the Bible and book of Habakkuk.

This study has shed light on the deep meaning of "Last Prophet" and revealed several holy texts on the prophecy of the Prophet Muhammad from the ancient religious books, which confirm the veracity and truthfulness of the holy Quran.

21.1 Last Prophet Concept and Truthfulness of the Holy Quran

As we know, in the Quran it is stated that Muhammad (Pbuh) is the messenger of God/Allah and that he is the last Prophet. In this survey, we will focus on the following verse (33:40):

(33:40) [مَا كَانَ مُحَمَّدٌ أَبَا أَحَدٍ مِنْ رِجَالِكُمْ وَلَكِنْ رَسُولَ اللهِ وَخَاتَمَ النَّبِيِّينَ وَكَانَ اللهُ بِكُلِّ شَيْءٍ عَلِيمًا]

Translation: "Muhammad is not the father of any man among you, but he is the Messenger of Allah and the last (end) of the Prophets. And Allah is Ever All-Aware of everything".

In this holy verse, Allah says that Muhammad is his Messenger and not the father of any man among his companions. He also added that he is the last prophet, which represents a very interesting information to explore and debate.

In fact, if another new religion, revealed by a new prophet, appeared after the death of Muhammad (Pbuh), then this verse would present a contradiction.

However, as we know, no new religion has been revealed after the Islam* and no Prophet has been reported except those who are well known, such as Jesus/Isa, Moses/Musa, Abraham/Ibrahim, etc. (see figure 1), and then the previous verse does represent a real proof on the truth of the holy Quran.

** The terme of religion refers to a serious revelation from God/Allah with a Divine holy book and which is effectively followed by a large population of people, such as Islam or Christianity for instance.*

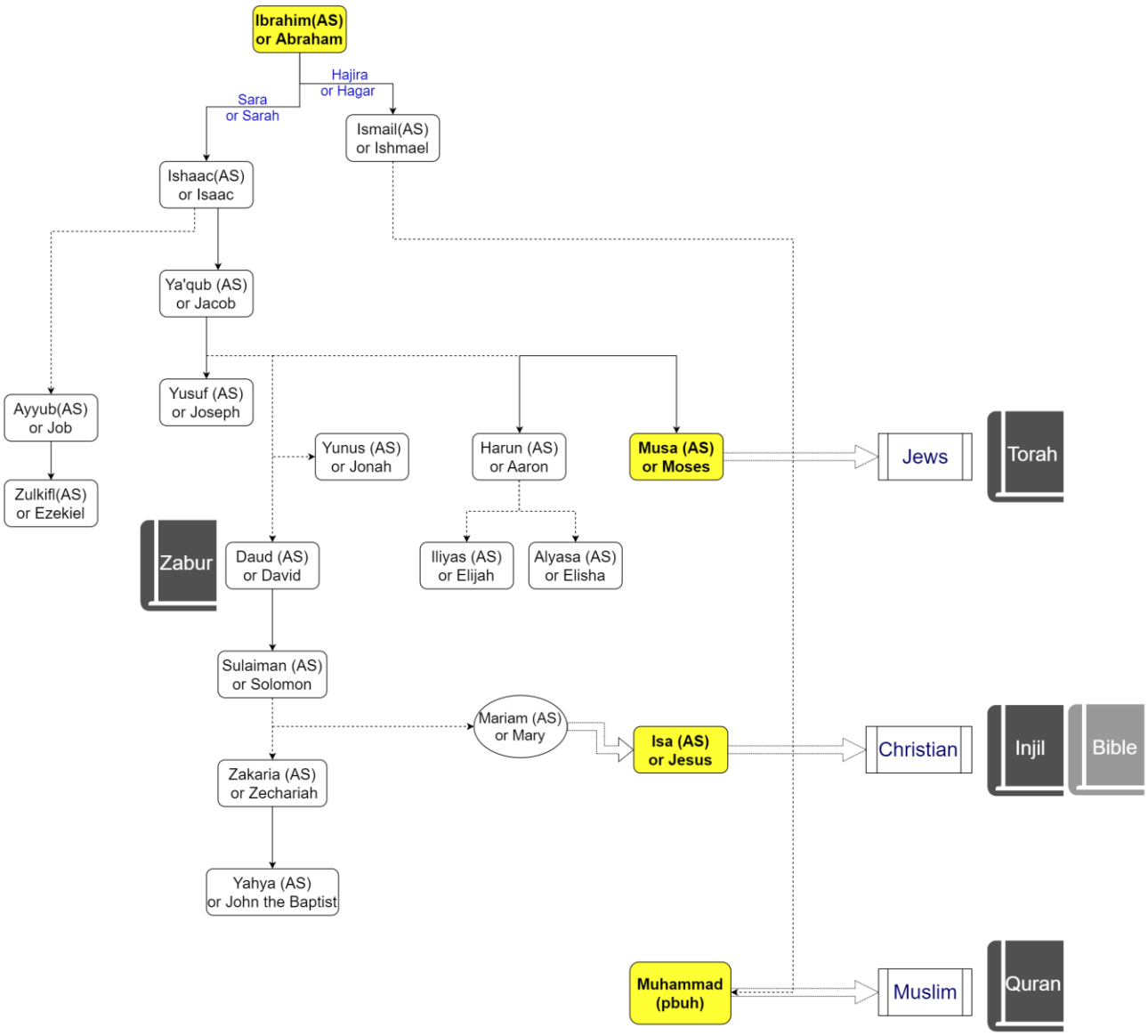


Figure 21.1: Prophets tree in accordance to Islam. *Photo from Islam-beliefs.net* (Islam-beliefs, 2023).

21.2 Other significations on the fact that there will be no more Prophets

The fact that Muhammad is the last Prophet and that there will be no other Prophet after him, involves the following important points:

- This Prophet is sent for all the humanity, as clarified by the Prophet (Pbuh);
- The holy Quran will be preserved against alterations since there will be no other Divine book after. This fact has also been cited in the Quran verse (15:9) "{إِنَّا نَحْنُ نَزَّلْنَا الذِّكْرَ وَإِنَّا لَهُ لَحَافِظُونَ}", which can be translated into "Indeed We have sent down the Quran, and indeed We Ourselves surely are its Guardians".

This is one of the miracles of the Quran, since no one has been able to change even one word of its text, and has remained in its original morphological form since the 6th Century (A. D).

- The end of the world is expected, as reported by the Prophet (Pbuh) "«بعثت أنا والساعة كهاتين»", which can be translated into "I have been sent and the Hour (end of the world) as these two (fingers)".
- There is no need to another new revelation or another new religion. This fact can be confirmed by observing the global religion expansion in the world, as displayed in figure 2, and where one can notice that Islam is the most growing religion in the world;

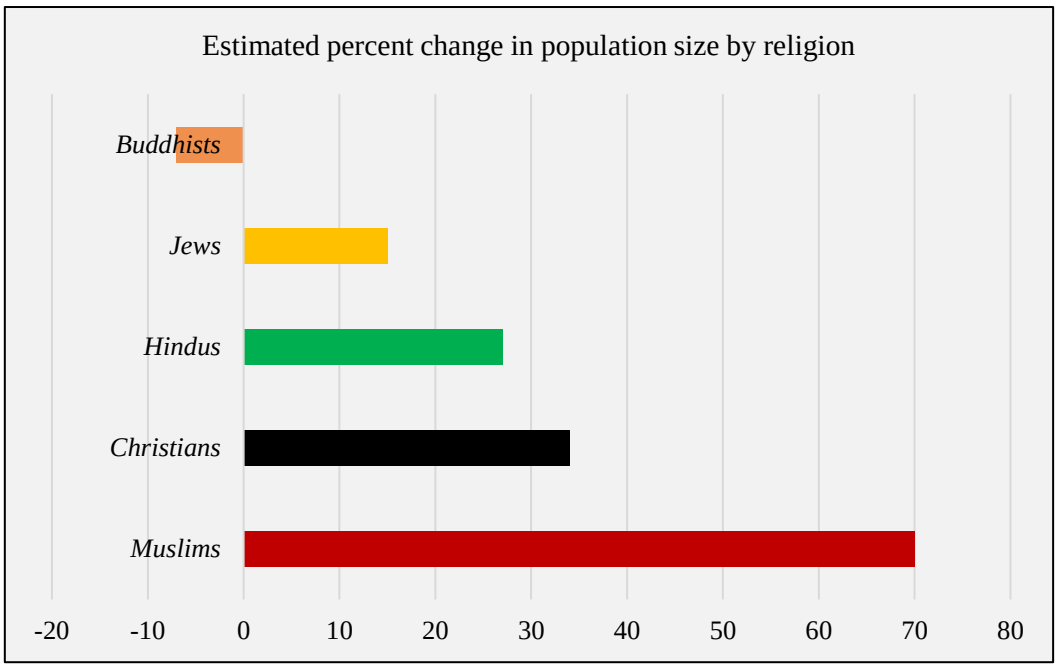


Figure 21.2: Estimated percent change in population size by religion. Source: *Pew Research center demographic projection*.

According to this study, one can conclude that a lot of information can be deduced from the fact that Muhammad is the last Prophet and that there will be no other Prophet after him, as discussed in this paper, but one of the most important points to note here is that this fact represents a real proof on the truthiness of the holy Quran (since no new religion has been revealed after the Islam).

21.3 Prediction of Muhammad in the Bible and ancient religious books

- **1st Prediction**

The Gospel of Barnabas is attributed to Barnabas and is considered as one of the disciples of Jesus. The total authenticity of this Gospel has not been established, but it appears quite interesting by explicitly predicting the Prophet Muhammad (Pbuh) in it. For example, the chapter 163 of the Gospel says:

"... Then Jesus said: 'So secret is predestination, 0 brethren, that I say to you, truly, only to one man shall it be clearly known. He it is whom the nations look for, to whom the secrets of God are so clear that, when he comes into the world, blessed shall they be that shall listen to his words, because God shall overshadow them with his mercy even as this palm-tree overshadows us. Yes, even as this tree protects us from the burning heat of the sun, even so the mercy of God will protect from Satan them that believe in that man.' The disciples answered, "0 Master, who shall that man be of whom you speak, who shall come into the world?" Jesus answered with joy of heart: 'He is Muhammad; Messenger of God, and when he comes into the world, even as the rain makes the earth to bear fruit

when for a long time it has not rained, even so shall he be occasion of good works among men, through the abundant mercy which he shall bring ..." (Quran Project, 2023).

According to this chapter, Jesus clearly foretells that there will be another Prophet coming after and his name is Muhammad, as well expressed in the previous paragraph (sentence in red).

- **2nd Prediction**
  Pr David Abdu Benjamin Keldani, BD, a Roman Catholic priest of the Uniate-Chaldean sect, wrote the book, 'Muhammad in the Bible.', where he cited one translated paragraph from the Gospel of John:

  "... and he shall give you another Parakletos/Periqlytos, that he may stay with you for ever" (John xiv. 16).

  Pr Keldani rewrote it in the following explanative form:

  "... and he shall send you another apostle whose name shall be Periqlytos, that he may remain with you for ever."

  In his book, he said: "There is not the slightest doubt that by "Periqlyte," Muhammad, i.e. Ahmad, is intended. The two names, one in Greek and the other in Arabic, have precisely the same signification, and both mean the "most Illustrious and Praised" (David Abdu Benjamin Keldani, 2007).

- **3rd Prediction**
  Deuteronomy 33:1-2 (fifth book of the Old Testament) combines references to Moses (Pbuh), Jesus (Pbuh), and Muhammad (Pbuh). It speaks about God's revelation coming from Sinai, rising from Seir (probably the village of Sair near Jerusalem) and shining forth from Paran. According to Genesis 21:21, the wilderness of Paran was the place where Ishmael (Pbuh) settled. In other words, it was in Arabia, and specifically in Makkah.
  Another sign of the prophet to come from Paran (Makkah) was that he would come with "ten thousands of holy ones." (Deuteronomy 33:2) That was the exact number of faithful who accompanied Muhammad (bpuh) to Makkah. The text says:

  "He shone forth from Mount Paran; he came from the ten thousands of holy ones, with flaming fire at his right hand."

  According to Islamic history, the city of Makkah (Paran) was liberated by Prophet Muhammad's 10,000 troops. If Muhammad (bpuh), who liberated the city of Paran with 10,000 believing, was not the one who fulfilled this Biblical prophecy, then who was that prophet? (Al-Rassi, M. S., 2019).

- **4th Prediction**
  Habakkuk 3:3 (known as book of "Habakkuk the prophet") speaks of God (God's help) coming from Teman (an oasis north of Madinah, according to Hastings' Dictionary of the Bible), and the "Holy One" coming from Paran. That "Holy One" who, under persecution, migrated from Paran (Makkah) to be received enthusiastically in Madinah was none other than Prophet Muhammad (bpuh) (Al-Rassi, M. S., 2019). See the Habakkuk 3:3 paragraph given below.

