Spring 2013

Romans 1:18-2:29: A Stylometric Reconsideration

Keith L. Yoder, *University of Massachusetts*

Romans 1:19-2:29: A Stylometric Reconsideration

Keith L. Yoder
University of Massachusetts, Amherst
2013 Eastern Great Lakes Biblical Society

Several have questioned the Pauline authorship of Romans 1:18-2:29. E. P. Sanders, for one, surmised that Paul borrowed existing homiletical material from Hellenistic Judaism and used it here virtually unchanged[1]. William O. Walker argued on linguistic and ideational grounds that the text was composed and interpolated into the existing letter by someone other than Paul[2]. Both agree that the text did not originate with Paul as author. Few, however, have followed this lead.

Authorship attribution has grown into field of strong current interest, especially in the application of computerized tools of modern data analysis to supplement traditional protocols of literary analysis. In this paper I use automated data analysis tools to test the Pauline authorship of Romans 1:18-2:29 following the common protocol of several model studies by leading exponents of automated authorship attribution such as Holmes[3], Tweedy[4], and Forsyth[5]. In New Testament studies, Mealand previously analyzed the Pauline corpus with the same automated tools[6], although his more recent studies reflect a slightly modified approach[7].

Here is a brief summary of the sequential protocol used in these studies:

1. The text(s) to be analyzed must be in machine-readable format, and stemmed or lemmatized as appropriate.
2. A common feature set is chosen from the most frequent words or lemmas from all the texts being studied, for English or Latin texts the first 50 or more, for Greek the first 30 or more, usually exclusive of "content" words. The use of other non-lexical feature sets is being explored in recent studies, but I will not use any of them here.[8]

---

[1] Ed P. Sanders, *Paul, the Law,and the Jewish People* (Philadelphia: Fortress, 1985), 123-135.

[2] William O. Walker, *Interpolations in the Pauline Letters*, (New York: Sheffield, 2001), 166-189.

[3] David I. Holmes, Lesley J. Gordon, and Christine Wilson (2001) "A Widow and her Soldier: Stylometry and the American Civil War", Literary and Linguistic Computing, 16(4):403–420; and David I. Holmes, and Daniel W. Crofts (2010) "The Diary of a Public Man: A Case Study in Traditional and Non-traditional Authorship Attribution", Literary and Linguistic Computing, 25(2):179–197.

[4] Fiona J Tweedie, David I. Holmes, and Thomas M. Corns (1998) "The Provenance of De Doctrina Christiana, attributed to John Milton: A Statistical Investigation", Literary and Linguistic Computing, 13(2):77–87.

[5] Richard S. Forsyth, David I. Holmes, and Emily K. Tse (1999), "Cicero, Sigonio, and Burrows: Investigating the Authenticity of the *Consolatio*", Literary and Linguistic Computing, 14(3): 375-400.

[6] Mealand, David L. (1995) "The Extent of the Pauline Corpus: A Multivariate Approach", Journal for the Study of the New Testament, 59(4):61-92.

[7] David L. Mealand (2011), "Is there Stylometric Evidence for Q?", New Testament Studies, 57(4):483-507; and Mealand, David L. (2012) "Hellenistic Greek and the New Testament: A Stylometric Perspective", Journal for the Study of the New Testament, 34(4):323–345.

[8] See Efstathios Stamatatos (2009) "A Survey of Modern Authorship Attribution Methods," Journal of the American Society for Information Science and Technology, 60(3):538–556

3. The texts are partitioned into appropriate chunks, usually of 1000 or more consecutive words, after excluding quotations from other works, if that is appropriate.
4. For each text partition, counts are made of each word or lemma in the feature set. The set of all the partitions and their respective counts become the samples for the study.
5. Often a set of control texts from the same era and language, and of similar genre, as the study texts is preprocessed and analyzed with the same tools to validate the methodology.
6. Multivariate data analysis tools are applied to the control and study samples in turn, usually in this order: (1) Principal Components Analysis or Correspondence Analysis, (2) Cluster Analysis, and (3) some form of Linear Discriminant Analysis.
7. Graphs and charts produced from these tools are examined for patterns inherent in the data, patterns unobservable by any other means, which give meaningful results in terms of authorship attribution or other text classification questions.

