Application of Computational Stylistics and Text Mining Techniques to Identify and Compare Salient Features of Different English Translations of the Holy Quran

Muhammad Badruddin Khan

Information Systems Department College of Computer and Information Sciences Imam Mohammad ibn Saud Islamic University (IMSIU), Riyadh, KSA

Abstract

The presence and role of religion in human life and history is an undisputable phenomenon. The religious text provides foundation for understanding the meaning of life and offers answers of fundamental questions related to human life. Religious text is claimed to be revealed in a particular language on some holy person and later, it was translated and interpreted in multiple languages. Since the text can have multiple meanings, the translation and interpretation depend on the background, style and understandings of the person who is translating and interpreting the text. Quran, the Holy book of the Muslims is a unique literary masterpiece in terms of its linguistics and style. More than forty English translations of the original Arabic text appeared over the past two centuries. These translations of the Holy Quran vary widely in terms of linguistics and style. Various translators adopted different strategies in translating the Quranic Arabic to English language. Comparative studies by linguistic researchers identified differences among these translated works in terms of translating Arabic figures of speech, homonyms, metaphors, polysemous words, and overall loyalty and preciseness to the Arabic text. The syntactic and semantic features of each translation also varies greatly. Existing comparative studies are mostly limited in nature with regard to the number of translation aspects compared and the size of corpus used for the study. These limitations are primarily due to the linguistic complexity and the sheer large volume of translated works of the Quranic text. The rise of computing power as well as artificial intelligence and machine learning advancements have resulted in crossing many limits. Text mining with its set of linguistic, statistical and machine learning techniques can help overcome the limitations faced by the traditional linguistic researchers. In this paper, computational stylistics and text mining techniques were utilized to perform a detailed comparative study on a large number of available translations of the Holy Quran. The quantification of literary works to measure style and compare translations was performed using internationally recognized metrics. Cosine similarity was used to discover textual level of similarity.

Key words: Text Mining, Computational stylistics, Cosine Similarity, Religious text, Holy Quran

1. Introduction

Muslims believe that The Holy Quran was revealed by Allah (SWT) in Arabic and it was sent for the guidance of whole mankind. In order to deliver its message to non-Arabic speaking nations, it was translated into numerous languages. English is the first language of almost 375 million people in this world and has its prominence as the leading language for international discourse. In 1734, an orientalist from England named George Sale produced the first translation of the Holy Quran directly from Arabic into English. In 1930, the English convert to Islam, Muhammad Marmaduke Pickthall's translated Holy Quran into English. Till present there are a number of English translations by different learned scholars. In this paper, different translations of the Holy Quran were compared based on their literary style by combining well established computational literature and text mining tools. What is expected to be discovered is as follows:

- Difference in literary style of each translator
- The literary style of translated work in relation to the changing nature of English of the past century.

13 famous English translations of the Holy Quran were selected as input data and computing power was applied to analyze the text. In Section 2 of this paper the related works are described, followed by the description of the text corpora and relevant text processing techniques in Section 3. In Section 4 the framework on the basis of which the presented work was implemented, is described. Finally, Section 5 concludes the paper after describing the experiments and their brief interpretations.

2. Background

Comparative translation has been traditionally identified as an important area of research [1]. In comparative translation studies, detailed comparative analysis in terms of literary
criticism, stylistics, and cognitive linguistics of various
translations of the same text are made [2]. For example, the
authors in [3] has made a comparative study of different
translation styles in court proceedings and their impact in
multilingual interaction in the court. In [4] attempted to
identify the stylistic traits of two English translators of some
Spanish and Portuguese fictions. And, in [5] the author
presented a comparative study of literary translations from
Arabic to English and French.
The Holy Quran, the most revered religious text in Islam is
a literary master-piece with regard to its linguistic style and
depth of content. An accurate and faithful translation of the
original Arabic text of the Quran is generally considered to
be a difficult task [6]. Nonetheless, it is being translated to
English and other languages of the world even fourteen
hundred years after its revelation. Over the past two
centuries, more than forty different English translations of
the Quran were published [7], [8]. Some of the notable
translation works are [9]–[11]. Various translators adopted
different strategies and styles in translating the Quranic
Arabic to the English language, making it an interesting
subject for the field of comparative translation studies.
Comparative studies by linguistic researchers identified
differences among these translated works in terms of
various syntactic and semantic features of each [6], [12]–
[22]. Dastjerdi and Jamshidian in in [15] investigated the
translation translation strategies of only two English translations of
the Quran. Rasekh et.al [12] looked into the different
translation strategies of four English translators with regard
to translating homonymous expressions in the Quran. Alqinai [14] explored the difference in translations of
connotations and synonyms in the Quran for five different
translators. Partini [17] analysed the usage of figurative
language in translating only two chapters of the Quran.
Unfortunately, most of these comparative studies are
limited in nature with regard to the number of translation
aspects compared and the size of corpus used for the study.
These limitations are primarily due the linguistic
complexity and the sheer large volume of the translated
works of the Quranic text. Natural language processing
(NLP) and text mining tools and technologies can help
overcome the limitations faced by the traditional linguistic
researchers [23]–[26]. NLP and text mining makes it
possible to extract and process specific linguistic features
from very large volumes of texts. Automatic detection of
similarity among multiple documents using different
approaches is discussed by [35]–[37]. In recent years, NLP
and text mining approaches have been successfully used in
wide variety of linguistic research. For example, Zaho in
[27] performed a stylistic analysis of Virginia Woolf’s
novel To the Light-house to identify the unique features of
Woolf’s writing. In the field of translation studies, Qin and
Wang [28] used NLP to study the linguistic features of
English to Chinese translations. Li et. al [29] analysed the
translation style of two English translation of Hongloumeng,
a classical Chinese novel by Xueqin Cao. Similar to these
works, we propose to utilize NLP and text mining to
perform a detailed comparative study on a much larger
scope involving all the publicly available translations the
Holy Quran written over a period of two centuries. Unlike
most of these researches, apart from exploring various
linguistic features of these different translations, the
proposed research will also delve into relating the impact of
the evolution of English language and its readership over
the past century and the translators’ choice of a specific
style.

