
Proceedings Paper:
Alrehaili, SM orcid.org/0000-0002-4957-2478 and Atwell, E orcid.org/0000-0001-9395-3764 (2014) Linguistics features to confirm the chronological order of the Quran. In: Second Workshop on Arabic Corpus Linguistics. Second Workshop on Arabic Corpus Linguistics, 22 Jul 2013, Lancaster, UK. UCREL , pp. 62-65.

Linguistics features to confirm the chronological order of the Quran

Sameer Alrehaili, Eric Atwell

University of Leeds
salrehaili@gmail.com

1 Introduction

This paper reports a corpus-based study of the chronological order of the chapters in the Arabic Quran. Unlike many text analysis tasks, analysis of the chorological order of a set of texts is about analysing how the text was written instead of what is being said; analysing the style rather than the content. The task involves dividing the texts into groups, then decided which group was written before others.

The aim is to find linguistics features in the texts of the Quran that are related to time. In other words, we want to label the Quran verses to the phase when they were revealed.

2 The order of texts in the Quran

There is a consensus among Muslims that the Arabic text of the Quran came from Allah, revealed to the Prophet Muhammad through visitations from his messenger the angel Gabriel over many years. The accepted standard order of suras or chapters in the Quran is not arranged according to the date in which the suras were revealed or even the place where they were revealed (Ali 2002; Akbar 2002). There is an agreement among scholars that the order of texts within the Quran was arranged by the Prophet Muhammad following Allah's command. He was instructed to put each text in a specified location. Although the chapters have been arranged in an order that is different from the sequence of revelation, scholars do not say that it has been arranged in the wrong way because it was revealed to respond to various events and incidents.

3 Dividing the corpus

We chose a copy provided by (Tanzil project 2007) due to it is verified manually and automatic. An XML file from Tanzil used with the Jquran Tree library from (Dukes 2011). This library allows us to access the Quran with several formats such as diacritics, removed diacritics and Buckwalter transliteration.

The Islamic scholar Bazargan proposed a "block scheme" to rearrange the Quran text into chronological units or blocks (Sadeghi 2011). A block is a set of texts that are believed to belong to the same time period. A Sura can be divided into one block or more, but a block cannot have verses from different Suras. We also divided the corpus into 3 different possible arrangements suggested by previous researchers. The first one arranges the text into 194 blocks; the second division is 22 phases, which is the first chronology proposed by Bazargan, while the third is the modified Bazargan division into 7 phases. Some Suras have been divided into several blocks and others taken intact. Therefore we encoded the specific order in an index file and read the text numbers in each block from that file.

```
1   1,96,1,5
2   2,74,1,7
3   3,103,1,2
4   4,51,1,6
5   5,102,1,2
6   6,52,1,8
7   7,112,1,4
8   8,88,1,5
9   8,88,8,16
10  9,86,11,17
```

Figure1: Example reordering of texts in blocks.

Then, we extracted a set of style markers to observe the behaviour of these markers over an ordered series of blocks. Frequencies of a range of style markers were computed, such as co-occurrence of key words, common letters and diacritics used in Arabic, frequencies of key concepts, parts-of-speech, and most frequent morphemes.

4 Generate markers of style

Decision of choosing markers required text-mining and some searching. Therefore, we did text-mining to get the most frequent words. We also did Background searching about this and found that according to (Jaffer and Jaffer 2009 ; Ahmed 2008) Meccan Suras tend to be short, whereas the Medinan tend to be long. In addition, (Sadeghi 2011) shows that a mean verse length has a gradual increasing against the time. Therefore, the mean verse length (MVL) can be used as a style marker. MVL can be calculated using the following formula:

$$\text{MVL} = \frac{Total\ number\ of\ words\ in\ the\ phase}{Total\ number\ of\ words\ in\ all\ phases}$$

As our pre-processing shows that the word of "Allah" has occurred about 2153 times in the Quran,

we think that if we take it as a marker of style would present an important style due to it occurs almost in all Suras. Furthermore, we also take the other names that belong to Allah in the Quran as a marker.Assume we have a vector V of words that refer to Allah. To compute the frequencies of these names is as shown in the below:

Allah names frequencies in phase "i" equal to

$$= \frac{\sum_{k=0}^{n} V_k \; in the \; phase \; i}{Total number of verse in the phase \; i}$$

Relative frequency of a tag is the number of occurrences for this tag in the text divided by total number of tags. An example of computing relative frequency of Noun in the verse number can be shown as the following: the first verse in the Quran, (bisomi {ll¬ahilr¬aHoma'nlr¬aHiymi) it is clear that it has 4 words and 5 morphemes and 3 types of part-of-speech tagsets. (bisomi) composed of prefixed preposition + noun. ({ll¬ahi) composed of proper noun. (lr¬aHoma'n) has an adjective. (lr¬aHiymi) also has an adjective. The relative frequency of noun for this verse is 1/5.