Here are the same steps as performed in this study of the authorship of Romans 1:18-2:29:

1. All the New Testament epistles from Romans through Jude were extracted from a CD-ROM from CCAT[9] containing the third edition of the United Bible Societies Greek text with punctuation and Old Testament quotes annotated.
2. A feature set of the most frequent 35 lemmas was selected from the 44,951 words in the epistles, exclusive of content words[10]; see Table 3 at the end of the paper for the list.
3. Since the text of Romans 1:18-2:29, contains 709 words after excluding an Old Testament quote, all the other texts in this study were partitioned into chunks of 709 consecutive words, exclusive of Old Testament quotes, and exclusive of non-Pauline interpolations identified in Walker's published works[11]. See Tables 1 and 2 at the end of the paper for complete partition definitions and word counts, and Walker interpolations.
4. For each of the resulting 52 text partitions, frequency counts were made of each of the 35 lemmas in the feature set, producing a total of 52 data samples to be analyzed. Table 4 at the end of this paper contains complete feature counts for all samples, comprising 50% of the original text.[12]
5. A control dataset was prepared of three epistle set of the Apostolic Fathers: Barnabas, 1 Clement, and the authentic letters of Ignatius. The text is based on an electronic version[13] of the Lake edition[14], preprocessed, lemmatized, and partitioned as above.

---

[9] The Center for Computer Analysis of Texts, *GNT_UBSC3 on CD-ROM*, Revision 2005; a prior online version, no longer available, also contained lemmatization and morphology.

[10] Mealand often confines his feature set to prepositions and conjunctions, not always the most frequent lemmas. The other studies cited above normally follow the "most frequent" rule, excluding any obvious content words, which is the route I took here.

[11] Walker, *Interpolations*, 90-236; also Walker (2003), "Galatians 2:7b-8 as a Non-Pauline Interpolation," CBQ 65(2003):568-587; and Walker (2007), "1 Corinthians 15:29-34 as a Non-Pauline Interpolation," CBQ 69(2007):84-103.

[12] This is a larger proportion than those in the studies cited in notes 3-5, to compensate for the necessity of using a partition size smaller than the normal 1000 word threshhold.

[13] Kirsopp Lake, The Apostolic Fathers. Print version in following note. Cited 17 December 2012. Online: http://www.ccel.org/ccel/lake/fathers2.pdf. Scanning errors were corrected by reference to the same texts in Thesaurus Linguae Graecae, Online: http://www.tlg.uci.edu/.

[14] 1 Clement, Ignatius, and Barnabas, *The Apostolic Fathers*. 1912. Edited and translated by Kirsopp Lake. 2 vols. LCL. New York: G. P. Putnam's Sons.

6. I apply the data analysis tools of Correspondence Analysis[15] and Cluster Analysis to the control texts, and then to the study texts. I apply Linear Discriminant Analysis only to the study texts, since the authorship of the control texts is not under question here.
7. The heart of this study will be the visual graphs and charts produced by and from the results of the three data analysis tools. They will reveal patterns inherent in the data that will show Romans 1:18-2:29 to be visually and statistically separate from the rest of Romans and the other undisputed Paulines, but visually and statistically close to Ephesians and Colossians, and to Hebrews.

The open source R[16] statistical software was used for all the data analysis and graphics presented below. All data preparation, counting, and assembling was done with Microsoft Excel. The data analysis tools Correspondence Analysis (CA) and Cluster Analysis (Cluster) are exploratory in nature, requiring no categorical or authorship assumptions about the samples being tested. Each of these finds and presents patterns inherent in the data samples themselves. The third tool, Linear Discriminant Analysis (LDA) requires prior group identification for the samples, which is why LDA is used after the other two. I will explain the workings of each of these tools in more detail as they are presented below.

Before applying the automated tools, it is informative to understand why they are needed. If we examine data from one, two, or even three variables, it would easy to rank our 52 samples in relative order for each of the variables, or graph the variables against each other in two-dimensional plots, and perhaps gain valid observations. But that approach fails if we need to take account of 10, 20, or in this case, 35 different variables; it would take 595 graphs to plot all the possible two-dimensional combinations of 35 different variables.

Even so, it turns out here that simply plotting the two most frequent lemmas, ὁ and καί, against each other in a two-dimensional graph, does give helpful information. Figure 1 below shows this plot, ὁ on the horizontal axis and καί on the vertical. Segments of the undisputed Paulines are enclosed in squares, while Romans 1:18-2:29 ("R-0") is encircled. Definite patterns are easily observed. With few exceptions, the Romans segments occupy the lower right quadrant, 1 and 2 Corinthians and Galatians the lower left quadrant, 1 Thessalonians and Philippians the upper left quadrant beside, but separate from, the Pastorals, while our R-0 segment lies in the upper right quadrant amidst an uninterrupted diagonal swath of Ephesians, Colossians, and 1 John. This pattern has a surprising likeness to the results of the multivariate tests applied below.

