Research on Humanities and Social Sciences
ISSN (Paper)2224-5766 ISSN (Online)2225-0484 (Online)
Vol.5, No.21, 2015
www.iiste.org

Authorship Verification in Arabic using Function Words: A Controversial Case Study of Imam Ali's Book Peak of Eloquence

Khalid Shakir Hussein
English Dept. Thi-Qar University, An-Nassiriyah, Iraq

Abstract
This paper addresses the viability of two multivariate methods (Principal Components Analysis and Cluster Analysis) in verifying the disputed authorship of a famous Arabic religious book called (Nahjul-Balagha/ Peak of Eloquence). This book occupies an exceptional position in the history of the huge debates held between the two basic Islamic sectors: Sunni'e and Shia. Therefore, it represents a serious challenge to the viability of the multivariate techniques in resolving certain types of historical and sectarian conflicts and controversies. Furthermore, verifying the authorship of this book could be a good opportunity to find out whether there are certain quantitative techniques of attribution that hold for different languages such as English and Arabic. Function words have been targeted in this paper as possible indicators of the author's identity. Accordingly, a set of Arabic function words would be tested using WordSmith Tools (version 5). It turned out that the multivariate techniques are most likely robust for addressing the type of issues raised about Nahjul-Balagha. Besides, it appeared that the statistical patterns of function word usages are quite sensitive to genre in Arabic.
Keywords: authorship attribution, authorship verification, stylometrics, computational stylistics.

1. Introduction

This paper is an attempt to test the capability and efficiency of the multivariate methods in settling down a real case study of verifying the authorship of a heatedly debated Arabic text. The book under investigation stands as a hallmark of the recently increased hassle between the two basic Islamic sectors: Sunni'e and Shia. Its authenticity is the real moot point that stimulates all kinds of sectarian dispute across the Islamic world. Here comes the role of the quantitative attributional verification to conduct a rather objective investigation of the problem.

Authorship verification is a particular case of authorship attribution. Generally speaking, the questions usually posed in authorship analysis might vary according to the circumstances of the cases under investigation. However, two questions are basically addressed in authorship analysis: "*which author, among a clear-cut set of candidate authors, has written the questionable document?*" " . . . *Did a particular author write the document?*" (Luyckx, and Daelemans 2008). The first question is approximated by attributing the disputed texts to one of the *n* candidate authors. The second is a consistency question approximated in cases where the researcher is faced with a set of textual samples attributed to a single author and he has to figure out whether a given disputed text sample belongs to this set or not.

2. Nahjul-Balagha Authorship Controversy

Nahjul-Balagha (Peak of Eloquence) is a religious text well-known for its highly distinctive *genre* and *style*. The distinctive figurative style of the book and the various genres it involves crucially contribute to its uniqueness within the rhetorical map of Arabic. It comprises a collection of *sermons*, *letters*, and *sayings* attributed to Imam *Ali Ibn Abi Talib*, cousin and son-in-law of Prophet *Muhammed*. This collection was compiled in a book in the fourth century (A.H) by a well-respected Shi'i scholar and poet Shareef Razi over 300 years after Imam Ali (Nahjul-Balagha 1998). The text he had selected fall into three different genres: 241 *sermons*, 79 *letters* and 489 *sayings*. These numbers may vary in different editions of Nahjul-Balagha.

Known as a literary masterpiece, this book is deemed exceptionally distinctive in its eloquent standards and style in Shia Islam. It is seen by Shia scholars as being above the words of human beings and below the words of Allah and third only to the Qur'an and Prophetic narrations (see http://www.nahjulbalagha.org/sermons.php).

Since the time it has been collected, Nahjul-Balagha was and still a subject of analysis and interpretation. The genuineness and authenticity of this book were not questioned by Shia scholars. They think that the references Al-Razi had made to the sources from which he collected the sermons, letters and sayings of Imam Ali were quite enough to guarantee the authenticity of the book (ibid.). However, Al-Razi did not dedicate an independent bibliography for the sources, he instead referred occasionally to certain books and their compilers as he incorporated some explanatory notes at the margins of the expounded sermons (see shttp://www.al-islam.org/nahjul/sources.htm) .

The first person that suspected the attribution of Nahjul-Balagha to Imam Ali was *Ibn Khalikan*, a Sunni scholar (see http://www.islamology.com/Resources/Nahj-Imam/main/main1.htm). Then a series of Sunni scholars followed Ibn Khalikan's doubts: Ibn Al-Athir Al-Jazari, Salah Al-Din Al-Safadi, and Al-Dhahabi.

Nevertheless, all the reasons such scholars used to back up their attributional doubts are more or less ideological and biased rather than professional and technical.

The most controversial sermon that received a huge amount of denial from the Sunni scholars is called Al-Shaqshaqiyyah (the roar of a camel) due to its sensitive subject matter. Imam Ali in this sermon unveiled quite clear and direct indications of the Caliphate being snatched from him by the two caliphs Abu-Baker and Umar. Most Sunni scholars abhor the downgrading of the two caliphs explicitly expressed throughout this sermon. Therefore, this sermon, in particular, was and still deeply doubted and repeatedly attacked by them and seen as being forged and unauthentic. Some other doubts are raised nowadays on Websites by some modern Wahhabi scholars claiming that the whole book has been authored collaboratively by a number of Shia orators and outspoken experts of literature (see www.sunniforum.com/what do Sunnis say about Nahjul-Balagha).

The researcher will apply two multivariate techniques (Principal Component Analysis and Cluster Analysis, henceforth PCA and CA) to address only two types of skeptical questions commonly raised about the authorship attribution of Nahjul-Balagha. First, does this collection of sermons, letters, and sayings share a single author? Or is it the product of a multiple-author conspiracy? Second, does the sermon "Al-Shaqshaqiyyah" belong to this collection?

3. Methodology

Six methodological procedures will be followed in analyzing Nahjul-Balagha corpus:

1. Since the samples selected for this study are machine-readable, the scanning or retyping processes could be a very threatening source of all types of errors. The researcher tried his best to check the authenticity of each sample making sure that each one is highly representative of the hard copy. What is more, all the non-authorial materials have been removed from the main body of the texts such as, *titles of sermons*, *author names*, *dates*, *Quranic verses*, *Prophetic narrations*, poetic lines, etc
2. Transcribing every individual sample into *plain text format*
3. Grouping all the samples into one master corpus
4. Analyzing samples with their master corpus via *WordSmith Tools (5.0)* for frequency and word count, besides producing some sort of charts representing basic statistical descriptions
5. Importing WordSmith Tools (5.0) outputs into an excel spreadsheet in a form of matrix
6. Conducting a thorough statistical analysis to the matrix using SPSS (14.0) (Principal Components Analysis/PCA and Cluster Analysis/CA).