  "... and the Holy One from Mount Paran. Selah His glory covered the heavens, and His praise filled the earth.[3] His radiance was like the sunlight; rays flashed from His hand, where His power is hidden.[4]". [Habakkuk 3:3-4].

  So according to the admission of the Bible scholars and the Bible itself in Genesis 21:21 Paran is in Arabia, the land where prophet Ishmael settled. Then who is the Prophet who rose from Arabia? Isn't he Muhammed (Pbuh)? (Anthony Matthew Jacob, s. d.).

21.4 Conclusion

Throughout this study, we explored the significance of Muhammad being considered as the final prophet and delved into the philosophical and logical implications of this fact, where we can derive four key implications and conclusions that strengthen the Quran truthiness.

On the other hand, we tried to explore some ancient prophecies about the Prophet Muhammad apparition in the ancient holy books, where we can also find four amazing prophecies on the coming of another prophet from the Arabic peninsula who cannot be anyone other than Muhammad (Pbuh), which strengthen the Prophet truthiness too.

Consequently, we can conclude that all these facts represent another strong proof on the veracity and truthiness of the holy Quran and the Prophet Muhammad (Pbuh), leading to a wise and thorough reflection on the holy words of our Creator.

22 Does the Heart have a Control on Mind and Emotions? A Scientific Evidence Supporting what is Said in the Holy Quran

During centuries, human beings have been raising a key question on the exact role of the heart: is it concerned with the mind or is it only a blood pump? And in the latter option, is the brain the main part of the mind?

According to the holy Quran, the Heart has an important role in mind, thoughts and wisdom (See Alaaraf and Alhadj chapters). Moreover, during the ancient civilizations it was also admitted that the heart was strongly linked with feelings and thoughts.

However, in the last centuries, the unique definition of the heart has become only medical and limited to the function of blood pumping.

Fortunately, recent research works in the field of neuro-cardiology have led to very interesting discoveries by showing the great role of the heart in mind and feelings. These discoveries, which have completely changed the definition and role of the heart, represent a real scientific revolution.

22.1 What is the main source of the mind and emotions: the Heart, the Brain or both?

The famous question, concerning the actual location of the mind center or the feeling center, has not been solved yet. In the holy Quran, the heart plays a key role in human being behavior, while in ancient medical science it was supposed to be only the brain. So, is it the heart, the brain or something else?

In this context, *Rollin McCRaty*, Director of the HeartMath research center, said: **"**Most of us have been taught in school that the heart is constantly responding to "orders" sent by the brain in the form of neural signals. However, it is not as commonly known that the heart actually sends more signals to the brain than the brain sends to the heart! Moreover, these heart signals have a significant effect on brain function – influencing emotional processing as well as higher cognitive faculties such as attention, perception, memory, and problem-solving. In other words, not only does the heart respond to the brain, but the brain continuously responds to the heart**"** said the HeartMath Institute Research Director (HartMath, 2022).

Consequently, it appears that the ancient medical definition of the heart, supposed to be a simple blood pump, is misleading. In fact, several research experiments showed that the heart is also responsible for many emotional functions.

In fact, for many years scientists studied heart from the physiological side and they considered it only a blood pumping machine, but starting from the twenty one century and because of the high development in heart transplantation and artificial heart surgeries, researchers started to notice a strange phenomenon which is a change in the patient's psychological status after heart transplantation, these psychological changes are very deep to the extent that after changing the old heart these changes may affect his personal believes (Al-Kaheel A., 2022).

Moreover, according to recent research works, in the new field of neurocardiology, scientists have discovered that the heart possesses its own nervous system (Armour, J. A., 2007) : a nervous network so sophisticated as to earn the description of an intelligent heart-brain. With mover than 40,000 neurons, this heart-brain gives the heart the ability to sense, process information, make decisions, and even to show a type of learning or memory.

Again, literature in this research field has shown that the heart communicates with the brain in several ways: through nerve impulses, via hormones, through pulse waves, and through electromagnetic fields (McCraty, R., 2015).

According to Fredrickson, the pattern of the heart's rhythmic activity and the corresponding pattern of cardiac afferent neural signals transmitted to the brain can facilitate or inhibit higher cognitive functions. In other words, during emotional stress, when the heart's rhythm is irregular or incoherent, this discordance communicates some signals to the brain that result in the inhibition of higher brain processes related to perception, reasoning, and creativity (Fredrickson BL, Branigan C., 2005).

Also, according to Fredrickson, this fact explains why we often cannot think clearly, make careless mistakes, and have little access to our creativity under stress situation. Hence, these negative emotional states tend to produce a rigid and limited patterns of thought and action by reducing the possibility to make wise judgments (Fredrickson BL, Branigan C., 2005).

Biochemically speaking, the heart manufactures and secretes oxytocin hormone, which is involved in cognition, tolerance, trust, etc. For instance, some previous research works reported that the rat heart is a site of oxytocin synthesis and release, since this hormone was detected in the four chambers of the rat heart (Jankowski, M., Hajjar, F., Kawas, S. A., Mukaddam-Daher, S., Hoffman, G., McCann, S. M., & Gutkowska, J., 1998).

Moreover, the heart produces and secretes several other hormones, such as atrial peptide or atrial natriuretic peptide, which inhibits the release of stress hormones and influences our motivation and behaviour (McCraty R, Atkinson M, Tomasino D, Bradley RT., 2009).

One of the important conclusions reported by researchers in this field (HeartMath LLC. www.heartmath.com) is as follows: "Although there is much yet to be understood, it appears that the age-old associations of the heart with thought, feeling, and insight may indeed have a basis in science".

22.2 Heart and Brain citation in the Quran

The term "**قلب**", meaning heart in English, in its singular or plural form, was cited 132 times in the holy Quran, while the brain was not cited in the holy Quran (to the knowledge of the author) except the term head in its singular or plural form.

Now, when one reads the holy Quran, we strangely discover that the heart has an important role in feeling and wisdom, as one can see in the following verses:

a. ***First verse (chapter 7, verse 179):***

وَلَقَدْ ذَرَأْنَا لِجَهَنَّمَ كَثِيراً مِنَ الْجِنِّ وَالْإِنْسِ لَهُمْ قُلُوبٌ لا يَفْقَهُونَ بِهَا... {لأعراف: من الآية179}

Translation (chapter 7, verse 179): [Certainly We have winnowed out for hell many of the jinn and humans: they have hearts with which they do not understand…].

b. ***Second verse (chapter 22, verse 46):***

أَفَلَمْ يَسِيرُوا فِي الْأَرْضِ فَتَكُونَ لَهُمْ قُلُوبٌ يَعْقِلُونَ بِهَا أَوْ آذَانٌ يَسْمَعُونَ بِهَا فَإِنَّهَا لا تَعْمَى الْأَبْصَارُ وَلَكِنْ تَعْمَى الْقُلُوبُ الَّتِي فِي الصُّدُورِ. {الحج:46}

Translation (chapter 22, verse 46: [Have they not travelled through the land so that they may have hearts by which they may exercise their reason, or ears by which they may hear? Indeed, it is not the eyes that turn blind, but it is the hearts in the breasts that turn blind!].

If we observe the first verse, we notice an important phrase: "they have hearts with which they do not understand", which involves that the heart plays an important role in faith.

Now, what is meant by Heart in the holy Quran? Is it only an abstract concept without any physical existence (i.e., representing the mind)? Or is it the real heart organ that is concerned with the cardiovascular system?

To respond to that question, let us observe the second verse (22:46). By reading the first sentence of this verse, we notice an important sentence: "they may have hearts by which they may exercise their reason", which shows that the heart is a key point of wisdom. However, by reading the second part of the verse, we find the following sentence: "it is the hearts in the breasts that turn blind", which shows that the meaning of the term "heart" is probably the well-known "heart organ" that is inside the breast. This Quranic clarification gives an interesting response to the previous question, showing that the mind and belief are linked to the heart.

Now, by comparing the Quran knowledge on the heart and the new neuro-cardiological discoveries, we observe a great compatibility. Hence, once again, recent scientific discoveries come together strengthening the knowledge embedded in the Quran and which was sent down 14 centuries before. Probably, we will continue to discover more and more amazing concordances between the Quran and future research discoveries.

22.3 Conclusion

Even though the relationship between the heart, brain and mind is still an enigma, which is not completely solved or even understood, it appears that the Quran shed light on a lot of questions that were mistakenly answered by some ancient medical scientists.

The recent research works in neurocardiology have led to very interesting results reinforcing the Quran theory by showing the important role of the heart in mind and feelings. This discovery, which has changed the interpretation of the heart role, represents a real scientific revolution.

That is, it appears that the new scientific discovery about the important role of the heart in feelings, wisdom and reasoning are confirming several Quran verses, which give to the heart a key importance in belief and decision making.

So, once again and according to the recent important discoveries on this field, it is evident that this ancient book, dating from the 7th century, could not be a human invention and one cannot find any explanation on how that knowledge was embedded in the Quran except by considering a Divine intervention.

23 Do Animals communicate with each other? A Scientific Evidence Supporting what was Revealed in the Quran

In this survey, we analyze the recent discoveries on animal communication, such as birds, dolphins, ants and honeybees communication. We also try to see the Quran point of view on the matter by exposing some pertinent verses reporting a speech/communication related to animals.

The recent works cited in this paper affirm that it does exist a real way of communication between animals, confirming the information revealed by the holy Quran on the subject.

The cited research works and related results, not only confirm the information embedded in the Quran, but also lead to an important conclusion about the Divinity of this noble book.

23.1 Animal communication in the Quran

The Quran mentions the language of birds. For instance, in the verse (27:16), Allah (*The most Gracious the most Merciful*) says:

وَوَرِثَ سُلَيْمَانُ دَاوُودَ وَقَالَ يَا أَيُّهَا النَّاسُ عُلِّمْنَا مَنطِقَ الطَّيْرِ وَأُوتِينَا مِن كُلِّ شَيْءٍ إِنَّ هَٰذَا لَهُوَ الْفَضْلُ الْمُبِينُ [27:16]

Translation: And Solomon (Sulaiman) inherited David. He said, "O people, we have been taught the **language of birds**, and we have been given from all things. Indeed, this is evident bounty." (Quran 27:16).

The Quran clearly mentions the language of birds, where it can be noticed that the birds speak with each other. Some birds like parrots can learn human words. Moreover, some humans, for example villagers of Kusköy in Turkey are used to communicate with each other by whistles like birds. However, the strange ability hold by the Prophet Sulaiman to have meaningful conversations with birds (see verse (27:16)) is amazing (Illias S, s. d.).

Here is a part of discussion between Suleiman (Solomon) and the Hoopoe bird:

فَمَكَثَ غَيْرَ بَعِيدٍ * لَأُعَذِّبَنَّهُ عَذَابًا شَدِيدًا أَوْ لَأَاذْبَحَنَّهُ أَوْ لَيَأْتِيَنِّي بِسُلْطَانٍ مُّبِينٍ *وَتَفَقَّدَ الطَّيْرَ فَقَالَ مَا لِيَ لَا أَرَى الْهُدْهُدَ أَمْ كَانَ مِنَ الْغَائِبِينَ فَقَالَ أَحَطتُ بِمَا لَمْ تُحِطْ بِهِ وَجِئْتُكَ مِن سَبَإٍ بِنَبَإٍ يَقِينٍ(27:20-22)

Translation: (Solomon) inspected the birds and said, "How is it that I cannot see the hoopoe. Is he absent? I shall certainly punish him severely or slaughter him unless he has a good reason for his absence." Not long after, the hoopoe came forward and said, "I have information which you do not have. I have come from the land of Sheba with a true report. (Quran 27:20-22).

In these verses, we notice a discussion between the Prophet Sulaiman and the hoopoe bird, where the bird not only understand what was said by Sulaiman, but also responds to him too.

Concerning the Prophet Dawud (David), we find the mention of the birds too:

وَلَقَدْ آتَيْنَا دَاوُودَ مِنَّا فَضْلًا يَا جِبَالُ أَوِّبِي مَعَهُ وَالطَّيْرَ وَأَلَنَّا لَهُ الْحَدِيدَ (34:10)

Translation: And We certainly gave David from Us bounty. [We said], "O mountains, repeat [Our] praises with him, **and the birds** [as well]." And We made pliable for him iron (Quran 34:10).