3. Text Corpus and its Processing

3.1 Quran Translation Corpus

13 English translations of the Holy Quran written by
different authors were collected for this study. The collected
translations were then converted and saved in a uniform
plain-text format to be used in our comparison framework.
Table 1 lists the 13 translated works and provides some
details of each translation with regards to the year of
publication, the name of the author, and whether the author
is a native English speaker or not along with their
nationality (where known).

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Year</th>
<th>Native English Speaker?</th>
</tr>
</thead>
<tbody>
<tr>
<td>q1</td>
<td>Ahmed Ali</td>
<td>2001</td>
<td>No (India)</td>
</tr>
<tr>
<td>q2</td>
<td>Arthur John Arberry</td>
<td>1955</td>
<td>Yes (UK)</td>
</tr>
<tr>
<td>q3</td>
<td>Abdul Majid Daryabadi</td>
<td>1957</td>
<td>No (India)</td>
</tr>
<tr>
<td>q4</td>
<td>Maulana Muhammad Ali</td>
<td>1917</td>
<td>No. (Pakistan)</td>
</tr>
<tr>
<td>q5</td>
<td>Muhammad Sarwar</td>
<td>1981</td>
<td>No. (Pakistan)</td>
</tr>
<tr>
<td>q6</td>
<td>Hamid S. Azz</td>
<td>2009</td>
<td>No.</td>
</tr>
<tr>
<td>q7</td>
<td>Firdulu Haq, from urdu</td>
<td>1990</td>
<td>No. (Pakistan)</td>
</tr>
<tr>
<td>q8</td>
<td>Mohammad Habib Shakir</td>
<td>1970</td>
<td>No (Egypt)</td>
</tr>
<tr>
<td>q9</td>
<td>Abdullah Yusuf Ali</td>
<td>1934</td>
<td>No (India)</td>
</tr>
<tr>
<td>q10</td>
<td>Muhammed Marmaduke William Pickthall</td>
<td>1930</td>
<td>Yes (UK)</td>
</tr>
<tr>
<td>q11</td>
<td>Ali Unal</td>
<td>2006</td>
<td>No (Turkey)</td>
</tr>
<tr>
<td>q12</td>
<td>Amatul Rahman Omar (Urdu) – retranslated to English by Allamah Nooruddin</td>
<td>2005</td>
<td>No. (Pakistan)</td>
</tr>
<tr>
<td>q13</td>
<td>Muhamed Ahmed &amp; Samira</td>
<td>1994</td>
<td>No.</td>
</tr>
</tbody>
</table>

*ID: The ID that is used in this report for identifying each translation

3.2 Text processing

Number of text processing steps were performed as given
below. All the translated documents were initially stored as
plain text documents. These documents were used in two
separate blocks to generate the required processed
documents for analysis. These two building blocks were
independent of each other and hence do not require to be
completed in sequence. The first set of processing block consisted of the following steps:

a. Remove numbers
b. Remove punctuations
c. Stemming
d. Remove stop words

The other text processing block performs the following steps:

a. Tagging
b. Lemmatization

4. Framework

4.1 Comparison Framework

In order to objectively compare various translation works of the Holy Quran a new framework was proposed and used. The framework consisted of text mining and quantitative analytical components. Figure 1 presents a sketch of the constructed framework. The components are described in subsequent sections.