3.1. Function Words in Arabic

Function words might be the commonest features that have been counted in the computational studies of authorship attribution. Ever since Mosteller and Wallace (1964) published their influential study of the Federalist Papers, function words have been the focus of interest in numerous papers and studies dealing with these words as possible indicators of authorial styles. Burrows (1987) has conducted another pioneeric study of function words demonstrating the efficacy of such words for attributing different texts and samples to different authors. Then, Grieve (2007) has produced an extensive quantitative evaluation of attributional techniques that underscored the powerful performance of function words in discriminating various authorial styles.

The appeal of function words in attributional studies lies in their being important markers of authorial individuality. Much has been written about the rationale behind the assumption that people tend to express themselves in stable and unique patterns of function words usage. This rationale almost always instigates three salient characteristics about function words: their high frequency, low semantic load, and the very fact that their usage lies beyond our conscious control (Zhao and Zobel 2005: 174-189).

Determining function words hierarchical lists in Nahjul-Balagha, however, is not an easy process. Arabic morphology is quite complex and demanding when it comes to the morphology of function words. There is a diversity of suffixes and prefixes that should be detached from the basic function words before measuring their frequencies and distributions. The inflectional morphology of Arabic created a dilemma for the researcher as to whether to target only the function words occurred distinctly stripped off any further affixational elements, for example (Min/from; Fi/in; Ina/is; . . . etc.) or to target the distinct ones along with their affixational occurrences, for example (Min/from- Minaa/from us- Minhu/from him- Minha/from her- Minhuma/from them (dual)- Minhuna/ form them (femi.)- Minki/from you (femi.), Minka/from you (masc.))

It was a crucial choice to be made, though the first choice sounds too easy to be worked out. On the one hand, selecting the non-affixational function words would save much time in figuring out the wordlist and even in conducting the statistical analyses. On the other hand, if we skip the affixed function words there might be a high risk of disturbing the actual statistics of these linguistic items leading to probably serious problems in attribution.

It would be more reasonable if we try our hands at both choices to see how far the effects would be on the attributional process. First, the researcher will consider Arabic function words that occur distinctly with no

affixes whatsoever. Then the function words will be accounted for, in the second choice, regardless of how many affixes are attached to them. Nevertheless, the traditional basic statistics of the book will be explored first to find out how far the traditional statistical characteristics might be of use in resolving the questions raised in this paper.

4. Results (Basic statistics)

The collected texts in Nahjul-Balagha have been segmented into eleven (5,000) token blocks or samples. The authentication of these samples was painstaking due to the many and various citations incorporated into the textual body of the samples in question. The citations removed by the researcher were a diversity of Quranic verses, poetic lines, major titles, Prophetic narrations and alike. Then the eleven samples were grouped into one master corpus. The latter was subjected to hierarchical processing for frequency and the table (1) below explores the basic statistical descriptions of the master corpus.

The mean of *word-length* is in particular appealing in the table below. The extremely narrow range of the word-length means (from 7.48 to 8.13) indicates a possibly significant characteristic of a single authorship. As for the mean of *sentence-length* in words, it is really statistically turbulent with a wide numerical band (from 22.80 to 34.81). There are (twelve) figures fluctuating between the statistical profiles of the samples. Therefore, there is a rather wide range of statistical transition. Hence, the mean of sentence-length does not help much in drawing any significant conclusion regarding the authorship of Nahjul-Balagha.

Table 1. Basic Statistics of Nahajul-Balagha Corpus

	1	2	3	4	5	6	7	8	9	10	11
File Size	45,145	46,039	46,862	45,882	45,930	45,398	47,406	45,681	44,223	44,640	33,615
tokens	5,016	5,015	5,023	5,017	5,016	5,001	5,021	5,025	5,049	5,025	3,857
types	3,341	3,425	3,410	3,190	3,346	3,112	3,367	3,248	3,038	2,994	2,288
type/token ratio (TTR)	66.61	68.30	67.89	63.58	66.71	62.25	67.06	64.65	60.18	59.58	59.32
standardised TTR	76.46	77.78	76.56	74.26	76.04	72.66	76.22	73.40	71.42	70.24	69.60
standardised TTR std.dev.	18.56	17.83	18.76	21.01	20.06	23.08	19.28	24.13	22.84	23.60	23.40
standardised TTR basis	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
mean word length (in characters)	7.70	7.89	8.05	7.89	7.86	7.79	8.13	7.80	7.51	7.66	7.48
word length std.dev.	2.85	2.88	2.98	2.92	2.91	2.98	3.04	2.96	2.81	2.88	2.80
sentences	217	178	156	220	197	187	176	163	145	173	121
mean (in words)	23.12	28.17	32.20	22.80	25.46	26.73	28.53	30.82	34.81	29.05	31.88
std.dev.	19.11	21.21	25.26	16.18	18.68	20.74	22.00	25.37	26.20	20.88	20.33

The lexical diversity is indicated by a comparatively narrow range (from 59.32 to 68.30). The nine figures continuum of this range suggests an exceptionally rich vocabulary attributed to the claimed author of the samples. One particular finding, however, might be reached if one gives a quizzical look at the values of the Type/Token Ratio (henceforth TTR). There is an observable shift in the TTR values scored in samples (9), (10), (11) showing (60.18; 59.58; 59.32) respectively (see **Figure 1** below). This range of values sounds quite stable and distinct in comparison to that of the rest of samples which ranges from (62.25) to (68.30).