More surprisingly Prophet Suleiman (Salomon) was even able to listen to ants communicating with one another, as cited in the verse (27:17-19):

وَحُشِرَ لِسُلَيْمَانَ جُنُودُهُ مِنَ الْجِنِّ وَالإِنسِ وَالطَّيْرِ فَهُمْ يُوزَعُونَ * حَتَّى إِذَا أَتَوْا عَلَى وَادِي النَّمْلِ قَالَتْ نَمْلَةٌ يَا أَيُّهَا النَّمْلُ ادْخُلُوا مَسَاكِنَكُمْ لا يَحْطِمَنَّكُمْ سُلَيْمَانُ وَجُنُودُهُ وَهُمْ لا يَشْعُرُونَ * فَتَبَسَّمَ ضَاحِكًا مِّن قَوْلِهَا وَقَالَ رَبِّ أَوْزِعْنِي أَنْ أَشْكُرَ نِعْمَتَكَ الَّتِي أَنْعَمْتَ عَلَيَّ وَعَلَى وَالِدَيَّ وَأَنْ أَعْمَلَ صَالِحًا تَرْضَاهُ وَأَدْخِلْنِي بِرَحْمَتِكَ فِي عِبَادِكَ الصَّالِحِينَ (27:17-19)

Translation: Solomon's army, consisting of human beings, jinn, and birds were gathered together in his presence in ranks. Until, when they came upon the Valley of Ants, **an ant said**, "O ants! Go into your nests, lest Solomon and his troops crush you without noticing". So [Solomon] **smiled, amused at her speech**, and said, "My Lord, enable me to be grateful for Your favor which You have bestowed upon me and upon my parents and to do righteousness of which You approve. And admit me by Your mercy into [the ranks of] Your righteous servants." (Quran 27:17-19)

In the previous verse, Allah (*The most Gracious the most Merciful*) reports the scenario between Suleiman and the ant, which was calling her community to enter their habitations to avoid being crushed by the Prophet and his army. The Prophet was able to understand its speech, which proves that ants have a developed means of inter-communication.

According to the holy Quran, it is stated that every creature is organized in communities, as it is the case with human beings. This fact is clearly mentioned in the verse (6:38)

وَمَا مِن دَابَّةٍ فِي الْأَرْضِ وَلَا طَائِرٍ يَطِيرُ بِجَنَاحَيْهِ إِلَّا أُمَمٌ أَمْثَالُكُم مَّا فَرَّطْنَا فِي الْكِتَابِ مِن شَيْءٍ ثُمَّ إِلَىٰ رَبِّهِمْ يُحْشَرُونَ

Translation: All the beasts on land and flying birds have different communities, just as you (people) do. Nothing is left without a mention in the Book. They will all be brought into the presence of their Lord (Quran 6:38)

23.2 Scientific discoveries about animal communication

Looking for animal behavior is an exciting experience, which shows how little we know about the many species that exist on earth and how little we know on their mean of communication.

We are still discovering more about the strange and high capabilities that animals possess to communicate each other.

In the songs and rituals they perform, one can detect a real meaning, which represents a real motivation in studying animal communication and facts that sustain the most enduring inquiries (Rogers L. J. & Kaplan G., s. d.).

As mentioned in (Rogers L. J. & Kaplan G., s. d.), most communications occur between members of the same species (intraspecies communication), but there are cases where one species communicates with another species (interspecies communication). (Rogers L. J. & Kaplan G., s. d.). We can quote for instance a communication between a furious dog and a female cat trying to protect its babies.

Scientists from different research fields might agree that the study of animal communication and the discovery of commonalities of some features of communication between animals, raise the possibility of considering the human language as only one of the different existing alternative modes of communication (Rogers L. J. & Kaplan G., s. d.).

23.2.1 Dolphins communication

Dolphins employ a highly developed acoustic communication system that uses many pertinent features associated with the term language (Fulton, J. T., s. d.).

As illustrated by Dr. Denise Herzing, Founder and Research Director of the Wild Dolphin Project [www.wilddolphinproject.org], dolphin communication employs the same mechanisms and signaling parameters as human speech except for the frequency range of 3900 and medium used (Fulton, J. T., s. d.).

Moreover, Dr Herzing asserts that the dolphin language presents a high entropy (suggesting that it could even approach some human languages), but without identifying any word in that language (Fulton, J. T., s. d.).

According to Lammers et al. (Lammers, M. O., Au, W. W., & Herzing, D. L., s. d.), if dolphins pay attention to the whistles structure with an important associated social role, then the evidence presented here indicates that there is considerably more to the social acoustic signaling behavior of some species of dolphins than meets the human ear (Lammers, M. O., Au, W. W., & Herzing, D. L., s. d.).

In figure 23.1 below, one can see a Markov model tree representing the probabilistic sequences of two dolphin whistles. Numbers in boxes represent whistle types. Percentages and direction of arrows shown represent the probability of one whistle type following a second whistle type. A curved arrow indicates the probability that a whistle of one type immediately follows itself. This figure shows a certain probabilistic structure between whistle types.

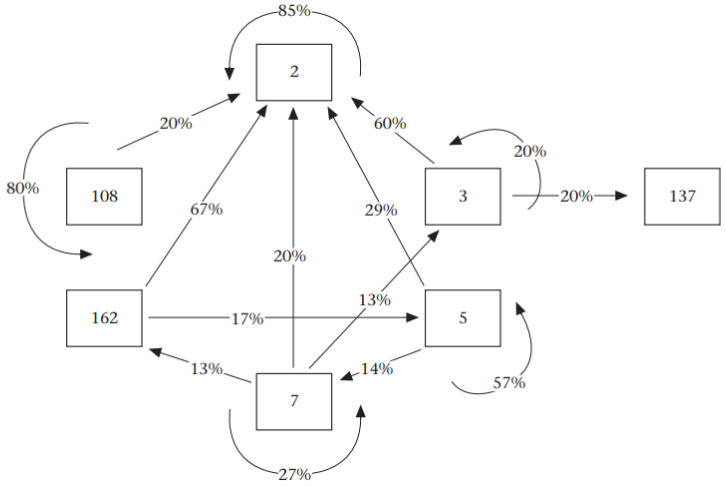


Figure 23.1 (McCowan, B., Hanser, S. F., & Doyle, L. R., s. d.): One set of two-whistle sequences shown as a probability tree based on a Markovian first-order (i.e. Shannon second-order entropy) analysis. Numbers in boxes represent whistle types. The number of whistles for each whistle type (WT) included in the diagram were: VVT2=188, VVT7=15, WT162=12, WT5=7, VVT108=5, WT3=5, WT137=1 (McCowan, B., Hanser, S. F., & Doyle, L. R., s. d.).

According to Ferrer-i-Cancho et al., it appears that the statistical properties of the mapping of dolphins whistle into meanings is really consistent with the hypothesis that dolphins whistles have some sort of meaning and that dolphins can communicate through them: the use of a specific whistle type is constrained by the behavioral context, where it can be shown that these constraints are sometimes shared by several individuals (Ferrer-i-Cancho, R., & McCowan, B., s. d.).

23.2.2 Bird communication

Birds produce a large variety of sounds, from high-frequency whistles to simple hooting. All of these sounds are used to communicate an information to other members of the same species. However, birds have another way of communication: birds can also communicate with their bodies through movement or color for example (Birdfact, s. d.).

Communication between individuals of any species, especially birds, relies on sending and receiving information in a format that can be understood by both parties. Moreover, the sense of hearing is highly developed in birds, so it is unsurprising that they use the sound signal to communicate with each other (Birdfact, s. d.).

Some investigations and analyses that were made on some animal sounds and their studies have provided evidence for syntax-like structures in their communication systems. In linguistics, syntax represents the rule of combination of meaningful sounds to form higher-order structures like phrases or sentences (Spiess, S., Mylne, H. K., Engesser, S., Mine, J. G., O'Neill, L. G., Russell, A. F., & Townsend, S. W., s. d.).

Furthermore, vocal communication in birds can roughly be divided into songs and calls. Although, these two terms are often used for the same meaning, scientists separate them depending on their real message and targeted function (Birdfact, s. d.).

An example of human-bird communication can be seen in the following video representing an interesting human-hawk example of communication: Can I talk to my Hawk?

https://www.youtube.com/watch?v=9mWzwWY1TuM (Mercer D., s. d.).

23.2.3 Ant communication

As other animal species, ants also do communicate between them by forming a strong community collaboration. We do not exactly know all their means of communication, but it is possible that they could communicate with different means and in different ways.

Recent research works have shown that ants can communicate via large arrays of pheromones and possess complex olfactory systems, with antennal lobes in the brain, with up to 500 glomeruli (Hart, T., Frank, D. D., Lopes, L. E., Olivos-Cisneros, L., Lacy, K. D., Trible, W., ... & Kronauer, D. J., s. d.).

So, it appears that odors could activate hundreds of glomeruli, which would pose challenges for higher-order processing. The researchers in (Hart, T., Frank, D. D., Lopes, L. E., Olivos-Cisneros, L., Lacy, K. D., Trible, W., ... & Kronauer, D. J., s. d.) generated transgenic ants expressing the genetically encoded calcium indicator in olfactory sensory neurons, and by using two-photon imaging, they mapped complete glomerular responses to four ant alarm pheromones.

Interestingly, alarm pheromones activated almost 6 glomeruli, and activity maps for the three pheromones inducing panic alarm in their study species converged on one glomerulus.

Their results showed that ants employ precise and stereotyped representations of alarm pheromones. Furthermore, they stated that a simple neural architecture is sufficient to translate pheromone perception into behavioral outputs (Hart, T., Frank, D. D., Lopes, L. E., Olivos-Cisneros, L., Lacy, K. D., Trible, W., ... & Kronauer, D. J., s. d.).

However, it is also possible that ants communicate by sounds too, and the important role that acoustic signaling has in ant communication is well established and it is unsurprising that other interacting species present adaptations that relate to the acoustic characteristics of the host (Schönrogge, K., Barbero, F., Casacci, L. P., Settele, J., & Thomas, J. A., s. d.).

For concreteness, one can hear a sample of ants sound in this link:

https://exploresound.org/2020/01/ants-have-an-acoustic-world-of-their-own/ (Hickling R., s. d.)

23.2.4 Bee communication

Really, bees do communicate and this fact was proven by scientific evidences. In fact, the Austrian scientist Karl von Frisch (living from 1886 to 1982) observed that the body movements of foraging bees on their return to the nest from a food source correlate with its direction and its distance (Von Frisch, K., s. d.-a) (Von Frisch, K., s. d.-b). Scientists have been amazed by the discovery of this encoded language, based on the waggle dance, and how it could be effectively used to transmit information about the localization of remote objects (see figure 23.2) [(Hrncir, M., Barth, F. G., & Tautz, J., s. d.).

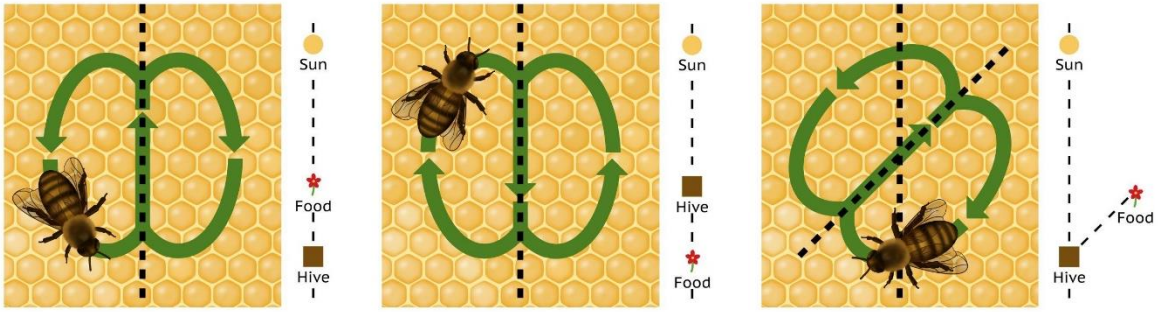


Figure 23.2 (BeesWiki, s. d.): Waggle dance is one of the main types of communication methods used by bees (BeesWiki, s. d.).

Further investigations using recording techniques have revealed that the bee's communication is quite complex and we are only beginning to understand a little of the large field of bee communication. In fact, the bee dance movement is only one chapter of the large story of communication processes that is used by hundreds of bees belonging to a single colony (Hrncir, M., Barth, F. G., & Tautz, J., s. d.).

23.3 Conclusion

In this survey we saw that the holy Quran mentioned the language of birds and ants. For instance, in the verse (27:16), the Quran clearly mentions the language of birds, where it can be noticed that the birds speak with each other.

In the verse (27:20-22), we noticed a discussion between the Prophet Sulaiman and the hoopoe bird, where the bird understanded what was said by Sulaiman, and even responded to him.

Moreover and more surprisingly, as cited in the verse (27:17-19), Prophet Suleiman (Salomon) was even able to listen to an ant that was talking to its community.

When we read the Quran, we understand that animals are grouped in well-organized communities. In fact, according to the holy book, and as it is clearly mentioned in the verse (6:38), it is stated that every creature is organized in communities, like human beings.

So, the holy Book states that animals, or at least those cited in the Quran, do communicate and speak with each other in their own language, even if we do not understand their speech.