So, the relative frequency of noun in phase number "i" is equal to =

$$\frac{Total number of noun in the phase \; i}{Total number of all \; tags \; in \; the \; phase \; i}$$

Arabic diacritics are very important because they can change the meaning of the text. An example of that can be seen in the words {ذَهَبَ, ذَهَب}, the first word means went while the second means a gold. Relative frequency of most common diacritics also used as a style feature as the following. Calculate the number of a vowel that we want to get it relative frequency in a specific group of text divided by the number of other all vowels in the same group.

For example, Table 1 shows key vowels and their overall frequencies in the Quran.

No	Vowel	Frequency	Description
1	a	122948	Fatha,Sound equivalent to a
2	i	45970	Kasrah, Sound like I or e
3	u	37320	Dmmah, Sound like u or o
4	aa	15955	Sound aa
5	ii	4194	Sound ii
6	final an	3741	Double Fatha
7	final in	2633	Double Kasrah
8	final un	2519	Double Dammah
9	uu	2034	Sound uu

Table 1: 9 vowels used as style metrics.

5 Representation

In order to represent the style of texts in terms of frequencies of stylometric markers within different temporal groups of text, we constructed a database with two tables. The first is Chapters or Suras. The second table is the Verses (Aya). Here we recorded 6236 verses along with several markers and orders. An example can be seen in Figure 2; Marker1, Marker2, and Marker4 are word counts, and the symbol of Fatha and Kasrah have been computed for each verse. Order5 is the revelation order adapted from (Tanzil 2007). Order33 field is the order first proposed by Bazargan and order44 the 7-phases order or Bazargan modified order.

idVerses	Marker1	Marker2	Marker4	order5	order33	order44
1	4	4	6	5	3	2
2	4	6	4	5	3	2
3	2	3	3	5	3	2
4	3	2	5	5	3	2
5	4	8	3	5	3	2
6	3	5	3	5	3	2
7	9	17	7	5	3	2
8	1	0	0	87	8	5
9	7	8	6	87	8	5
10	8	15	7	87	8	5

Figure2: example different orders and markers for each verse.

Now, we can use SQL to produce these combinations of ordered markers. For example, assume we want the style of Marker1 with the order44; we only need to write the following SQL statement.

```
select order4, sum(Marker1) from verses group by order4;
```

6 Features that support the 7-phases

No	Fatha	Damma	Kasrah
1	6.2784	1.4744	2.0256
2	7.7691	2.0829	2.8564
3	10.574	3.1684	3.854
4	13.3318	3.8167	4.8654
5	21.1552	6.0359	7.7309
6	27.7882	8.9608	10.7717
7	38.237	12.3888	14.2328

Table2: 3 vowel style markers increase monotonically

morphemes	concept of Allah	Allah	related verse
5.4773	0	0.02	0.3608
7.0674	0.0298	0.04	0.6939
10.6673	0.0426	0.05	0.645
13.8051	0.0696	0.05	0.9072
22.1818	0.1066	0.29	1.3171
29.7479	0.1365	0.88	1.7625
43.8586	0.0894	1.34	1.79

Table 3: Another 4 style markers increase

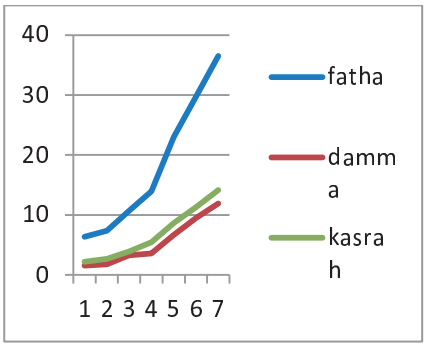

Figure3: relative frequencies of 3 stylometric features over reversal 7 phases order

As a result of this study, an API has been produced to test markers computed during the implementation of the project, as well as a web user interface to make the experiments available for interested researchers. The basic markers include key word counts, mean verse length, three most common vowel symbols in Arabic, the number of morphemes, and the occurrence of "Allah"; these support the accepted text chronology. We also investigated a novel type of feature which gave frequency distributions inconsistent with the accepted chronology: conceptual semantic metrics such as the concept of "Allah" (including variants of the name). Other applications can be tested using these markers like the Arabic poetry.

References

A. J. a. M. Jaffer, "Quranic Sciences", London: ICAS Press, 2009.

B. Sadeghi, "The Chronology of the Qurān: A Stylometric Research Program," Arabica, pp. 210-299, 2011.

K. Dukes, "Java API -Quran Java API," 2011. [Online]. Available: http://corpus.quran.com/. [Accessed 01 05 2012].

M. Ahmad, "Statistical profile of Holy Quran and Symmetry of Makki and Madni Suras," Pakistan Journal of Commerce and Social Sciences, pp. 1-16, 2008.