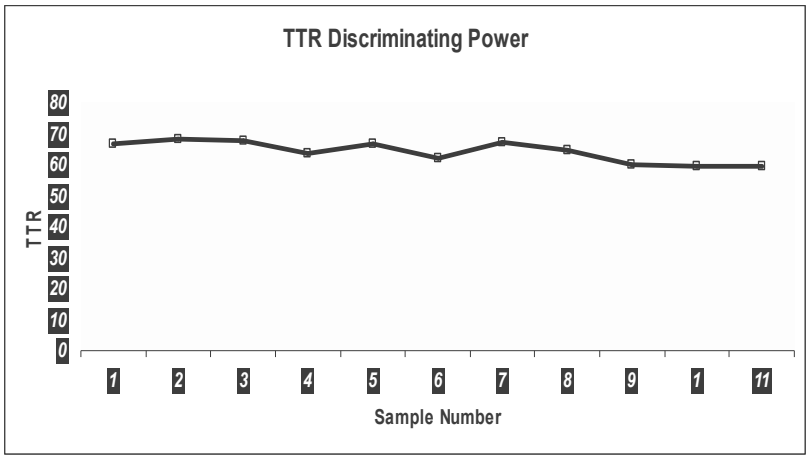

Figure 1. TTR Discriminating Power in Nahajul-Balagha Corpus

It seems that TTR values are perhaps highly sensitive to genre: the samples (9, 10) constitute the *letter-blocks*, and (11) the *saying-block* in Nahjul-Balagha, whereas the remaining ones represent the *sermon-blocks*. Nevertheless, TTR does not provide us with any decisive and finite conclusions about the two questions raised about the authorship of Nahjul-Balagha. How about the Hapax Legomena/Type Ratio (henceforth HTR)?

The eleven samples reveal insightful and noticeable harmony in the HTR values (see Table 2 below). The statistical consistency of the values provide us with an unavoidable avenue to draw rather concrete evidence of *unitary authorship* of Nahjul-Balagha. All the samples share a great deal of similar HTR values ranging from (0.85) to (0.88). **Figure 2** shows some sort of a straight line extending throughout the stable node-samples with roughly indiscriminate plottings.

Table 2. HTR Caculated for the Master Corpus of Nahajul-Balagha

Author	Segment	Types	Hapax Legomena	HTR
Imam Ali	1	3341	2926	**0.87**
	2	3425	3042	**0.88**
	3	3410	3033	**0.88**
	4	3190	2774	**0.86**
	5	3346	2958	**0.88**
	6	3112	2718	**0.87**
	7	3367	2961	**0.87**
	8	3248	2866	**0.88**
	9	3038	2629	**0.86**
	10	2994	2568	**0.85**
	11	2288	1978	**0.86**

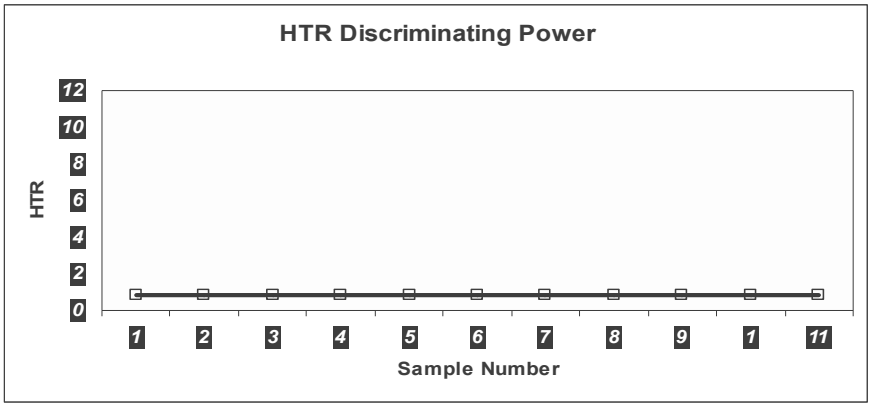

Figure 2. HTR Discriminating Power in Nahajul-Balagha Corpus

The turbulences seen in TTR values disappear in an exceptional way in the HTR plottings. This underscores the significance of HTR as a highly robust statistical feature of great usefulness in attributing samples.

Incorporating *Al-Shaqshaqiyyah* in sample (1), the HTR of this sample (0.87) stands high in favour of a strong attributional affinity that holds between this sample and the other ones. If *Al-Shaqshaqiyyah* was forged and unauthentic, it would not share the same range of unique occurrences with the other sermons. This might well be considered a plausible answer for the second question raised about Nahjul-Balagha in this study.

4.1 Distinct Function Words Frequency

Going through the distinct function words frequency, more than 70 function words were tabulated after processing the master corpus. However, some function words had zero frequency for at least one of the eleven samples. Thus, they were removed from the list. The list was left with only 40 function words with non-zero frequencies all over the samples. Below is a table showing the list of these function words, their distributions and percentages all through the master corpus.

Table 3. The Top-40 Distinct Function Words of Nahjul-Balagha-Corpus

WordSmith Tools --			14/7/2013
N	Word	Freq.	%
1	و	8509	19.0826807
2	لاَ	1505	5.657052803
3	مِنْ	1008	1.608604789
4	فِي	898	1.433062553
5	مَا	760	1.171345115
6	عَلَى	683	1.089957356
7	مَنْ	383	0.587268412
8	إِلَى	368	0.543668315
9	لَمْ	320	0.510668159
10	قَدْ	316	0.504284799
11	أَنْ	304	0.48513478
12	عَنْ	250	0.398959517
13	إِلاَّ	234	0.37342611
14	حَتَّى	190	0.303209245
15	أَوْ	178	0.284059167
16	كَانَ	170	0.279257264
17	إِنَّ	157	0.250546575
18	ذلِكَ	156	0.248950735
19	الَّذِي	136	0.217033982
20	ثُمَّ	132	0.208650623
21	بَعْدَ	117	0.200582613
22	بِمَا	106	0.181920611
23	لَوْ	96	0.169158831
24	هُوَ	94	0.156392127
25	عِنْدَ	93	0.153200448
26	بَيْنَ	93	0.153200448
27	لَيْسَ	91	0.150008783
28	هذَا	90	0.148221263
29	إِذَا	84	0.143625423
30	أَمَّا	83	0.138050399
31	فِيَما	79	0.128050399
32	غيرَ	72	0.122454559
33	كُلَ	71	0.121362219
34	معَ	62	0.114900343
35	أنتَ	53	0.097983576
36	قبلَ	52	0.095387743
37	إذ	51	0.092270252
38	أنا	50	0.088845017
39	التي	42	0.064258496
40	كيف	32	0.058258496

The function words above account for 37.64% of all the words in the master corpus. This percentage still falls within the limits set by Burrows' guidelines (1992), but it is quite expected that the percentage will go a little bit down after removing 30 function words with zero frequencies. Three content words have been removed from the list: *Allah*- 901 tokens, 4th word on the list; *Adduniah/ the world*- 155 tokens, 27th word on the list, and *Alhaq/the truth*- 66 tokens, 70th word on the list.

4.1.1 Analysis Matrix

The researcher designed the analysis matrix that will encompass the frequencies scored for the (40) function words. Below is a sample of the matrix, the first four words in the hierarchy as well as the segment length. The complete matrix can be found in the Appendix. The segment length is used in the matrix because the textual body can not be evenly divided into (5,000) word segments.