On the other hand, and through this investigation, we could cite different high-quality scientific research works related to animal communication, such as dolphins, birds, ants and honeybees, illustrated in the following publications: (Fulton, J. T., s. d.), (Lammers, M. O., Au, W. W., & Herzing, D. L., s. d.), (Ferrer-i-Cancho, R., & McCowan, B., s. d.), (Birdfact, s. d.), (Spiess, S., Mylne, H. K., Engesser, S., Mine, J. G., O'Neill, L. G., Russell, A. F., & Townsend, S. W., s. d.), (Mercer D., s. d.), (Hart, T., Frank,

D. D., Lopes, L. E., Olivos-Cisneros, L., Lacy, K. D., Trible, W., ... & Kronauer, D. J., s. d.) , (Hickling R., s. d.), (Von Frisch, K., 1965-b) (Von Frisch, K., 1965-a) and (Hrncir, M., Barth, F. G., & Tautz, J., 2005), and which explicitly show that those animals do communicate within their community and do possess a real organized way of communication with a specific language, as seen in sections 23.2.1, 23.2.2, 23.2.3 and 23.2.4.

Consequently, the scientific discoveries in this research field clearly confirm the main concept of animal communication revealed in the holy Quran, and which was reported 14 centuries before.

Now, the rising question would be: Is it possible that the holy Quran, with all that embedded knowledge, could be written by an illiterate human being from the 6th century?
The response is obviously: No. Moreover, it is not surprising to see many other scientific discoveries sustaining the Divine origin of this holy book.

24 Effect of the holy Quran in Soul Appeasement and Treatment of Anxiety: An experimental Evidence on the Divinity of the Book

The holy Quran is a fascinating book that creates several positive effects on the reader. Such impacts have been reported by a lot of researchers and can even be felt by reading the holy book either in its original version or in its translated version.

In this paper, we try to expose some related works on the appraisement of the Soul and on the reduction of anxiety, based on the reading or listening of the holy Quran.

The reported results of those research works have unanimously shown that the holy Quran does have a real impact on the reduction of anxiety by treating different negative psychological feelings and depressive disorders. Moreover, it provides hope, confidence and motivation.

24.1 Experimental studies

Four experimental studies were reported by several researchers to try evaluate the effect of holy Quran recitation on the reduction of anxiety. Those studies are described as follows:

24.1.1 Study 1

The first study was conducted in 2016, with the objective of determining the effect of the Quran recitation on mental health of the medical staff of Mazandaran University of Medical Sciences.

Hence, 80 medical staff of the University were gathered according to inclusion criteria's. Inclusion criteria's were (i) complete satisfaction to study, (ii) no having mental or physical disorders. Then participants were randomly distributed into two groups (40 participants in control and experimental group). Experimental group listened to some verses of the Holy Quran for 3 months at the beginning of each working day for 3 minutes, while the control group didn't receive the Quran recitation.

Results showed that the mean of mental health and all its domains, after hearing the verses of the Quran, in experimental group (listening to the Quran) was higher than the control (not listening to the Quran) group ($p < 0.05$) (Darabinia M., Gorji A.M.H., Afzali M.A., 2017).

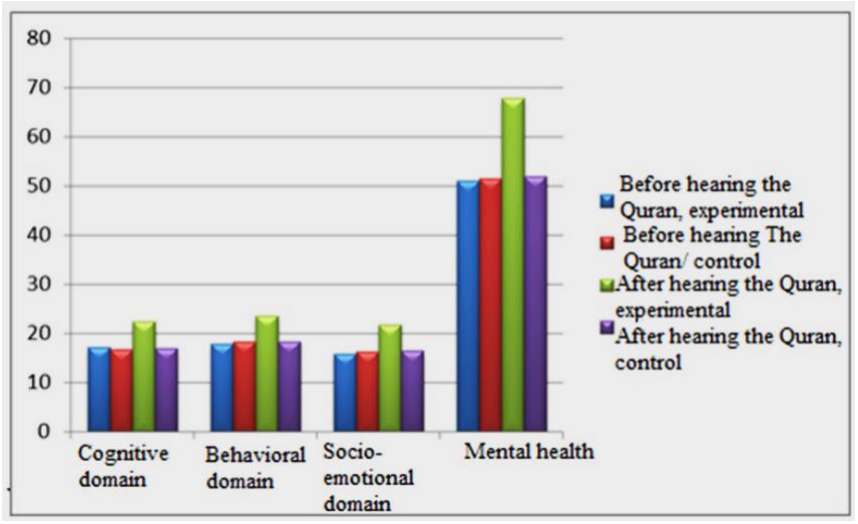


Figure 24.1: The mean of mental health before and after the playing verses of the Quran (Darabinia, 2017). We can notice that the mental health is better after Quran recitation (see difference between histograms green and blue) (Darabinia M., Gorji A.M.H., Afzali M.A., 2017).

Their study revealed the positive effect of hearing the Quran on the mental health of participants. As a conclusion and on the basis of their findings; it can be deduced that hearing the Quran recitations improves the mental state of people.

Furthermore, regarding the close connection of the teaching staff with the students of the University where the experiments were done, **they showed that hearing the Quran can make the staff to feel more satisfied and do their tasks with more optimism**; thus, the students will be pleased as well. As a consequence, they recommended the use of Quran recitations to reinforce positive emotions and psychological comfort for the University staffs (Darabinia M., Gorji A.M.H., Afzali M.A., 2017).

24.1.2 Study 2

Another review study was performed by Ashraf Ghiasi on articles published between January 1990 and September 2017. Several online databases including Scopus and Google Scholar were searched with the keywords of "Quran," "anxiety," "clinical trial." The risk of bias across all included studies was assessed using the Cochrane Collaboration's risk of bias tool.

In their study, the authors reported that from 973 articles found in the initial search, 28 randomized controlled trials and quasi-experiments were selected for the systematic review. Also, In most studies, State-Trait Anxiety Inventory was used to measure participants' anxiety.

The results of this review revealed a positive effect of listening to Holy Quran recitation in reducing anxiety in various settings (Ghiasi A., Keramat A., 2018).

The current evidence indicates that listening to Holy Quran recitation is a useful non-pharmacological treatment for reducing anxiety. However, due to the limited number of studies in this area, further research is needed to obtain more accurate evidence (Ghiasi A., Keramat A., 2018).

24.1.3 Study 3

Another different research work evaluated the effect of Quranic therapy on psychological diseases. The experiments have been conducted on 121 patients from both genders. There were different sessions with the patients, who were given some verses from the Holy Quran for listening during a specific time. Thereafter, every patient was given a remedy program. This study aimed to measure the effectiveness of patients to receive treatment through Quran. The results of the effectiveness factor came after ability and willingness and gave a result of 92.6% for those who support the fact that the Quran has a significant healing influence. **The authors report that some of the patients who regularly attended Quranic therapy sessions have been successfully cured**; and 81.8% of the sample believe that Quranic therapy supports their health needs. They also concluded by stating that this study has empirically proved that the sound of the Holy Quran is an effective treatment for spiritual and psychological issues (Saged, A.A.G., Mohd Yusoff, M.Y.Z., Abdul Latif, F. et al., 2020).

24.1.4 Study 4

In 2022, Gavgani et al. (Gavgani Z., V., Ghojazadeh, M., Sadeghi-Ghyassi, F., & Khodapanah, T., 2022) tried to evaluate the effects of listening to Quran recitation on reducing preoperative anxiety, since such anxiety is a very common unpleasant reaction among patients waiting to undergo a surgery. A systematic review, for collecting the data, was performed in Medline, EMBASE, Cochrane Library, PsycINFO, Arab World Research Source, and other relevant databases.

Randomized trials about the effects of listening to Quran on preoperative anxiety reduction in elective surgery were selected without language or date restriction.

The Cochran's Q statistic and the I^2 index with 50% threshold were used for calculating the heterogeneity and inconsistency index. Furthermore, subgroup analysis was conducted based on the surgery type and the funnel plot was used to assess the possibility of publication bias. Basically, twelve studies were included in the qualitative synthesis and nine were included in the quantitative synthesis.

The meta-analysis showed a significant anxiety reduction with listening to Quran recitation. The heterogeneity between the included studies was statistically significant ($Q = 23.05$, $I^2 = 65.29$, $p = 0.003$). The pooled effect size of anxiety was $d = -8.893$ (95% confidence interval (CI) = −10.763 to −7.022) ($p < 0.001$). So, the analysis showed that listening to Quran recitation reduces anxiety in both major and minor surgeries.

The findings of this statistical investigation indicated that listening to Quran recitation can be considered as a non-invasive and peaceful intervention to reduce preoperative anxiety among patients waiting to undergo a surgery (Gavgani Z., V., Ghojazadeh, M., Sadeghi-Ghyassi, F., & Khodapanah, T., 2022).

24.1.5 Summary of the four studies

In study 1, the researchers showed that hearing the Quran can make the staff to feel more satisfied and do their tasks with more optimism.

In study 2, the results revealed a positive effect of listening to Holy Quran recitation in reducing anxiety in various settings.

In study 3, the authors reported that some of the patients who regularly attended Quranic therapy sessions have been successfully cured.

In study 4, the results indicated that listening to Quran recitation can be considered as a non-invasive and peaceful intervention to reduce preoperative anxiety.

Hence, the four experiments have shown that listening to the recitation of the Quran does have a real positive impact on the soul appraisement and anxiety reduction.

24.2 Quran effect in coping with anxiety

The way people cope with anxious events represents a major factor in whether or not their health is affected. For Instance, if we take the example of divorce, most people would say that divorce is one of the most destructing life crises, since it affects all the family and generates a serious stressful environment (Ayad, A., 2008).

That is, let us look at the Talaq chapter (verses 2 and 3) in the Quran. Reading these verses makes you feel as if a merciful hand is caressing you and giving you hope for the future. In fact, the verses are filled with positive attitude - assurances from Allah that things will get better, that there is a foreseeable end to the current problems and sadness (Ayad, A., 2008).

فَإِذَا بَلَغْنَ أَجَلَهُنَّ فَأَمْسِكُوهُنَّ بِمَعْرُوفٍ أَوْ فَارِقُوهُنَّ بِمَعْرُوفٍ وَأَشْهِدُوا ذَوَيْ عَدْلٍ مِّنكُمْ وَأَقِيمُوا الشَّهَادَةَ لِلَّهِ ۚ ذَٰلِكُمْ يُوعَظُ بِهِ مَن كَانَ يُؤْمِنُ بِاللَّهِ وَالْيَوْمِ الْآخِرِ ۚ **وَمَن يَتَّقِ اللَّهَ يَجْعَل لَّهُ مَخْرَجًا** ﴿٢﴾ **وَيَرْزُقْهُ مِنْ حَيْثُ لَا يَحْتَسِبُ ۚ وَمَن يَتَوَكَّلْ عَلَى اللَّهِ فَهُوَ حَسْبُهُ** ۚ إِنَّ اللَّهَ بَالِغُ أَمْرِهِ ۚ قَدْ جَعَلَ اللَّهُ لِكُلِّ شَيْءٍ قَدْرًا ﴿٣﴾ (65:2-3)

***Translation**: So when they are about to reach their appointed term, hold them back with kindness or separate them with kindness, and make two just men among you as witnesses, and establish the testimony for Allah; with this is advised whoever believes in Allah and the Last Day; **and whoever fears Allah – Allah will create for him a way of deliverance** (2). **And will provide him sustenance from a place he had never expected**; **and whoever relies on Allah – then Allah is Sufficient for him**; indeed Allah will accomplish His command; indeed Allah has set a proper measure for all things (3). [65:2-3]*

Amazingly, by only reading these verses, with a sincere belief in Allah's promise, power, mercy and wisdom, is sufficient to reduce stress and give hope, which will help the depressed persons to cope with an anxious situation, even in the most difficult situations (Ayad, A., 2008).

In a psychological point of view, the cognitive appraisal as introduced by Professor Richard Lazarus (Lazarus, R. S., & Folkman, S., 1984), could describe how different changes and various encountered circumstances influence individuals. So, some people perceive any problem as menacing and stressful; while others approach their problem with a fighting spirit, favoring adjustment and adaptation. Thus, perceiving stressful situations as harmful complicates our ability to cope with these situations. On the other hand, seeing them as challenging enables us to deal efficiently with the events (Ayad, A., 2008).

However, for believers, the way of coping with emotions depends primarily on the degree of faith. A deep trust in Allah, associated with the fact that this world is transient, can give a real strength and feeling of peace and satisfaction. A true believer never falls into despair, since Allah is present and since He promises to reward patience in this life and in the hereafter (Ayad, A., 2008). See the following verse of the Quran:

قُل لَّن يُصِيبَنَا إِلَّا مَا كَتَبَ اللَّهُ لَنَا هُوَ مَوْلَانَا ۚ وَعَلَى اللَّهِ فَلْيَتَوَكَّلِ الْمُؤْمِنُونَ ﴿9:51﴾

Translation: ***Say: "Nothing will happen to us except what Allah has decreed for us: He is our protector": and on Allah let the Believers put their trust.*** (Quran 9: 51)

In this way, the fact of reading the Quran, associated with a great faith in Allah, can influence the response to a stressful situation, in changing the fear or anxiety into calm, peace and hope. It also highly contributes in improving the coping attitude by producing a relaxing feeling and by calming down the stress.

24.3 Discussion

In this scientific survey, we noticed some real positive effects of the holy Quran (reading, listening or recitation) on the appeasement of anxiety and the improvement of motivation in stressful or difficult situations.