M. M. Akbar, "Authenticity of Quran", Niche of Truth, 2002.

M. M. Ali, "Holy Quran: English Translation and Commentary", U.S.A: Ahmadiyya Anjuman Isha'at Islam Lahore Inc, 2002.

"Tanzil Quran Navigator," 2007. [Online]. Available: http://tanzil.info/. [Accessed 01 06 2012].

Appendix

```
<quran>
  <sura index="1" name="الفاتحة">
    <aya index="1" text="بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ" />
    <aya index="2" text="ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ" />
    <aya index="3" text="ٱلرَّحْمَٰنِ ٱلرَّحِيمِ" />
    <aya index="4" text="مَٰلِكِ يَوْمِ ٱلدِّينِ" />
    <aya index="5" text="إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ" />
    <aya index="6" text="ٱهْدِنَا ٱلصِّرَٰطَ ٱلْمُسْتَقِيمَ" />
    <aya index="7" text="صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ ٱلْمَغْضُوبِ عَلَيْهِمْ وَلَا ٱلضَّآلِّينَ" />
  </sura>
  <sura index="2" name="البقرة">
    <aya index="1" text="الٓمٓ" bismillah="بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ" />
    <aya index="2" text="ذَٰلِكَ ٱلْكِتَٰبُ لَا رَيْبَ فِيهِ هُدًى لِّلْمُتَّقِينَ" />
    <aya index="3" text="ٱلَّذِينَ يُؤْمِنُونَ بِٱلْغَيْبِ وَيُقِيمُونَ ٱلصَّلَوٰةَ وَمِمَّا رَزَقْنَٰهُمْ يُنفِقُونَ" />
```

Figure 4: xml file that used, downloaded from Tanzil project.

```
1  Location location = new Location(113, 3, 3);
2  Verse verse1 = Document.getVerse(location).removeDiacritics();
3  Verse verse2= Document.getVerse(location). toUnicode();
4  Verse verse3 = Document.getVerse(location). toBuckwalter();
5  System.out.println(verse1);
6  System.out.println(verse2);
7  System.out.println(verse3);
```

Figure 5: obtaining a token in different formats using JquranTree.

Format	Output
RemoveDiacritics	غاسق
Unicode	غَاسِقٍ
Buckwalter	gaAsiqK

Table 3: several formats for a word.

```
1,1,bisomi,1,1,1,2,P,prefixed preposition <i class="ab">bi</i>,N,genitive masculine noun
2,1,{ll~ahi,1,1,2,1,PN,genitive proper noun &rarr; <a href="/concept.jsp?id=allah">Allah</a>
3,1,{lr~aHoma`ni,1,1,3,1,ADJ,genitive masculine singular adjective
4,1,{lr~aHiymi,1,1,4,1,ADJ,genitive masculine singular adjective
5,2,{loHamodu,1,2,1,1,N,nominative masculine noun
```

Figure 6: Part-of-Speech information tags.

Figure 6, shows the Part-of-Speech tags for each token in the Quran, in this example, we only take a screen shot for the first verse in the first Sura, the first column represents the token location, the second represents the Verse number, while the third is the Buckwalter transliteration, followed by the location for word, verse, Sura. Then the number of tags followed by the tags names.

No	Tag	Frequency	Description
1	N	25137	Noun
2	PRON	24691	Personal pronoun
3	V	19356	Verb
4	P	13007	Preposition
5	CONJ	9450	Coordinating conjunction
6	PN	3911	Proper noun
7	REL	3575	Relative pronoun
8	REM	2925	Resumption particle
9	NEG	2688	Negative particle
10	ACC	2283	Accusative particle
11	ADJ	1961	Adjective
12	EMPH	1244	Emphatic lam prefix
13	T	1166	Time adverb
14	DEM	1059	Demonstrative pronoun
15	COND	1049	Conditional particle
16	INTG	946	Interogative particle
17	SUB	684	Subordinating conjunction
18	LOC	669	Location adverb
19	RES	558	Restriction particle
20	CERT	414	Particle of certainty
21	VOC	376	Vocative particle
22	RSLT	350	Result particle
23	PRO	332	Prohibition particle
24	PRP	319	Purpose lam prefix
25	CIRC	293	Circumstantial particle
26	SUP	235	Supplemental particle
27	PREV	162	Preventive particle
28	FUT	161	Future particle
29	RET	122	Retraction particle
30	EXP	104	Exceptive particle
31	INC	90	Inceptive particle
32	CAUS	88	Particle of cause
33	IMPV	78	Imperative lam prefix
34	EXL	66	Explanation particle
35	AMD	65	Amendment particle
36	INT	47	Particle of interpretation
37	ANS	40	Answer particle
38	EXH	40	Exhortation particle
39	SUR	35	Surprise particle
40	AVR	33	Aversion particle
41	INL	30	Quranic initials
42	EQ	6	Equalization particle
43	COM	3	Comitative particle
44	IMPN	2	Imperative verbal noun

Table 4: Part-of-Speech tags used in the Quranic Arabic Corpus http://corpus.quran.com/