Sample Matrix – Nahjul-Balagha Corpus

Author	Text	Segment	Wa	La	Min	Fi	Segment Length in Tokens
A	NB	1	790	131	110	87	5,016
B	NB	2	835	111	115	96	5,015
C	NB	3	843	133	132	92	5,023
D	NB	4	762	145	118	82	5,017

4.1.2 Statistical Analysis through PCA

Principal components analysis was used to deal with the over-dimensionality of the 40 function words that should be reduced into a more manageable number of components (factors). It is usually the first two dimensions or components that sufficiently explain the most interesting variables and determine which variable is correlated more highly with one component rather than with another (see Craig and Burrows 2001). This is the only way by which the researcher can draw conclusions about the behavior of the function words all through the eleven samples. Some function words are expected to stand out constituting salient markers with crucial role in attributing Nahjul-Balagha.

Conducting PCA on the corpus of the eleven samples indicates that the first two components could capture 42.70% of the total variance within the data environment. These two components are quite enough to determine which variables are highly correlated with each one of them. The first component accounted for 23.30% of the total variance and the second factor 19.40%. Table (4) below explains the percentage of variance captured by each component.

Table 4. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	9.320	23.301	23.301	9.320	23.301	23.301
2	7.763	19.406	42.707	7.763	19.406	42.707

Then, the researcher can apply PCA one step further to check the way the function words behave in respect of the two principal components specified above. **Figure 3** below plots the statistical behavior of the function words throughout the eleven samples of Nahjul-Balagha.

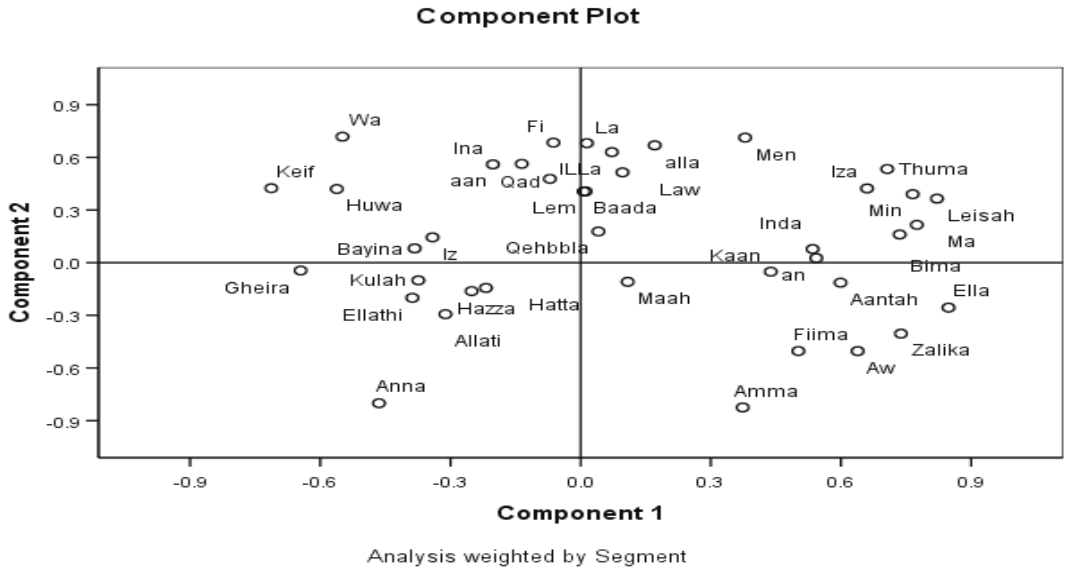

Figure 3. The 40 distinct function words behavior

The first two principal components are represented in the figure above by the x-axis and y-axis. The working principle is rather simple and attractive: those variables or function words that come close together tend to have a similar plotting behavior. Moreover, they tend to be found in one group of samples more often than in another. The function words in **Figure 3** are not treated equally, only those variables that exist at either end of the two axes will be given an authorial weight in controlling the markers of the authorial style.

It is evident that the function words found on the far right of **Figure 3** are Ella/to, Leisah/not, Min/from, Zalika/that, Bima/as. These words together with those located on the far left of the figure (Keif/how, Gheira/but, Huwa/he, and Wa/and) are highly correlated with component (1).

As for component (2), its top is occupied by *Wa/and, Men/who, Fi/in, La/no* and *Alla/on* and the function words *Amma/either, Anna/I am, Aw/or,* and *Fiima/while* move significantly towards the bottom of it. Table (5) below indicates how far the variables have significantly influenced the variance across both components.

Table 5. Components Matrix of PCA

	Component	
	1	2
Wa	-.549	.718
La	.014	.681
Min	.765	.390
Fi	-.063	.684
Ma	.775	.216
Alla	.171	.668
Ella	.848	-.256
Men	.379	.712
Lem	.012	.406
Qad	-.071	.477
An	.543	.026
Aan	-.136	.563
ILLa	.072	.630
Hatta	-.218	-.144
Aw	.639	-.504
Kaan	.438	-.051
Ina	-.202	.560
Zalika	.738	-.404
Ellathi	-.388	-.200
Thuma	.707	.534
Baada	.007	.406
Huwa	-.562	.420
Bima	.735	.160
Law	.097	.514
Inda	.535	.078
Bayina	-.383	.081
Leisah	.821	.365
Hazza	-.251	-.162
Iza	.660	.423
Amma	.374	-.825
Fiima	.502	-.504
Kulah	-.374	-.101
Gheira	-.645	-.045
Maah	.108	-.109
Aantah	.599	-.114
Qehbbla	.041	.178
Anna	-.465	-.801
Iz	-.341	.144
Allati	-.312	-.293
Keif	-.713	.424

Plotting the statistical behavior of the 40 function words, the researcher needs to verify how much of each principal component is referenced by the individual samples. **Figure 4** plots the eleven samples through a scatterplot giving us a representative graph of the way these samples behave based on 42.70% of the data variance.