Several experimental investigations were made on real patients, as reported by Darabinia in the 1st study (Darabinia M., Gorji A.M.H., Afzali M.A., 2017), by Ghiasi in the 2nd study (Ghiasi A., Keramat A., 2018), by Saged in the 3rd study (Saged, A.A.G., Mohd Yusoff, M.Y.Z., Abdul Latif, F. et al., 2020) and by Gavgani in the fourth study (Gavgani Z., V., Ghojazadeh, M., Sadeghi-Ghyassi, F., & Khodapanah, T., 2022). Strangely, the four experiments have shown that listening to the recitation of the Quran does have a real positive impact on the soul appraisement and anxiety reduction, and these discoveries were scientifically proved by statistics.

That is, what could be the secret in such spiritual strength? Is it only the effect of words? But what types of words could influence so deeply the soul and treat many psychological pains? Sincerely, we do not see any possible cause except the fact that the holy Quran should be sent down by a Super Power Creator (Allah, praise be upon Him) who embedded all the necessary treatment required in such situations in His holy book.

25 Statistical Investigation on Ancient Quran Folios: Case of Birmingham and Sanaa Parchments

In 2014, some scientists of the University of Birmingham discovered that four folios containing some ancient Quran manuscripts can be dated sometime between 568 and 645CE. This means that the animal from which the skin was taken was living sometime between these dates. Similarly, in 1965 heavy rains damaged the roof construction of the Western Library in the Great Mosque of Ṣanaa, where scientists discovered some Quran folios that should belong to the period between 578 CE and 669 CE. This means that those Quran manuscripts are probably dated from the period of the Prophet's companions.

In this investigation, we want to check whether those ancient texts are similar to the present Quran or not, and if the two ancient manuscripts, discovered in Birmingham and Sanaa, contain similar text or not.

The first results, based on character analysis and word analysis, have shown that the two old folios are very similar to their corresponding part contained in the present Quran (Uthmanic compilation). So, it appears that the morphological skeleton of the analysed Quran text has been safely preserved during the last 14 centuries.

25.1 Introduction

A brief description and history on the two old parchments, namely: Birmingham Quran and Sanaa Quran, are depicted below.

25.1.1 Introduction on the Birmingham Quran manuscript

The Birmingham Quran manuscript consists in four pages made of parchment, written in ink, and containing parts of chapters 18, 19 and 20 of the holy Quran (Fedeli A., 2015). The manuscript forms part of the University of Birmingham's Mingana Collection of Middle Eastern manuscripts, held in the Cadbury Research Library (Birmingham University, 2015).

Sir Cadbury named this collection "the Mingana Collection" after its first curator (Hopwood D., 1961). The collection came to the University of Birmingham in the late 1990s. Concerning the palaeographic aspect of the manuscript (*Titled Hejazi text*), the handwriting geometry suggests that it may have been created in the Hejaz area in the west of the Arabian Peninsula, which includes the sacred cities of Mecca and Medina. In fact, there are several old manuscripts dating from the first centuries after the Hijra, where we can clearly see the difference in the palaeographic style (Awwad K., 1982). The palaeography can give a quite good estimation on the probable date of the manuscript, but the radiocarbon dating is usually more accurate with regards to the parchment dating. This last technique is widely used in archaeological dating (Taylor R.E. and Aitken M.J., 1997).

Thus, the radiocarbon analysis, made at the Radiocarbon Accelerator Unit of Oxford University (Ramsey B. C., Higham T. F. G., Brock F., Baker D., & Ditchfield P., 2009), yielded the following technical dating results (Higham T. F., Bronk Ramsey G. C., Chivall D., Graystone J., Baker D., Henderson E., Ditchfield P, 2018) (Birmingham University, 2016):

- Radiocarbon Reference: OxA-29418
- Radiocarbon Result: $\delta 13$ C = -21.0 ‰ 1456+/-21
- Calibrated Date Range: 95.4% probability to be between 568 and 645 AD

Hence, the manuscript has been radiocarbon-dated by the University of Oxford (Radiocarbon Accelerator Unit) to the date range of 568–645 CE with a 95.4% degree of confidence. The radiocarbon result means that the animal from which the skin was taken was living sometime between these specific dates. This places the discovered parchment close to the death of the Prophet who lived between 570 and 632 CE.

Some researchers argued that the manuscript is among the earliest written textual document of the Quran known to survive, which was written few years after the Prophet death. They also claim that it should probably be the oldest Quran manuscript in the UK.

In this investigation, we try to check whether the ancient text is similar to the corresponding Sanna Quran part and to the present Uthmanic Quran by means of comparative analysis.

25.1.2 Introduction on the Sanaa Quran manuscript

In 1965 heavy rains damaged the roof construction of the Western Library in the Great Mosque of Ṣanaa. A single window was discovered to contain a substantial cache of used Arabic manuscripts, almost all being ancient manuscripts of the Quran spanning the first few Islamic centuries. A Colloquium on the Islamic City organised by the World of Islam Festival Trust, sponsored by UNESCO, was held at the University of Cambridge, in July 1976. Drawing a wide variety of experts from both the Muslim and non-Muslim world, a number of specific research activities were recommended, amongst which was highlighted the pressing need to conserve the rich corpus of Quranic texts discovered in the Mosque of Ṣanaa (UNESCO, 1980).

The Ṣanaa collection was made known to the general public with the publication of Maṣāḥif Ṣanʿāʾ in 1985, an exhibition catalogue presenting some of the findings of the project. A single palimpsest folio from the part of the codex located in Dār al-Makhṭūtāt (i.e., *DAM 01-27.1*), *folio 21a* according to Sadeghi and Goudarzi's classification, was displayed along with some brief comments regarding the script and its contents. The folio was tentatively dated to the first half of the 1st century of *hijra* (Sanaa, 1985).

In 2010 Sadeghi and Bergmann had published their article analysing the four auction folios, specifically the *Sotheby's 1993* / *Stanford 2007* folio, where details were given of a radiocarbon study corroborating the early date already assigned to the manuscript. Analysis was done at the Accelerator Mass Spectrometry Laboratory of the University of Arizona (Sadeghi, B., & Bergmann, U., 2010). According to Sadeghi and Bergmann, the results indicated that the parchment had a 95% (2σ) probability of belonging to the period between 578 CE and 669 CE.

In our investigation, we are particularly interested in the parchment referenced by the reference 029006B, by making a statistical comparison of this manuscript with its corresponding part in Birmingham folios and the present Uthmanic compilation.

25.1.3 Goal of this work

We try to conduct a statistical analysis for evaluating the difference, if any, between the discovered parchments and the present compilation of the Quran. The used discrimination technique is based on the computation of characters and words that are different, within the manuscripts.

25.2 Notes on the ancient Arabic handwriting

Over the time, a great variety of manuscript copies of the Quran survived, along with commentaries written by several ancient scholars. Furthermore, a large exploration of the ancient Arabic manuscripts shows how the Quran, in terms of calligraphy, evolved over time in different regions of the world (Ansorge, C., 2011).

25.2.1 Hijazi script

The earliest Quran manuscripts were probably written in the 7th century. Its script represents a very early Quran text written on a single large sheet of parchment and folded in half. This primitive script is known as hijāzi script, which is originated from the Arabian Peninsula around Mecca and Medina. The hijāzi script is characterized by tall and sloping letters representing the consonants. Also, one can find few dots and other markings indicating pronunciation or pauses (Ansorge, C., 2011).

25.2.2 Diacritics and rasm dots

The new Arabic writing structure has undergone a gradual improvement, which may be noticeable in modern printed Quran books (Denffer A.V., 2009). The first change, in the writing structure, was the addition of the diacritical marks (tashkeel), corresponding to vowels. The second one was the addition of dots, for discriminating between some close consonants.

So, the ancient Arabic characters are a bit different from what we are used to write nowadays. In fact, as displayed in figure 25.1, there is a correspondence between the ancient characters (*at the bottom*) and the new equivalent ones (*at the top*).

Figure 25.1: Ancient Arabic characters and their corresponding recent Arabic characters.

One of the particularities of the ancient Arabic scripture is the fact that the text was not vocalised and did not contain Rasm dot marks. For example, in the ancient script, the word *Alaiha* (عَلَيْهَا) was written as follows: .

25.2.3 Note on the "silent alif"

The natural elongation, called: *Al-Madd Al-tabe**e**e*, is the act of elongating the sound of the three basic *vowels* (Muqith M. A., 2011), consisting in stretching the sound of a character to two basic ones or more. However, the Ancient Arabic script was not vocalised and often not accompanied by elongation marks. For the case of the vowel A (*fatha* in Arabic), th*e corresponding elongation mark is called "silent alif". But we* have noticed that in the ancient manuscripts, the silent *alif (elongated A)* was not very used as it is the case in the recent Arabic text. In fact, prior to the orthographic reforms carried out under the auspices of the Umayyad caliph Abd al-Malik (AH 80), the Arabic script did not contain a lot of woyels," i.e., there were no short vowels and no grapheme to represent the hamza or the long alif, example: the verb ***qaala*** (قَالَ) in the modern Arabic script, was often written ***qala*** (قَلَ) without silent *alif* in the ancient Arabic one.
That is why, it is not surprising to see the ancient Quran manuscript without diacritics or without elongation marks, since most of these marks were invented several centuries afterward, probably under the auspices of the Umayyad caliph Abd al-Malik in 80 AH (Powers D. S., 2011).

25.3 Analysis of the Birmingham Quran

In this investigation we will make a comparative analysis on the four ancient Birmingham folios with regards to their corresponding verses in the present Quran (*Hafs Recitation*). Our discrimination technique is based on the computation of characters and words that are different, within the manuscripts. The two first folios correspond mainly to chapter 18 (Al-Kahf), and the two last folios correspond mainly to chapter 20 (Taha) and the end of chapter 19 (Maryam).

The following table (table 25.1.a) consists of two columns; the left one contains a part of Birmingham Quran folio and the right one contains the corresponding verses in the present Quran (Hafs recitation with Uthmanic rasm) without vocalisation (i.e. without diacritics).

Table 25.1.a: Example of comparison between a part of Birmingham Quran and the present Quran without vocalisation

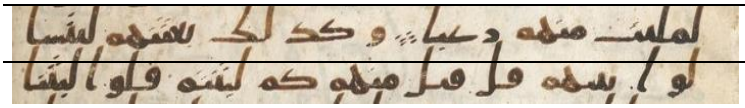
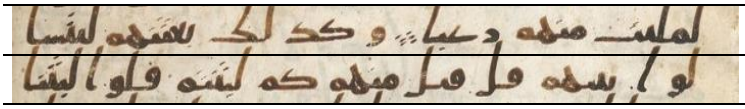
Birmingham Quran	**Present Quran** (old rasm without vocalisation)
	لملئت منهم رعبا (18) وكذلك بعثنهم ليتسا لوا بينهم قال قائل منهم كم لبثتم قالوا لبثنا

Table 25.1.b: Example of comparison between a part of Birmingham Quran and the present Uthmanic Quran

Birmingham Quran	**Present Uthmanic Quran**
	فِرَارٗا وَلَمُلِئۡتَ مِنۡهُمۡ رُعۡبٗا ١٨ وَكَذَٰلِكَ بَعَثۡنَٰهُمۡ لِيَتَسَآءَلُواْ بَيۡنَهُمۡۚ قَالَ قَآئِلٞ مِّنۡهُمۡ كَمۡ لَبِثۡتُمۡۖ قَالُواْ لَبِثۡنَا

By comparing all the folios with the present Uthmanic rasm, we have obtained the following statistics that are presented in table 25.2.

Table 25.2: Statistics of all Birmingham folios

Manuscript	**Number of verses**	**Number of lines**	**Similarity** in terms of words	**Similarity** in terms of characters without considering the "silent *alif*"	**Similarity** in terms of characters by considering the "silent *alif*"
Birmingham Folio 1	7	24	100%	100%	98.08%
Birmingham Folio 2	9	23	100%	100%	98.74%
Birmingham Folio 3	19	23	100%	100%	99.5%
Birmingham Folio 4	28	23	100%	100%	98.6%
All Folios	**63**	**93**	**100%**	**100%**	**98.71%**

Note: The similarity in terms of words (or characters) is equal to: *100% - the number of different words (or characters) that are different in percent*.

As one can see in table 25.2, which represents a statistical comparison between this ancient Quran folios and the present one compiled by Uthman, it appears two important conclusions:

- The two analysed text documents are similar in terms of words (*similarity of 100%*);
- The two analysed text documents are quite similar in terms of characters (*similarity of about 100% without considering the sildent alif and about 99% by considering the sildent alif*);

Consequently, and since the ancient Birmingham scripture was found to be morphologically similar to the present holy scripture, it appears that the skeletal morphology of the Quran, for the investigated chapters, has been safely preserved during the last 14 centuries.