The quantification of literary works to objectively measure style and compare translations requires internationally recognized metrics. The metrics will be described in one of the later sections of this paper. These metrics will be used in context of demonstration of the new framework as well as the experiment results.

Fig. 1 Quran translations lexical comparison framework

4.2 Computational analysis

In this work, a quantitative approach to objectively compare literary works, was adopted. The work relies on quantifiable linguistic properties of the literary works for this purpose. The following linguistic properties were computed on the output generated by the previous component of the comparison framework:

- Readability
- Lexical diversity
- Similarity

Each of these linguistic properties can be measured by a set of predefine indexes. The indexes will act as the building blocks of the computational analysis component of the framework to identify and compare the literary styles of various translated works. Here the explanation about these indexes is given in detail:

- **Readability.**

  Readability is the ease with which the reader can understand a literary work. Readability of a text depends on the complexity of the words and sentence structure. There are various indexes to measure readability. For this research
work, readability was measured using the following indexes:

a. **Flesch Readability Ease Index**: The Flesch Reading Ease index is one of the most popular and widely used readability index [30,31]. It attempts to predict the difficulty of a text. It is expressed on a scale of 0 to 100 with 0 equivalent to 12th grade of formal education and 100 equivalent to 4th grade education. Flesch Readability Ease is computed by:

\[
FRI = 206.835 - \left( \frac{1.015 \times W}{St} + \frac{84.6 \times SY}{W} \right)
\]

where, \(W\) is the number of words, \(St\) is the number of sentences, and \(SY\) is the number of syllables. The higher the value of the index the easier the document will be to read.

b. **Gunning FOG Index**: Gunning Frequency of Gobbledygook (FOG) index measures the amount of difficult to understand “fog” in a given text. It is given by:

\[
4 \times \left( \frac{W}{St} + \frac{100 \times W_{3Sy}}{W} \right)
\]

where, \(W\) is the number of words, \(St\) is the number of sentences, and \(W_{3Sy}\) is the number of words with at least 3 syllables. The higher the value of the index the more difficult the document will be to understand.

c. **Björnsson’s Läsbarhet Index (LIX)**: LIX is given by:

\[
\text{LIX} = \frac{W}{St} + \frac{100 \times W_{7c}}{W}
\]

where, \(W\) is the number of words, \(St\) is the number of sentences, and \(W_{7c}\) is the number of words with at least 7 letters. Text with LIX < 25 are considered very easy to read, LIX around 40 is normal, and LIX > 55 are very difficult to read.

d. **Linsear Write Index**: The Linsear Write index is given by:

\[
\text{LWI} = 100 - \left( \frac{100 \times W_{cSy}}{W} \right) - \left( \frac{3 \times (100 \times W_{3Sy})}{W} \right)
\]

where, \(W\) is the number of words, \(St\) is the number of sentences, and \(W_{cSy}\) is the number of words with less than 3 syllables. The higher the value of the index the more difficult the document will be to read.

- **Lexical Diversity**.
  Lexical diversity of a document is the measure of richness of vocabulary in a given text [32]. High lexical diversity of a text usually positively correlates to higher linguistic competency of the author. In this work, Lexical diversity was measured using the following indexes:

a. **Type Token Ratio**: Type Token Ration (TTR) is the relationship between the total number of words (tokens) and the number of unique word stems (types). TTR is given by:

\[
TTR = \frac{V}{N}
\]

where, \(N\) is the number of tokens and \(V\) is the number of types.

b. **Uber Index**: Dugast’s Uber Index is given by:

\[
\frac{(\log N)^2}{\log N - \log V}
\]

where, \(N\) is the number of tokens and \(V\) is the number of types.

c. **Yule’s K Index**: This index measures the concentration of high frequency words in a text. A high value of K implies that the authors vocabulary is highly concentrated on a few repeated words, where as a relatively smaller K value implies a well diverse set of vocabulary. Yule’s K is measured by:

\[
K = \frac{10^{4.6 \times (\sum_{x=1}^{n} f_x x^2) - N}}{N^2}
\]

where, \(N\) is the number of tokens, \(X\) is a vector of the frequencies of each type and \(f_x\) is the frequencies for each \(X\).

d. **HD-D Index**: The Hypergeometric Distribution of Diversity (HD-D) index approximates the rate of change of TTR as the text grows in length. For each unique word stem, the probability is computed (using the hypergeometric distribution) of drawing it at least one time when drawing 42 tokens from the text randomly. The sum of these probabilities is the HD-D value.