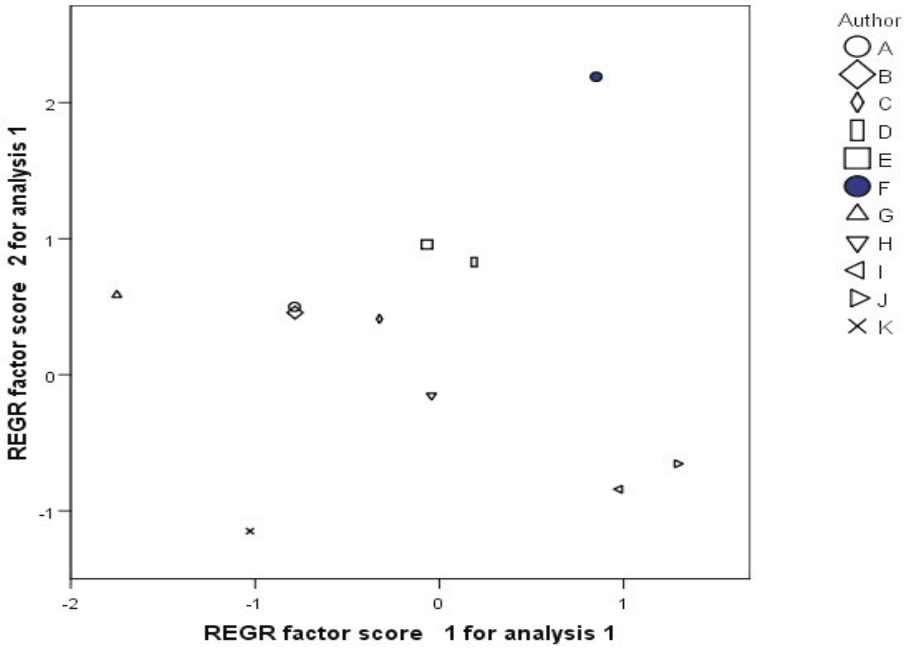

Figure 4. Eleven Samples behavior

The scatterplot in **Figure 4** above does not show clear cut clusters, though there is a major cluster lurking behind this chaos. It is plotted at the center of the figure with six samples orbiting around (1, 2, 3, 4, 5 and 8). While the remaining samples (6, 7, 9, 10, 11) go in diverse and scattered directions along the two axes: (6) jumps to the top right corner, (7) diverts to the left on the further side of the horizontal axis, (9) and (10) tend to the lower right, and (11) is on the lower left.

The anomalous behavior of (6, 7, 9, 10 and 11) hints that the way function words are used in these samples is inconsistent with that in (1, 2, 3, 4, 5 and 8). This could cast some serious doubts about the claimed *unitary authorship* of Nahjul-Balagha. However, the researcher would not hasten to draw any premature conclusions about the two questions raised in this case study. Further statistical analysis is required to confirm the outputs of the PCA.

4.1.3 Statistical Analysis through CA

Hierarchical cluster analysis is used to trace possible groups that might be observed in the numerical tables that we consider for the eleven samples. This analysis is always used regardless of any assumptions a researcher might make about the corpus he is interested in (Baayen 2008). Consequently, the researcher will allow this multivariate technique to trace groups or any possible clusters by calculating the degrees of similarity or difference between all the samples under consideration.

The significance of this analysis lies in its outputs that might confirm or disconfirm the researcher's findings in PCA. The outputs are usually presented in a graphic representation that visualizes the distance at which clusters come close to each other. This graphic representation is called *a dendrogram*. Dendrograms are usually read from left to right. The groups of membership will be allowed to form relying on the statistical characteristics of the eleven samples. No presuppositions are undertaken by the researcher, the individual samples will behave in a way that complies with their own genuine statistical features.

The groups that will be allowed to form by CA are measured in accordance with the statistical characteristics of the individual samples including the disputed sermon of ~~(~~*Al-Shaqshaqiyyah*~~)~~ incorporated in the body of the first sample. It is a privilege that CA does not pay attention to any assumptions might be taken about the data. It is quite independent of PCA in terms of the way the latter addresses the targeted samples. It will figure out any possible similarity calculations between the samples that PCA may have missed. Nevertheless, CA credits or discredits the outputs produced by PCA. It should be noted that the researcher is conducting CA assuming that he knows nothing about the clustering possibilities that the eleven samples might show.

The dendrogram reported in **Figure 5** below represents the findings of CA conducted on the eleven samples. What is immediately apparent is that there is a sort of clustering that identifies (1, 2, 3, 4, 5 and 8) rather closely with each other, though their subgroups are exhausted by some further lines at which they meet: (1) and (2) are identified most strongly on the first vertical line and then cluster with the 3rd sample. Similar clusters continue to merge indicating similarity among the individual samples: (4) and (5) cluster on the second vertical line and then align themselves with (8) on the fourth vertical line and with (7) on the fifth vertical line. The latter in turn meets with (11) on the sixth vertical line. It is notable that (9) and (10) cluster on the third vertical line, as

for (6), it turns out to be an outlier clustering with little similarity to any of the samples involved.

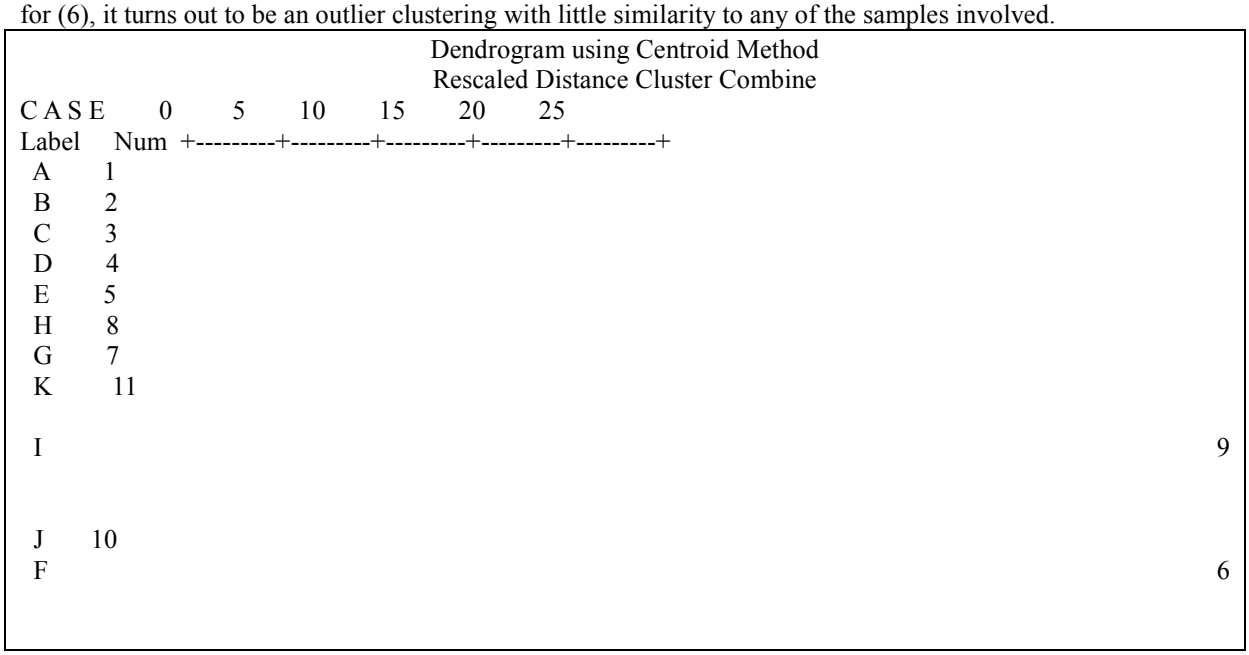

Figure 5. Dendrogram Using Centroid Method.