25.4 Analysis of the Sanaa Quran

In this investigation we will make a comparative analysis between the present Uhmanic Quran and its available corresponding part found in Sanaa. We should note that due to the limited available data in Sanaa Quran, we could only make a comparison from the end of chapter 19 and the beginning of chapter 20 in in the Sanaa Quran. Our discrimination technique is based on the computation of characters and words that are different, within the manuscripts. Once again, the diacritics and modern writing rasm have been removed to put the two texts in the same writing conditions (i.e. diacritics removal).

25.4.1 Statistical analysis of the Sanaa folio

In summary, by taking the Sanaa Quran and comparing this last one with the present Uthmanic Quran without vocalisation, we have obtained the following statistical results.

Table 25.3: Similarity between the ancient Sanaa folios and the present Uthmanic Quran

Comparison between the ancient Sanaa folios and the present Uthmanic Quran	**Similarity in %**
Similarity in terms of words	100 %
Similarity in terms of characters without considering the "silent *alif*"	100 %
Similarity in terms of characters by considering the "silent *alif*"	99.07%

Note: The similarity in terms of words (or characters) is equal to: *100% - the number of different words (or characters) that are different in percent*.

So, as one can see in table 25.3, all the similarity ratios are almost equal to 100%, showing that the old Sanaa parchment is almost identical to the present Quran in terms of skeletal morphology.

25.4.2 Comparison between Sanaa and Birmingham folios

In this investigation, we make a comparison between the two ancient folios. However, due to the limitation in the available data, the comparison is made between the verses [19:91] and [20:39].

The statistical results of this comparison are represented in table 25.4.

Table 25.4: Similarity between the Sanaa Quran folios and Birmingham Quran folios

Comparison between the ancient folios and the current holy Quran	**Similarity in %**
Similarity in terms of words	100 %
Similarity in terms of characters without considering the "silent *alif*"	100 %
Similarity in characters by considering the "silent *alif*"	99.84%

Note: The similarity in terms of words (or characters) is equal to: *100% - the number of different words (or characters) that are different in percent*.

We represented the images of these ancient folios in figure 25.2, the left side contains the 3rd and 4th Birmingham Quran folios, while the right side contains its corresponding verses in Sanaa Quran.

Figure 25.2: Comparison between Birmingham Quran (left) and Sanaa Quran (right), corresponding to the end of chapter 19 and beginning of chapter 20.

By comparing the two old parchments, one can notice that the different similarity measures are almost equal to 100%, showing that the two parchments, Sanaa and Birmingham manuscripts, present a very high similarity, despite their different geographical origins.

25.5 Discussion

In summary, during this investigation, we have undertaken a comparative analysis between the Birmingham Quran folios, Sanaa Quran folios and their corresponding verses in the Uthmanic compilation of the present Quran, which correspond to the end of chapter 19 and the beginning of chapter 20, where the comparative analysis is based on characters and words. This investigation is important and interesting because the parchments of Birmingham and Sanaa Quran, have been carbon-dated to the first century of Hijra.

Based on the obtained statistical results, we have identified two significant conclusions:

- The two analysed manuscripts: Birmingham and Sanaa folios, appear similar in terms of characters and in terms of words;
- The ancient texts (end of chapter 19 and beginning of chapter 20) appear identical to the corresponding present Uthmanic Quran, with regards to the skeletal morphology of the analysed texts.

Consequently, the old parchments corresponding to chapter 19 and 20, show that the consonantal morphology of the Quran has been safely preserved for those chapters, during the last 14 centuries without alteration.

That is, if the radiocarbon dating is quite accurate, we can say that this new discovery confirms that the present holy book should represent an authentic copy of the first original Quran that was recited by the Prophet fourteen centuries ago.

From another vision, and by reading the verse (15:9) of the Quran: « إِنَّا نَحْنُ نَزَّلْنَا الذِّكْرَ وَإِنَّا لَهُ لَحَافِظُونَ», it is stated that the holy Scripture is protected and preserved by His Creator; which strengthens the result of this scientific investigation, at least for the investigated chapters.

26 General Conclusion and Discussion

In this survey, we tried to see whether the Quran could be a simple invention of the Prophet (i.e. written by the Prophet) or really a book from God (i.e. a divine book sent down by Allah) as claimed in the Islamic religion.

This survey has to be considered as a pure scientific investigation without any form of theological or ideological point of view. Also, the author does not discuss his personal beliefs on the subject, but only what the scientific results of this investigation show.

Again, it is important to recall that the consequences of this research work could be very heavy, that is why we should comment the different results with care, rigor and objectivity.

So, in this second edition of the book, we can find a brief description of the two analysed books (i.e.Quran and Hadith), we can also find the description of the 13 series of experiments that were conducted during this investigation and some new scientific knowledge that were embedded in the Quran.

The 13 series of experiments are depicted as follows:

- 1st Series of Experiments, in chapter 7: Global Analysis
  This is a global authorship discrimination
- 2nd series of experiments, in chapter 8: Big Segments based Segmental analysis
  This is a segmental authorship discrimination
- 3rd series of experiments, in chapter 9: Automatic authorship attribution with several features and several classifiers
- 4th series of experiments, in chapter 10: Short Segments based Segmental Authorship Attribution
- 5th Series of Experiment, in chapter 11: Stylometric Comparison between the Quran and Hadith based on Successive Function Words
- 6th Series of Experiment, in chapter 12: Authorship Identification of 7 Books – A Fusion Approach
- 7th Series of Experiments, in chapter 13: Authorship Discrimination using the Leave-One-Out Validation
- 8th Series of Experiments, in chapter 14: Authorship Discrimination based on Gaussianity and Interpolability
- 9th Series of Experiments, in chapter 15: A Mysterious Numerical Structure in the Quran making it different from other Human Books
- 10th Series of Experiments, in chapter 16: Authorship Attribution based on the Interrogative Form
- 11th Series of Experiments, in chapter 17: Investigation on the Quran/Hadith Authorship Using Visual Analytics Approaches
- 12th Series of Experiments, in chapter 18: Authorship Discrimination based on Word Transition Probability
- 13th Series of Experiments, in chapter 19: Authorship Discrimination based on Deep Learning Technology

As one could notice in all the results reported in those 13 series of experiments, the conclusions are absolutely the same, namely: The Author's style of the Quran is completely different from the author's style of the Hadith (i.e. Prophet), which leads to the following major conclusion: The Quran could not be written or invented by the Prophet Muhammad. Hence, the claim that the Quran was sent down by God (Allah) is widely strengthened and appears to be true.

At the end of this book, more specific research works with regards to some important information and knowledge embedded in the holy Quran, have been reported. The specific works discussed are presented and commented in the following chapters:

- Chapter 20: Book Analysis based on Embedded Scientific Knowledge
- Chapter 21: About the "Last Prophet" Meaning and Prediction of a Prophet called Muhammad in the Ancient Religious Books
- Chapter 22: Does the Heart have a Control on Mind and Emotions? A Scientific Evidence Supporting what is Said in the Holy Quran
- Chapter 23: Do Animals communicate with each other? A Scientific Evidence Supporting what was Revealed in the Quran
- Chapter 24: Effect of the holy Quran in Soul Appeasement and Treatment of Anxiety: An experimental Evidence on the Divinity of the Book
- Chapter 25: Statistical Investigation on Ancient Quran Folios: Case of Birmingham and Sanaa Parchments

- In chapter 20, titled: 'Book Analysis based on Embedded Scientific Knowledge', we could see several new scientific knowledge that were embedded in the Quran and commented by several famous researchers, and which have further confirmed this conclusion by showing that the holy book could not be written by a human being, but it should belong to a super-power intelligence that has an extreme power and very large scientific knowledge (i.e. Allah/ God).

- In chapter 21, titled: 'About the "Last Prophet" Meaning and Prediction of a Prophet called Muhammad in the Ancient Religious Books', we explored the significance of Muhammad being considered as the final prophet and delved into the philosophical and logical implications of this fact, where we could derive four key implications and conclusions that strengthen the Quran truthiness. On the other hand, we tried to explore some ancient prophecies about the Prophet Muhammad apparition in the ancient holy books, where we could also find four amazing prophecies on the coming of another prophet from the Arabic peninsula who cannot be anyone other than Muhammad (Pbuh), which strengthen the Prophet truthiness too.

- In chapter 22, titled: 'Does the Heart have a Control on Mind and Emotions? A Scientific Evidence Supporting what is Said in the Holy Quran', we cited some recent research works in neurocardiology that have led to very interesting results reinforcing the Quran theory by showing the important role of the heart in mind and feelings. This discovery, which has changed the interpretation of the heart role, represents a real scientific revolution. That is, it appears that the new scientific discovery about the important role of the heart in feelings, wisdom and reasoning are confirming several Quran verses, which give to the heart a key importance in

belief and decision making. So, once again, it is evident that this ancient book, dating from the 7th century, could not be a human invention.

- In chapter 23, titled: 'Do Animals communicate with each other? A Scientific Evidence Supporting what was Revealed in the Quran', we could see that the holy Quran mentioned the language of birds and ants. In the verse (27:20-22) for instance, we noticed a discussion between the Prophet Sulaiman and the hoopoe bird, where the bird understood what was said by Sulaiman, and even responded to him. So, according to the holy book, it is clearly stated that every creature is organized in communities, like human beings. So, the holy Book states that animals, or at least those cited in the Quran, do communicate and speak with each other in their own language, even if we do not understand it. Furthermore, and through the cited investigations, we could see different scientific research works related to animal communication, such as dolphins, birds, ants and honeybees, and which explicitly showed that those animals do possess a real organized way of communication. Consequently, the scientific discoveries in this research field clearly confirm the main concept of animal communication revealed in the holy Quran, and which was reported 14 centuries before.

- In chapter 24, titled: 'Effect of the holy Quran in Soul Appeasement and Treatment of Anxiety: An experimental Evidence on the Divinity of the Book', we noticed a real positive effect of the holy Quran on the appeasement of anxiety and the improvement of motivation. In fact, several experimental investigations were made on real patients, and strangely, all experiments have shown that listening to the recitation of the Quran does have a real positive impact on the soul appraisement and anxiety reduction. But what could be the secret in such spiritual strength? Sincerely, we do not see any possible cause except the fact that the holy Quran should be sent down by the Creator (Allah, praise be upon Him) who embedded all the necessary treatment required in such situations in His mysterious book.

- In chapter 25, titled: 'Statistical Investigation on Ancient Quran Folios: Case of Birmingham and Sanaa Parchments', we made a comparative analysis between the ancient Birmingham Quran folios, Sanaa Quran folios and their corresponding verses in the Uthmanic compilation of the present Quran, where the comparative analysis was based on characters and words. That is, according to the radiocarbon dating of the ancient folios and the obtained statistical results, we were able to confirm that the present holy book (at least in its skeletal morphology) should represent an authentic copy of the first original Quran that was recited by the Prophet fourteen centuries ago.

Finally, we can say that the discoveries reported in this works have shed light on an old enigma that had not been solved (scientifically) for 14 centuries.

Again, the different investigations and results cited in this book can only reflect the authenticity, nobility, supremacy, and divinity of the holy Quran.

Nevertheless, it is clear that further research in this domain is essential, since continued scientific investigations will probably yield increasingly significant and impressive findings regarding this extraordinary holy Book.

27 Personal Feeling

Many questions could interest the curiosity of the researcher inside his personal field/box of interest (*in a specific limited field*), leading him to try making an effort to understand what is happening inside that field/box of interest...

When he manages to understand some of the related theories (*in that specific limited field*), he may experience a great feeling of success. That is good… However, when he raises his eyes above the limits of his box of interest, he may experience a strange feeling of sadness due to the ignorance about what could exist outside that box. But since he decided to raise his eyes (*towards the huge universe*), that is far better…

Thus, it has been important for me to try looking for the truth through the scientific analysis of what does exist between our hands (*i.e. the hoy Quran*). Consequently, several investigations have been conducted and analysed carefully to avoid misinterpretations or subjective assessments.

Also, by preparing this book (*3rd edition*), I tried to do my best to avoid any error or mistake, due to the delicacy of the subject. However, I still present my sincere apologies in case of any unintentional mistake found in this document.

Hopefully, may I ask this little book to continue its journey across the earth globe, illuminating the minds and hearts of readers.

References

Abdelkafy. (2013). *Last visit in*. http://www.abdelkafy.com

Al-Batineh, M. (2019). Using Forensic Stylistics and Corpus Analysis for Detecting and Investigating Plagiarism in Literary Retranslation : The Plagiarism Controversy of John Payne and Sir Richard Burton. *New Voices in Translation Studies*, *21*.

al-Ghazali, M. (s. d.). *From Wikipedia, the free encyclopedia (website). Last visit in 2021*.

Al-Kaheel A. (s. d.). *Al-Kaheel A. : Hearts Wherewith they Understand, https://kaheel7.net/?p=2303&lang=en. Last visit on september 12, 2022.*

Almujaiw, S. (2017). Grammatical construction of function words between old and modern written Arabic : A corpus-based analysis. *Corpus Linguistics and Linguistic Theory*, *13*(2).