- **Similarity**.
  Similarity of two documents, i.e., in this research work, two translations of the Quran is measured by computing Term Frequency-Inverse Document Frequency (TF-IDF) vectors for each document and calculating the cosine similarity of the resulting vectors. Term frequency (TF) of a term \(t\) in document \(d\) is obtained by finding the frequency \(y\) in \(d\). So,

\[
TF(t, d) = f(t, d)
\]
Inverse Document Frequency (IDF) is the ratio of total number of documents to the number of documents containing the term $t$. Mathematically,

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

(9)

where, $N$ is the total number of documents and $|\{d \in D : t \in d\}|$ is the number of documents where $t$ appears. TF-IDF is then given by:

$$tf(t, d) \times idf(t, d)$$

(10)

The cosine similarity of two translations represented as TF-IDF vectors is then computed using:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

(11)

where, $A$ and $B$ are two vectors representing two different translations of the Quran; $A_i$ and $B_i$ are the elements of vector $A$ and $B$ respectively; and $n$ is the number of elements in each vector.

5. Experimental Results

5.1 Comparison Based on Readability

Readability indexes provide an insight to the structure of the sentences and their perceived level of difficulty. Texts with similar level of readability index indicate similarity of writing style in terms of sentence structure and meaning. Figure 2, 3, 4, and 5 present the comparison of different Quranic translations with respect to four different readability indexes.

Gunning FOG and Linsear Write indexes estimate the years of formal education needed to understand the text on a first reading. An index value of 6 indicates the text requires a reading skill level of a native English speaker who has completed the primary education (Grade 6). As shown in Figure 3 and Figure 4, the translation by Ahmad & Samira (q13) and Nooruddin (q12) requires about 12 years of formal education, in contrast with the translation of Pickthall (q10) which requires less than 6 years of formal education. The results are in accordance with findings from Flesch Readability Ease Index. Most other translations such as the ones by M. Sarwar(q5), S. Aziz(q6), or F. Haq(q7) require 8-10 years of formal education.
Similar levels of readability are indicated by the LIX index in Figure 5. This index again shows that translation by Pickthall (q10) is very easy to read.

5.2 Comparison based on Lexical Diversity

High level of lexical diversity of a literary work is an indication of literary competency of the author and the richness of the content. In this paper, the selected translations were compared based on Type Token Ratio (Figure 6), Uber Index (Figure 7), Yule’s K index (Figure 8), and HD-D index (Figure 9). In research findings, the TTR did not vary much with a range from 0.03 (the translation of Mohammad Habib or q8) to 0.05 (the translations of Ahmed Ali or q1 and Muhamed Ahmed and Samira or q13). Uber Index provides a better measure of lexical diversity with respect to the length of text such as in this study. As shown in Figure 7, the translation with lowest lexical diversity is the translation by Mohammad Habib (q8) whereas, the translation by Muhamed Ahmed and Samira (q13) has the highest diversity. As with TTR and Uber index, the translation of Mohammad Habib (q13) showed relatively low diversity in terms of Yule’s K and HD-D indices.

5.3 Comparison Based on Similarity
The similarity figure (Figure 10) was created by mapping the similarity indexes (using Equation 9) of each translation on a two-dimensional plot using the multi-dimensional scaling algorithm presented in [33] and [34]. The two-dimensional plot reveals few interesting similarities. For example q12 and q13 are similar. Moreover q6 and q8 become another pair based on similarity. Another similar translations are q2 and q7. One possible reason may be q2 (1955) was the standard of translation for q7 (1990). The translation by Pickthall (q10) and Yusuf Ali (q9) were written during the same time period (1930’s), but they are very different based on their textual features. This is interesting because in spite of this dissimilarity, their readability and lexical diversity indexes are very similar.

6. Conclusions

The unstructured data in its digital form is one of the venue where the application of various metrics and machine learning techniques is possible. Numerous translations of the Holy Quran are available in English language and this paper presents the work that has used them for analytical purposes. With the passage of time, languages also evolve and therefore the readability easiness of old translation may differ from a new translation. In this work, different translations were compared with respect to different indexes and some interesting insights were discovered. Similarities of different translations were computed using cosine similarity from machine learning domain and were reported graphically in this paper. The work can be extended further by using different available lexical databases like WordNet. By the combination of lexicons and machine learning techniques, semantic similarities of translations can be discovered that will help in understanding different translators’ preferences and make clusters of English translations based on semantics of the religious text.

References


Dr. Muhammad Badruddin Khan obtained his doctorate in 2011 from Tokyo Institute of Technology, Japan. He is a full-time assistant professor in department of Information Systems of Al-Imam Muhammad Ibn Saud Islamic University since 2012. The research interests of Dr. Khan lie mainly in the field of data and text mining. He is currently involved in number of research projects related to machine learning and Arabic language including pandemics prediction, Arabic sentiment analysis, improvement of Arabic semantic resources, Stylometry, Arabic Chatbots, trend analysis using Arabic Wikipedia, Arabic proverbs classification, cyberbullying and fake content detection, and violent/non-violent video categorization using Youtube video content and Arabic comments, and has published number of research papers in various conferences and journals. He is also co-author of a book on machine learning.