It looks observable that there is an *outdistancing* effect pushing the samples (6, 7, 11) in particular further into the margins of the figured analysis environment. This effect could be seen even in the ascending numbers of the vertical lines, the samples meet first at the first line and then ascend up to the sixth line. Though the overall statistical behavior of these samples does not show a quite promising uniqueness in terms of the function words usage, the researcher can build an argument based on the highly distinctive and unique cluster of the first and second samples formed on the first vertical line. If sample (2) is confidently attributed to Imam Ali, it is highly possible, then, that (1) belongs to him on the ground of the intensive degree of similarity that holds between the two samples, though this is not enough. The scattered plottings of the other samples, however, suggest an evident disapproval of attributing them to a single author.

4.2 Affixed Function Words Frequency

Now it is the time for the researcher to try the second choice. The researcher thinks that all the statistical disharmony that hangs over the samples checked above is brought into the analysis environment by, first, ignoring those function words attached to affixes (prefixes or suffixes) and, second, by restricting frequency only to those function words stripped off any sort of affixational attachments. It is unreasonable to account for (Fi/ in), for example, only if it occurs alone as a distinct preposition. The type of information that (Fi) conveys is still stable no matter how many affixes are attached to it. Therefore, its function is lying there intact even when a series of suffixes is annexed to it: (Fi*hi*/ in it – Fi*ha*/ in her – Fi*na*/ in us – Fi*him*/ in them (masc.) – Fi*huna*/ in them (femi.) – Fi*huma*/ in them (dual)).

Even the difference in the frequencies of the 40 distinct function words listed above is too considerable to be ignored or mitigated. As one can see from the frequency list of the affixed function words below, (Fi), for example, occurs 898 times as a morphologically distinct preposition, and 1056 as an affixed preposition all through the master corpus of Nahjul-Balagha. The same can be observed for the other function words: (Min/from) rises in frequency from 1008 as a distinct to 1383 as an affixed; (Illa/to) from 368 to 549, and so on and so forth. Unless this difference in frequency is taken into consideration, the attributional statistics will not be trustworthy.

The function words in Table (6) attain comparatively a considerable percentage: they account for 42.35% of the total words in the master corpus. This percentage is soundly higher than that of the distinct function words, 37.64%, a matter which enhances the credibility of the scored affixed words.

Table 6. The Top-40 Affixed Function Words

WordSmith Tools -- 14/7/2013

N	Word	Freq.	%
1	و	8509	19.0826807
2	لاَ	1505	5.657052803
3	مِنْ	1383	5.308604789
4	فِي	1056	2.433062553
5	مَا	760	1.171345115
6	عَلَى	725	1.089957356
7	إِلَى	549	0.587268412
8	إِنَّ	392	0.548968315
9	مَنْ	383	0.543668315
10	عَنْ	346	0.510668159
11	لَمْ	320	0.504284799
12	قَدْ	316	0.48513478
13	أَنْ	304	0.398959517
14	إِلاَّ	234	0.37342611
15	كَانَ	211	0.303209245
16	حَتَّى	190	0.298059167
17	أَوْ	178	0.284059167
18	كُلِّ	168	0.248950735
19	ذٰلِكَ	156	0.217033982
20	الَّذِي	149	0.210650623
21	ثُمَّ	132	0.169158831
22	غَيْرِ	125	0.156392127
23	بَعْدَ	106	0.153200448
24	بَيْنَ	104	0.150008783
25	هٰذَا	101	0.148412943
26	هُوَ	98	0.145221263
27	أَنْتَ	97	0.143625423
28	بِمَا	96	0.141050399
29	لَوْ	94	0.132454559
30	عِنْدَ	93	0.130454559
31	لَيْسَ	90	0.128454559
32	إِذَا	83	0.114900343
33	أَمَّا	83	0.114900343
34	فِيمَا	72	0.083175255
35	مَعَ	63	0.077983576
36	قَبْلَ	58	0.071387743
37	إِذْ	50	0.065270252
38	أنا	49	0.05450172
39	الَّتِي	41	0.054258496
40	كَيْفَ	32	0.051258496

4.2.1 Statistical Analysis Through PCA

The behavior of the affixed function words in the scattorplot below (**Figure 6**) shows slightly different statistical patterns in terms of the first two principal components accounting for 45.83% of the total variance (see Table 7 below).

Table 7. Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	9.881	23.525	23.525	9.881	23.525	23.525
2	9.373	22.317	45.843	9.373	22.317	45.843

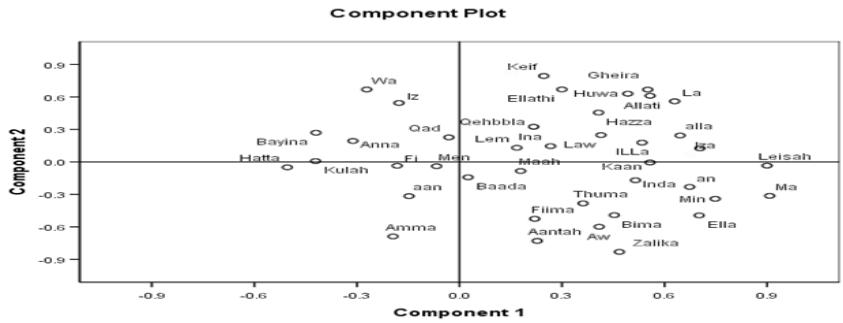

Figure 6. The 40 affixed function words behavior

The function words *Hatta, Kulah,* and *Bayina,* on the far left of the same figure, and *Leisah, Ma,* and *Min*, on the far right, are clearly correlated with component (1). Component (2) is headed at the top by *Keif, Wa,* and *Huwa*, and at the bottom we find *Amma, Zalika,* and *Aantah*. Table (8) below indicates how much the variance in components (1) and (2) is conditioned by the affixed function words.