Alqaradawi. (2013). *Last visit in*. https://www.al-qaradawi.net/.

Al-Qarni, A. (s. d.). *Last visit in 2021*. http://en.wikipedia.org/wiki/Aaidh_ibn_Abdullah_al-Qarni.

Al-Rassi, M. S. (s. d.). *Al-Rassi, M. S. The Amazing Prophecies OF Muhammad in the Bible, 2019. [Dawood, A. A.] David Abdu Benjamin Keldani, Muhammad in the Bible. Adam Publishers 2007, page 189.*

Al-Shreef. (2009). *A. Al-Shreef. Is the Holy Quran Muhammad's invention?*

Amr-Khaled. (2013). *Last visit in*. http://amrkhaled.net

Ansorge, C. (2011). *Ansorge, C., 2011. Faith & Fable : Islamic manuscripts from Cambridge University Library. Cambridge University Publisher.*

Anthony Matthew Jacob. (s. d.). *Anthony Matthew Jacob, Is Prophet Muhammed (s.a.w.a.) In the Bible ? Publisher World Islamic Network, 2017*.

Armour, J. A. (s. d.). *Armour, J. A. (2007). The little brain on the heart. Cleveland Clinic Journal of Medicine, 74, S48-51.*

Awwad K. (1982). *Awwad K., 1982. Aqdam al-makhtutat al-arabiyya fi maktabat al-alam (The Oldest Arabic Manuscripts in the World's Libraries), Baghdad, 1982.*

Ayad, A. (s. d.). *Ayad, A. (2008). Healing body & soul : Your guide to holistic wellbeing following Islamic teachings. International Islamic Publishing House.*

BeesWiki. (s. d.). *BeesWiki, 2023. How Do Bees Communicate ? BeesWiki, https://beeswiki.com/how-do-bees-communicate/ Last access in ugust 2023.*

Birdfact. (s. d.). *Birdfact. How Do Birds Communicate ? (All Methods Explained), Birdfact https://birdfact.com/articles/how-do-birds-communicate, Last updated : 3 August 2022.*

Birmingham University. (2015). *Birmingham University, 2015. Birmingham Quran manuscript dated among the oldest in the world. July 2015. Http://www.birmingham.ac.uk/news/latest/2015/07/quran-manuscript-22-07-15.aspx*.

Birmingham University. (2016). *Birmingham University, 2016. About the Birmingham Quran, FAQs, www.birmingham.ac.uk/ facilities/cadbury/quran-manuscript/faqs.aspx. Last access in july 2016.*

Clement, R., & Sharp, D. (2003). Ngram and Bayesian Classification of Documents for Topic and Authorship. *Literary and Linguistic Computing*, *18*(4), 423-447,.

Contributors. (2015). Contributors of Wikipedia. Central limit theorem. *Consultation on August*, *27*. https://en.wikipedia.org/wiki/Central_limit_theorem.

Corney, M. (2003). *Analysing E-Mail Text Authorship for Forensic Purposes* [Master Thesis,]. Queensland University of Technology.

Cox, I.J., Miller, M.L., Bloom, J.A, Kalker, T, Fridrich, J., & J. (2007). *Digital Watermarking and Steganography*. Elsevier.

Darabinia M., Gorji A.M.H., Afzali M.A. (2017). *Darabinia M., Gorji A.M.H., Afzali M.A. The effect of the Quran recitation on mental health of the Iranian medical staff. Journal of Nursing Education and Practice 2017, Vol. 7, No. 11. Http://jnep.sciedupress.com.*

Dasarathy, B. V. (1994). *Decision Fusion"*. IEEE Computer Society Press.

David Abdu Benjamin Keldani. (s. d.). *David Abdu Benjamin Keldani, Muhammad in the Bible. Adam Publishers 2007, page 189.*

Denffer A.V. (2009). *Denffer A.V., 2009. An Introduction to the Sciences of the Quran. Publisher : Islamic Foundation (UK).*

Eder, M. (2010). Does size matter ? : Autorship attribution, short samples, big problem. *Digital Humanities 2010 Conference*, 132-135.

Ellis, G., Mansmann, F., & VisMaster, V. A. (2010). Mastering the Information Age. *Chapter*, *2*. http://www.vismaster.eu/book/chapter-2-visual-analytics.

El-Zohairy, N. (2008). *A Dictionary of Function Words in Arabic*. Liban Publisher.

Evert, S. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, *32*(suppl_2), 4-16,.

Farringdon, J. M. (1996). *Analyzing for Authorship : A Guide to the Cusum Technique*. University of Wales Press.

Fedeli A. (2015). *Fedeli A., 2015. Phd thesis dedicated to the Qurʾan folios of Birmingham : Alba Fedeli, Early Qur'ānic manuscripts, their text, and the Alphonse Mingana papers held in the department of special collections of the University of Birmingham, 2015*.

Ferrer-i-Cancho, R., & McCowan, B. (s. d.). *Ferrer-i-Cancho, R., & McCowan, B. (2009). A law of word meaning in dolphin whistle types. Entropy, 11(4), 688-701*.

Foster, D. (2001). *Author Unknown : On the Trail of Anonymous*. Henry Holt and Company.

Fredrickson BL, Branigan C. (s. d.). *Fredrickson BL, Branigan C (2005) "Positive emotions broaden the scope of attention and thought-action repertoires." Cognition and Emotion 19 : 313-332*.

Fulton, J. T.,. (s. d.). *Fulton, J. T., 2021. Deciphering the Dolphin Language Appendix U, 2021*.

Galton, 1889] F. Galton. (1889). Book. *Natural Inheritance*, 66.

García-Barrero, D., Feria, M., & Turell, M. T. (2012). Using function words and punctuation marks in Arabic forensic authorship attribution. *IAFL Conference Proceedings*.

Gavgani Z., V., Ghojazadeh, M., Sadeghi-Ghyassi, F., & Khodapanah, T. (2022). *Gavgani Z., V., Ghojazadeh, M., Sadeghi-Ghyassi, F., & Khodapanah, T. (2022). Effects of listening to Quran recitation on anxiety reduction in elective surgeries : A systematic review and meta-analysis. Archive for the Psychology of Religion, 44(2), 111–126. Https://doi.org/10.1177/00846724221102198*.

Ghiasi A., Keramat A.,. (2018). *Ghiasi A., Keramat A., The Effect of Listening to Holy Quran Recitation on Anxiety : A Systematic Review, 2018, Iranian Journal of Nursing and Midwifery Research | Published by Wolters Kluwer – Medknow*.

Greenacre, M. (2014). Hierarchical cluster analysis, online documentation. *Chapter*, *7*, 7 1-7 11.

H. Hadjadj, & Sayoud, H. (2018). Authorship Attribution of Seven Arabic Religious Books-A Fusion Approach. *Conference Paper: 6ème Conférence Internationale En Automatique & Traitement de Signal*.

Hart, T., Frank, D. D., Lopes, L. E., Olivos-Cisneros, L., Lacy, K. D., Trible, W., ... & Kronauer, D. J. (s. d.). *Hart, T., Frank, D. D., Lopes, L. E., Olivos-Cisneros, L., Lacy, K. D., Trible, W., ... & Kronauer, D. J. (2023). Sparse and stereotyped encoding implicates a core glomerulus for ant alarm behavior. Cell, 186(14), 3079-3094*.

HartMath. (s. d.). *HartMath : The Heart–Brain Connection, https://www.heartmath.com/science/. Last visit in august 2022*.

Hawashin, B., Mansour, A., & Aljawarneh, S. (2013). An Efficient Feature Selection Method for Arabic Text Classification". *International Journal of Computer Applications, 83-Number 17*, 1-6.

Hickling R. (s. d.). *Hickling R. Ants Have an Acoustic World of Their Own. Robert Hickling – Exploresound, https://exploresound.org/2020/01/ants-have-an-acoustic-world-of-their-own/ Last access in August 2023.*

Higham T. F., Bronk Ramsey G. C., Chivall D., Graystone J., Baker D., Henderson E., Ditchfield P. (2018). *Higham T. F., Bronk Ramsey G. C., Chivall D., Graystone J., Baker D., Henderson E., Ditchfield P., 2018. Radiocarbon Dates from the Oxford AMS System, Archaeometry Journal : Datelist 36. Vol. 60, Issue3, June 2018, Pages 628-640.*

Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing, 13*(3), 111-117.

Hopwood D. (1961). *Hopwood D., 1961. 'The Islamic Manuscripts in the Mingana collection', Journal of the Royal Asiatic Society XCIII/3-4.*

Hrncir, M., Barth, F. G., & Tautz, J. (s. d.). *Hrncir, M., Barth, F. G., & Tautz, J. (2005). 32 vibratory and airborne-sound signals in bee communication (hymenoptera). Insect sounds and communication : Physiology, behaviour, ecology, and evolution, 421.*

Huang, X., & Pan, W. (2003). Linear regression and two-class classification with gene expression data". *Bioinformatics, 19*(16), 2072-2078,.

Huberty, C. J. (1994). *Applied discriminant analysis*. Wiley.

Ibnu-Kathir. (s. d.). *تفسير القرآن العظيم*. دار طيبة للنشر والتوزيع, الطبعة : الثانية 1420هـ - 1999 م, موقع مجمع الملك فهد لطباعة المصحف الشريف.

Ibrahim., I. 1997] I. A. (1997). *A brief illustrated guide to understanding Islam. Library of Congress, Catalog Card Number : 97-67654, Published by Darussalam, Publishers and Distributors*. http://www.islam-guide.com/contents-wide.htm.

Ibrahim, I. A. (1996). *A brief illustrated guide to understanding Islam. Library of Congress, Catalog Card Number : 97-67654, Published by Darussalam, Publishers and Distributors*. http://www.islam-guide.com/contents-wide.htm,

Illias S. (s. d.). *Illias S., 2018. The Language of the Birds, Article date : September 19, 2018. Signs & Science : The Language of The Birds (signsandscience.blogspot.com) https://signsandscience.blogspot.com/2018/09/the-language-of-birds.html*.

Islahi, I. 1989] A. A. (1989). *Fundamentals of Hadith Interpretation – an English translat. Of "Mabadi Tadabbur-i-Hadith* (T. M. Hashmi, Trad.). Al-Mawrid.

Islam-beliefs. (s. d.). *Islam-beliefs. Prophetic Lineage in Quran. Https://islam-beliefs.net/en_US/prophetic-lineage-in-islam/. Last access in April 2023*.

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 14, number 1*.

Jankowski, M., Hajjar, F., Kawas, S. A., Mukaddam-Daher, S., Hoffman, G., McCann, S. M., & Gutkowska, J. (s. d.). *Jankowski, M., Hajjar, F., Kawas, S. A., Mukaddam-Daher, S., Hoffman, G., McCann, S. M., & Gutkowska, J. (1998). Rat heart : A site of oxytocin production and action. Proceedings of the National Academy of Sciences,*.

Juola, P. (2006). JGAAP: Authorship Attribution. *Foundations and TrendsR in Information Retrieval*, *1*(3), 233-334,.

Juola, P. (2009). JGAAP: A System for Comparative Evaluation of Authorship Attribution. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, *1*(1).

Kaheel, A. (2015). *The Marvels of number 7 in the noble Quran*. http://kaheel7.com/Book/Marvels_BookSeven.pdf.

Kalgutkar, V., Stakhanova, N., Cook, P., & Matyukhina, A. (2018). Android authorship attribution through string analysis. *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 1-10 ,.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & K.R.K. (2001a). Murthy, Improvements to Platt's SMO Algorithm for SVM Classifier Design". *Neural Computation*, *13*, 637-649,.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & K.R.K. (2001b). Murthy, Improvements to Platt's SMO Algorithm for SVM Classifier Design". *Neural Computation*, *13*, 637-649,.

Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., & Stein, B. (2020). Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In *CLEF*.

Kim 2003] Y. Kim, B. Moon. (2003). New Usage of Sammon's Mapping for Genetic Visualization. *Conference: Genetic and Evolutionary Computation - GECCO 2003, Genetic and Evolutionary Computation Conference*, 1136-1147.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(ue: 9), 1464-1480. https://doi.org/10.1109/5.58325

Kroopnick, Kroopnick, M. H., Chen, J., & Jaehwa Choi, C. M. D. (2010). Assessing Classification Bias in Latent Class Analysis : Comparing Resubstitution and Leave-Out Methods. *Journal of Modern Applied Statistical Methods. May*, *9*(1), 2-331 52-63.

Lachenbruch, P. A. (1967). *An almost unbiased method of obtaining confidence interval for the probability of misclassification in discriminant analysis* (p. 639-645).

Lammers, M. O., Au, W. W., & Herzing, D. L. (s. d.). *Lammers, M. O., Au, W. W., & Herzing, D. L. (2003). The broadband social acoustic signaling behavior of spinner and spotted dolphins. The Journal of the Acoustical Society of America, 114(3), 1629-1639.*

Lazarus, R. S., & Folkman, S. (1984). *Lazarus, R. S., & Folkman, S. (1984). Stress, appraisal, and coping. Springer publishing company.*

Li, J., Zheng, R., & Chen, H. (2006). From fingerprint to writeprint. *Communications of the ACM*, *49*(4), 76-82.