Table 8. Components Matrix of PCA

	Component	
	1	2
Wa	-.213	.773
La	.592	.683
Min	.836	-.331
Fi	-.030	.202
Ma	.904	-.265
alla	.627	.413
Ella	.651	-.512
Men	.686	-.107
Lem	.409	.321
Qad	.254	.229
an	.633	-.151
aan	.187	-.155
ILLa	.698	.137
Hatta	-.574	-.185
Aw	.343	-.614
Kaan	.324	.163
Ina	.547	.106
Zalika	.426	-.843
Ellathi	-.087	.564
Thuma	.681	-.318
Baada	.272	.163
Huwa	.284	.756
Bima	.541	-.485
Law	.536	.428
Inda	.575	-.195
Bayina	-.479	.296
Leisah	.922	.020
Hazza	.115	.321
Iza	.819	.160
Amma	-.190	-.754
Fiima	.112	-.654
Kulah	-.487	.028
Gheira	.259	.652
Maah	.192	-.279
Aantah	.284	-.742
Qehbbla	.413	.423
Anna	-.279	-.270
Iz	.132	.613
Allati	.117	.631
Keif	-.279	.873

Compared with the way the distinct function words behave, the vast majority of the affixed function words shown in **Figure 6** tend to display a quite notable rush to the right side of the figure. This sharp shift in behavior gives rise to dramatic repercussions over the way the individual samples behave as it could be observed in **Figure 7**" below.

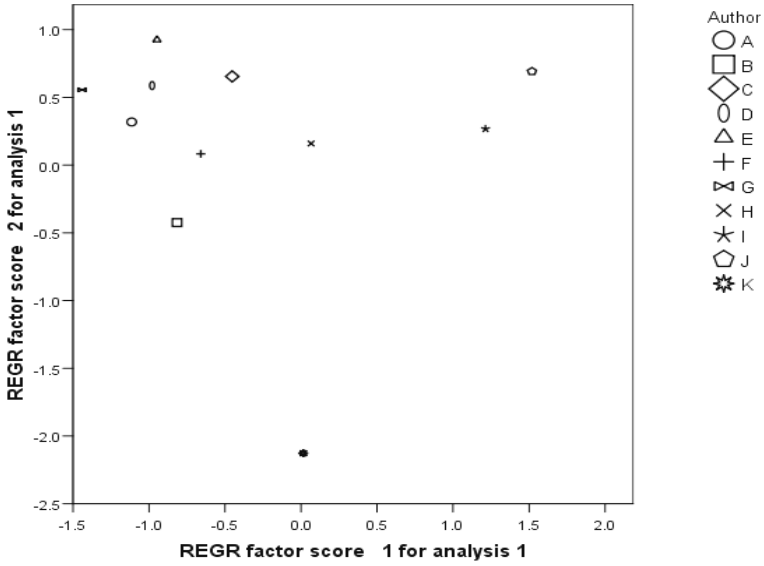

Figure 7. Eleven Samples Behavior.

It is apparent by the vertical axis in **Figure 7** above that samples (1, 2, 3, 4, 5, 6, 7, 8) share noticeable similarities in the patterns of using function words. This most likely consolidates a sort of stylistic harmony that made them grouped in one cluster plotted in the upper left. Also interesting to note is the way samples (9) and (10) divert from all the other samples identifying themselves in one cluster up to the right corner of the plotting. Whereas sample (11) goes down and resides as a single plot recognized on the horizontal axis.

The differences in samples behavior are dramatically associated with considerable differences in *genre*. The latter plays an undeniable role in the way the function words are used. The 241 *sermons* were segmented into eight subsequent samples represented by the first eight segments in Nahjul-Balagha corpus. Samples (9) and (10) represent the 79 *letters*. As for (11), it comprises all what is left of the 489 *sayings*. Therefore, the stylistic differences reflected by the different plottings of the eleven samples might well be attributed to the different writing styles of the three different genres: *sermons, letters,* and *sayings*. Does that mean the subconscious use of function words in Nahjul-Balagha corpus is significantly affected by the text genre?

4.2.2 Statistical Analysis Through CA

At the present analysis, there is no need to run a big risk in drawing concrete conclusions as to the extent to which genre is influential in Nahjul-Balagha corpus. A cluster analysis is needed to verify the findings of PCA and to investigate if the function words usage varies distinctively according to the genre used in the corpus. The dendrogram in **Figure 8** represents the findings of conducting CA over the eleven samples.

The first thing to be recognized in this dendrogram is the existence of two major clusters: the first encompasses the samples (1, 2, 3, 4, 5, 6, 7, 8), and the second holds for two closely identified samples (9) and (10) with a third sample (11) laxing as an outlier. The disputed sermon included in the first sample is still strongly identified with the other samples, especially with the second one. What holds the researcher's interest is the way all the samples are identified with each other: every two samples are intensively aligned on the first vertical line. It is notable that samples (9, 10) cluster independently with a strong correlation that binds the two samples on the first vertical line.