Love, H. (2002). *Attributing Authorship : An Introduction*. Cambridge University Press.

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005). Author identification on the large scale. *Joint Annual Meeting of the Interface and the Classification Society of North America (CSNA.*

McCowan, B., Hanser, S. F., & Doyle, L. R. (s. d.). *McCowan, B., Hanser, S. F., & Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems : Information theory applied to bottlenose dolphin whistle repertoires. Animal behaviour, 57(2), 409-419.*

McCraty, R. (s. d.). *McCraty, R. (2015). Science of the heart : Exploring the role of the heart in human performance. HeartMath Research Center, Institute of HeartMath.*

McCraty R, Atkinson M, Tomasino D, Bradley RT. (s. d.). *McCraty R, Atkinson M, Tomasino D, Bradley RT. The coherent heart : Heart–brain interactions, psychophysiological coherence, and the emergence of system-wide order*.

McLachlan 2003] G.J. McLachlan, R. W. Bean., D. Peel. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, *41*(ues 3–4), 379-388.

McLnchlan 2001] G. J. McLachlan, D. Peel., S. K. Ng. (2001). *On Clustering by Mixture Models,Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifikation e.V.* University of Munich.

McMenamin, G. R. (2002). *Forensic Linguistics—Advances in Forensic Stylistics*. CRC Press.

Mercer D. (s. d.). *Mercer D. 2022, Can I talk to my Hawk ? | Bird of Prey Communication—Mercer Falconry, https://www.youtube.com/watch?v=9mWzwWY1TuM, oct. 2022.*

Mills, D. E. (2003). *Authorship Attribution Applied to the Bible* [[Mills 2003]]. Graduate Faculty of Texas, Tech University.

Milne, & Milne, W. E. (2012). *Numerical Calculus—Approximations, Interpolation, Finite Differences, Numerical Integration, and Curve Fitting*.

Mohamed-Cherif, M. I. (2007). *Interogative styles in rethorical research and its secrets in the Quran* [PhD thesis,]. International Islamic University.

Mohd Yakub, Zulkifli Yusoff 2008, The Numerical inimitability in Quran : Its meaning and regulations “Inductive analytical study”. (2008). *Int. Conference on Numerical Inimitability in Quran 7-9 Nov. 2008*.

Mosteller, F., & Wallace, D. L. (1964). *Inference and Disputed Authorship : The*. Addison-Wesley.

Muangprathub, J., Kajornkasirat, S., & Wanichsombat, A. (2021). Document Plagiarism Detection Using a New Concept Similarity in Formal Concept Analysis. *Journal of Applied Mathematics*.

Muqith M. A. (2011). *Muqith M. A., 2011 Al-Madd Al-Tabee’ee. Heesbees : All for Quran and Tajweed. January 2011. Https://heesbees.wordpress.com/tag/short-and-long-arabic-vowels/*.

Naghich, N. (2012). Interogative style in the Prophetic statements. *MSc Thesis*.

Nasr, S. H. (2004). *The heart of Islam : Enduring values for humanity* (Harper, SanFrancisco, c2002).

Norusis 2008], M. Norusis. (2008). Cluster Analysis, Chapter 16.,pp:361-391. SPSS 17.0 Statistical Procedures Companion. *Marija Norusis*.

Ouamour, Ouamour, S., & Sayoud, H. (2013). Authorship Attribution of Short Historical Arabic Texts Based on Lexical Features. *CyberC – International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery CyberC Conference*.

Peng, Peng, F., Schurmans, D., Keselj, V., & Wang, S. (2003). Language independent authorship attribution using character level language models. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 267-274,.

Powers D. S. (2011). *Powers D. S., 2011. Review : La transmission écrite du Coran dans les débuts de l’islam : Le codex Parisino-petropolitanus. Texts and Studies on the Qurʾān, 5 by François Déroche. Islamic Law and Society, Vol. 18, No. 2 (2011), pp. 281-285.*

Quran Project. (s. d.). *Old and New Testament Prophecies of Muhammad, https://www.quranproject.org/Quran-Project-Appendix-Old-and-New-Testament-Prophecies-of-Muhammad—227-d, Last access in April 2023*.

Ramsey B. C., Higham T. F. G., Brock F., Baker D., & Ditchfield P. (2009). *Ramsey B. C., Higham T. F. G., Brock F., Baker D., & Ditchfield P., (2009). Radiocarbon dates from the Oxford AMS system : Archaeometry datelist 33. Archaeometry, 51(2), 323-349. Archaeometry, 51(2), 323-349.*

Rogers L. J. & Kaplan G. (s. d.). *Rogers L. J. & Kaplan G.. (2002). Songs, roars, and rituals : Communication in birds, mammals, and other animals. Harvard University Press, 2002*.

Sadeghi, B., & Bergmann, U. (2010). *Sadeghi, B., & Bergmann, U. (2010). The Codex of a Companion of the Prophet and the Qurān of the Prophet. Arabica, 57(4), 343-436.*

Saged, A.A.G., Mohd Yusoff, M.Y.Z., Abdul Latif, F. et al. (2020). *Saged, A.A.G., Mohd Yusoff, M.Y.Z., Abdul Latif, F. et al. Impact of Quran in Treatment of the Psychological Disorder and Spiritual Illness. J Relig Health 59, 1824–1837 (2020). Https://doi.org/10.1007/s10943-018-0572-8*.

Sanaa. (1985). *Sanaa, 1985. Maṣāḥif Ṣanʿā', 1985, Dār al-Athar al-Islamiyyah : Kuwait, p. 59, Plate 4.*

Sanderson, C., & Guener, S. (2006). Short text authorship attribution via sequence kernels, markov chains and author unmasking : An investigation. *Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP*, 482-491.

Sari, Y., Stevenson, M., & Vlachos, A. (2018). Topic or style ? Exploring the most useful features for authorship attribution. *Proceedings of the 27th International Conference on Computational Linguistics*, 343-353 ,.

Sarwar, R., Urailertprasert, N., Vannaboot, N., Yu, C., Rakthanmanon, T., Chuangsuwanich, E., & Nutanong, S. (2020). $ CAG $: Stylometric Authorship Attribution of Multi-Author Documents Using a Co-Authorship Graph. *IEEE Access*, *8*, 18374-18393,.

Sayoud, H. (2003). *Automatic speaker recognition – Connexionnist approach"* [PhD thesis,]. USTHB University.

Sayoud, H. (2012). Author Discrimination between the Holy Quran and Prophet's Statements. *LLC Journal, Literary and Linguistic Compting*, *27*(4), 427-444,.

Sayoud, S. 2010] H. (2010). Investigation of Author Discrimination between two Holy Islamic Books. *IET (Ex-IEE) Teknologia Journal*, *1*(ue. 1), ,.

Sayoud, S. 2012] H. (2012). Author discrimination between the Holy Quran and Prophet's statements. *Literary and Linguistic Computing*, *27*(4), 427-444. https://doi.org/10.1093/llc/fqs014

Schönrogge, K., Barbero, F., Casacci, L. P., Settele, J., & Thomas, J. A. (s. d.). *Schönrogge, K., Barbero, F., Casacci, L. P., Settele, J., & Thomas, J. A. (2017). Acoustic communication within ant societies and its mimicry by mutualistic and socially parasitic myrmecophiles. Animal Behaviour, 134, 249-256.*

Schuster, T., Schuster, R., Shah, D. J., & Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, *46*(2), 499-510,.

Shlens, J. (2003). A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS - Derivation, Discussion and Singular Value Decomposition. *J. Shlens*, *1*.

Signoriello, D. J., Jain, S., Berryman, M. J., & Abbott, D. (2005). Advanced text authorship detection methods and their application to biblical texts. *Proceedings of SPIE, 6039*, 163-175.

Spiess, S., Mylne, H. K., Engesser, S., Mine, J. G., O'Neill, L. G., Russell, A. F., & Townsend, S. W. (s. d.). *Spiess, S., Mylne, H. K., Engesser, S., Mine, J. G., O'Neill, L. G., Russell, A. F., & Townsend, S. W. (2022). Syntax-like Structures in Maternal Contact Calls of Chestnut-Crowned Babblers (Pomatostomus ruficeps). International Journal of Primatology, 1-20.*

Stamatatos, E. (2007). Author identification using imbalanced and limited training texts. *Proceedings of the 4th International Workshop on Text-Based Information Retrieval*, 237-241.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science & Technology*, *60*(3), 538-556.

Stamatatos, Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, *35*(2), 193-214,.

Stylianou, Y., & et al. (2005). *Y. Stylianou, Y. Pantazis, F. Calderero, P. Larroy, F. Severin, S. Schimke, R. Bonal, F. Matta, , and A. Valsamakis,"GMM- Based Multimodal Biometric Verification," Final Project Report 1, Enterface'05, July 18th—August 12th, Mons, Belgium, 2005.*

Suganya, S. 2012] R., & R. Shanthi, F. C.-M. A.-. (2012). *A Review, International Journal of Scientific and Research Publications*, *2*(ue 11).

T. Kardi, K-Means Clustering Tutorial. E-Book by Teknomo. (2007). *Kardi*. http://people.revoledu.com/kardi/tutorial/kMean/index.html

Tambouratzis, G. (2004). Discriminating the Registers and Styles in the Modern Greek Language-Part 1 : Diglossia in Stylistic Analysis. *Literary and Linguistic Computing, 19(2), 197-220.*, 197-220.

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Tambouratzis, D., & Carayannis, G. (2000). Discriminating the Registers and Styles in the Modern Greek Language. In *Proceedings of the Workshop on Comparing Corpora (held in conjunction with the 38th ACL Meeting* (p. 35-42).

Tambouratzis, G., Markantonatou, S., Vassiliou, M., & Tambouratzis, D. (2003). Employing Statistical Methods for Obtaining Discriminant Style Markers within a Specific Register. In *Proceedings of the Workshop on Text Processing for Modern Greek : From Symbolic to*

Statistical Approaches (held in conjuction with the 6th International Conference of Greek Linguistics), Rethymno (p. 1-10).

Taylor R.E. and Aitken M.J. (1997). *Taylor R.E. and Aitken M.J. 1997. Chronometric dating in Archaeology. Advances in Archaeological and Museum Science, volume 2. Oxford University, England.*

Tchechmedjiev, A., Schwab, D., & Goulian, J. (2013, mars 24). Fusion strategies applied to multilingual features for an knowledge-based word sense disambiguation algorithm : Evaluation and comparison". *CICLING' 2013 Conference: 14th International Conference on Intelligent Text Processing and Computational Linguistics*.

UNESCO. (1980). *UNESCO, 1980. « Recommendations » in R. B. Serjeant (Ed), The Islamic City, 1980, UNESCO: France, pp. 207-208.*

Verlinde, P. (1999). *A Contribution to Multimodal Identity Verification using Decision Fusion* [PhD thesis,]. Ecole Nationale Supérieure des Télécommunications.

Von Frisch, K. (s. d.-a). *Von Frisch, K. (1965). Die Orientierung der Bienen unterwegs zum Ziel. Tanzsprache und Orientierung der Bienen, 331-537.*

Von Frisch, K. (s. d.-b). *Von Frisch, K. (1965). Die tänze der bienen (pp. 3-330). Springer Berlin Heidelberg.*

Whiteman, & Whiteman, M. (1967). *Philosophy of Space and Time and the Inner Constitution of Nature : A Phenomenological StudyRelié– 1967*. Humanities Press.

Wiki1. (2012). *Quran. The free encyclopedia. Wikipedia, Last modified in*. http://en.wikipedia.org/wiki/

Wiki2. (2012). *Hadith. The free encyclopedia. Wikipedia, Last modified in*. http://en.wikipedia.org/wiki/

Wiki_COS. (2013). *Wikipedia, Cosine similarity, From Wikipedia, the free encyclopedia. The web page was last modified on*. http://en.wikipedia.org/wiki/Cosine_similarity.

Wiki_REG. (2013). *Linear regression, From Wikipedia, the free encyclopedia. The web page was last modified on*. http://en.wikipedia.org/wiki/Linear_regression.

Witten, I., Frank, E., Trigg, L., Hall, M., Holmes, G., & Cunningham, S. J. (1999). Weka : Practical machine learning tools and techniques with Java implementations. In N. Kasabov & K. Ko (Éds.), *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems* (p. 192-196,).

Zangerle, E., Mayerl, M., Specht, G., Potthast, M., & Stein, B. (2020). *Overview of the style change detection task at PAN 2020*. CLEF.

Zouhir, A., El Ayachi, R., & Biniz, M. (2021). A comparative Plagiarism Detection System methods between sentences. *Journal of Physics*, *1743*(1), 012041.

الحمد لله رب العالمين

Book Edited by Pr Halim Sayoud

Publisher EDT - SCHOLARPAGE

DOI: 10.5281/zenodo.14865663

February 2025