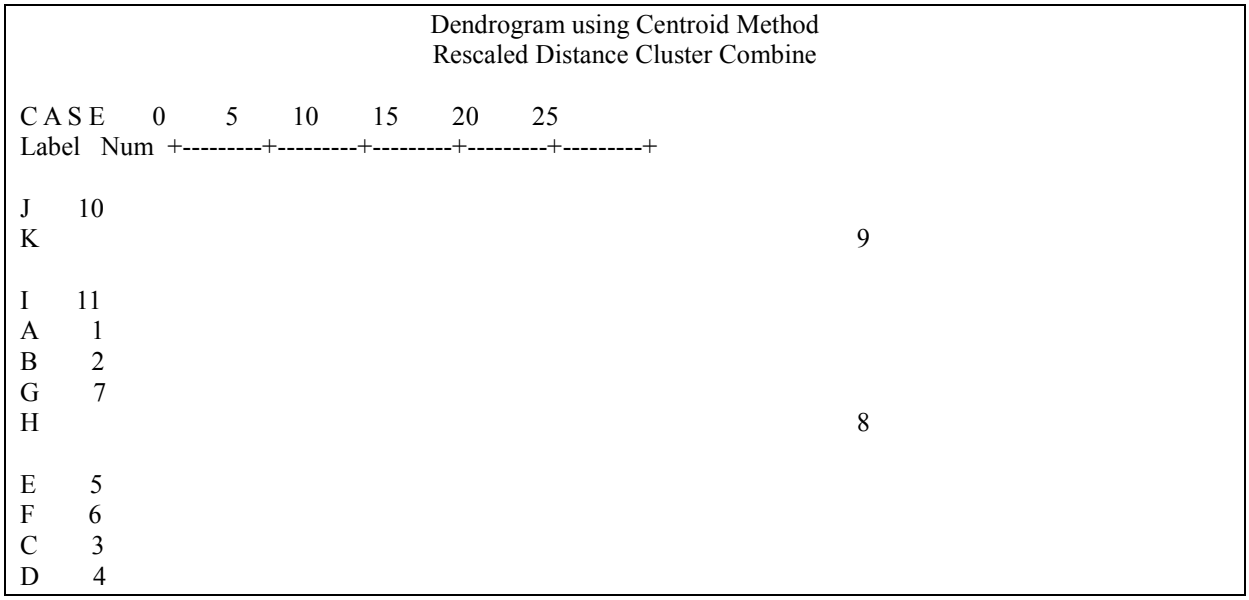

Figure 8. Dendrogram using Centroid Method.

5. Conclusions

It turns out that the multivariate methods (PCA and CA) are perhaps robust for addressing the issues raised about Nahjul-Balagha. It is most likely that the eleven samples do share a single author. The stylistic continuities discerned throughout the stable statistical plottings and clusters of the samples would seem to suggest *unitary authorship*, even though the samples cross genre-lines.

Unlike the rest of the sermons attributed successfully only by targeting the *affixed* function words, "*Al-Shaqshaqiyyah*", interestingly, identifies itself most closely with the rest of the sermons regardless of the type of function words used whether *distinct* or *affixed*. This case study does confirm that the author of Nahjul-Balagha preserved approximately the same unique patterns of function words usage, though these patterns show a discriminatory change when crossing over genre lines (from the territory of *sermons* to that of *letters* and *sayings*). Therefore, it appears that the multivariate methods based on function words usage are highly sensitive in Arabic to genre.

References

Baayen, R. 2008. *Analyzing Linguistic Data*. Cambridge: Cambridge University Press.

Burrows, J. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an experiment in Method*. Oxford: Clarendon Press.

Burrows, J. F. 1992. "Not unless you ask nicely: the interpretive nexus between analysis and information". *Literary and Linguistic Computing*, 7(2), (91-109).

Craig, H.,& Burrows, J. 2001. "Lucy Hutchinson and the authorship of two seventeenth century poems: a computational approach". *The Seventeenth Century*, 16, (259-282).

Grieve, J., 2007. "Quantitative authorship attribution: an evaluation of techniques". *Literary and Linguistic Computing*, 22(3), (251-270).

Mosteller, F., and Wallace, D. 1984. "Inference and disputed Authorship: The Federalist". (Reading, MA: Addison-Wesley Pub. Co., 1964); 2nd edition as *Applied Bayesian and Classical Inference: The Case of the 'Federalist' Papers*. New York: Springer-Verlag.

Ibn Abi Talib, A. 1998. *Nahjul-Balagha*, 3rd ed., Qum: Al-Thaqalein Library.

Luyckx, K. and Daelemans, W. 2008. "Authorship attribution and verification with many authors and limited data". *Proceedings of the 22 nd International Conference on Computational Linguistics*. Manchester: Coling, August, (513- 519).

Zhao, Y. and Zobel, J. 2005. "Effective and scalable authorship attribution using function words". In G.G. Lee et al. (Eds.): AIRS 2005, LNCS 3689 (174-189).

Appendix

Nahjul-Balagah Matrix

Author	Text	Segment	Wa	La	Min	Fi	Ma	alla	Ella	Men	Lem	Qad	an	aan	ILLa
A	NB	1	790	131	110	87	51	53	32	56	19	38	36	30	22
B	NB	2	835	111	115	96	64	61	31	35	43	26	32	28	20
C	NB	3	843	133	132	92	65	68	44	41	45	33	27	31	25
D	NB	4	762	145	118	82	68	76	38	34	24	41	41	27	41
E	NB	5	776	136	128	95	64	72	33	34	37	27	27	43	24
F	NB	6	848	199	133	90	80	75	44	25	33	28	53	18	24
G	NB	7	931	142	111	134	64	57	31	30	19	28	24	42	25
H	NB	8	750	121	136	103	67	76	41	28	29	28	49	40	18
I	NB	9	691	146	132	96	88	57	47	36	35	38	67	33	25
J	NB	10	719	151	149	110	82	80	48	43	23	23	38	31	30
K	NB	11	564	90	119	71	67	50	45	21	19	21	36	23	17

Hatt a	A w	Kaa n	In a	Zalik a	Ellat hi	Thum a	Baad a	Huw a	Bim a	La w	Ind a	Bayin a	Leisa h	Hazz a	Iz a	Amm a	Fiim a
34	18	11	48	4	10	14	12	6	6	9	6	17	7	9	8	14	1
21	18	14	14	10	17	6	13	9	10	12	6	6	6	6	7	14	10
26	11	7	27	15	10	10	19	10	2	7	14	15	7	3	12	5	10
16	15	10	49	10	16	12	7	15	8	7	11	8	11	15	7	8	6
16	15	13	51	11	13	14	14	7	12	5	11	5	5	13	13	3	3
37	10	19	28	26	32	11	24	12	9	20	0	15	6	19	8	15	9
9	5	18	44	9	12	11	15	13	9	9	3	12	5	5	5	9	6
17	29	23	30	20	5	10	17	12	11	13	7	4	8	6	6	11	4
18	30	17	43	28	10	16	15	9	13	8	6	11	11	5	12	13	11
11	20	21	32	23	9	21	15	5	16	13	18	4	14	9	14	14	9
18	29	17	26	25	15	9	10	5	10	2	10	7	6	11	7	21	11

Kulah	Gheira	Maah	Aantah	Qehbbla	Anna	Iz	Allati	Keif	Segment Length in Tokens
8	12	6	1	7	6	8	1	2	5,016
26	9	6	5	13	6	5	5	2	5,015
29	8	1	4	5	1	4	2	1	5,023
6	16	7	1	2	4	3	8	6	5,017
16	17	5	1	3	3	7	3	7	5,016
10	13	13	1	4	6	13	5	4	5,001
20	11	4	1	6	6	1	4	4	5,021
16	11	1	12	4	4	3	2	3	5,025
8	8	3	17	7	3	2	2	1	5,049
13	9	10	4	4	2	2	4	1	5,025
16	12	7	6	3	8	2	5	1	3,857