[15] The studies cited in notes 3-6 all use Principal Components Analysis (PCA); in Mealand's more recent studies, cited in note 7, he switches to Correspondence Analysis (CA) instead of PCA. I follow Mealand in using CA, since preliminary PCA testing on my dataset produced negative cross-validation results. The end results of the two were very similar, but CA seemed mathematically safer.

[16] R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

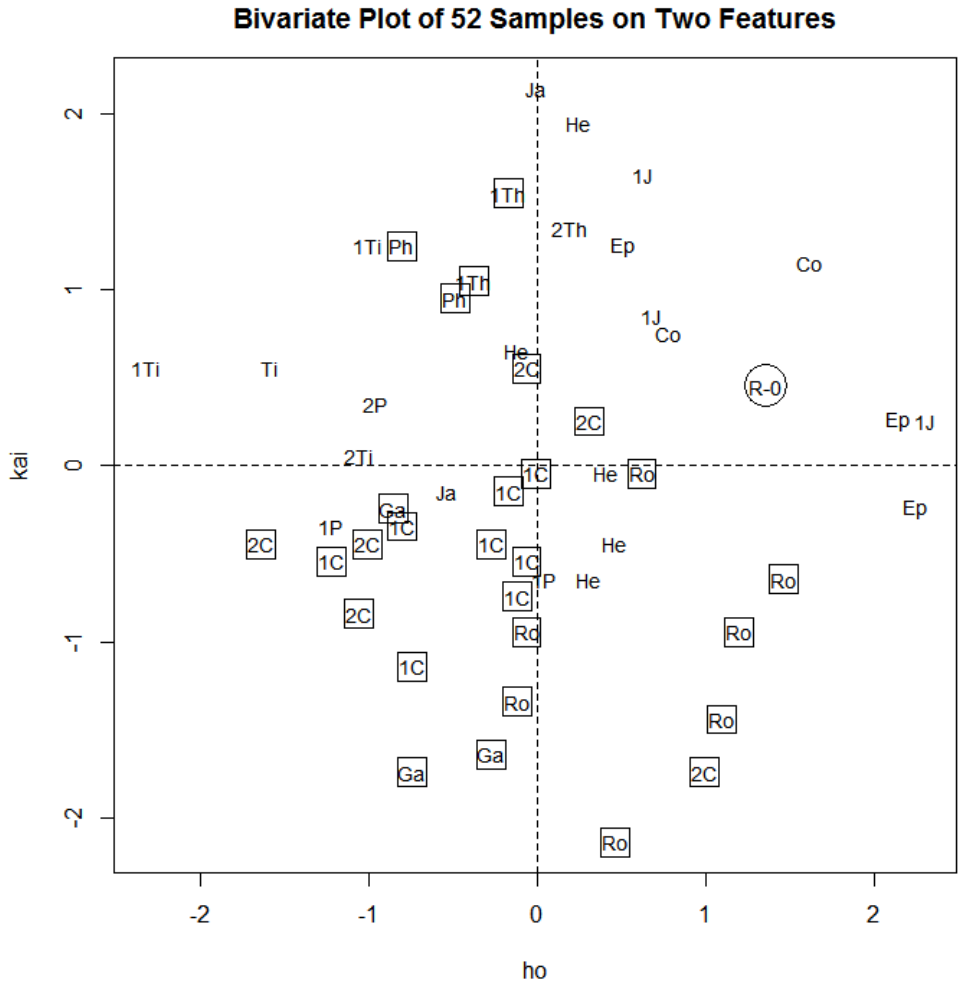

Figure 1: Bivariate Plot of 52 NT Epistle 709-Word Partitions on the Two Most Frequent Lemmas
Data zero-mean centered and scaled to unit variance
Romans 1:18-2:29 encircled, Undisputed Paulines enclosed in squares
Plotting by R Base package, function "plot"

Before testing the New Testament data, the analysis tools will be applied to the control data set of Apostolic Fathers texts: 1 Clement, Barnabas, and the seven authentic letters of Ignatius. The control set was prepared the same way as the New Testament data, using the same 35 lemma feature set. Figure 2 plots the first two principal coordinates produced by the R Correspondence Analysis (CA) tool "ca"[17]. CA partially solves the problem of too much dimensionality noted in the previous section, that is, which of the 35 variables or which of the 595 possible pairs of variables yields the most information about the whole dataset. CA extracts variance information from the entire feature set (here, the 35 lemmas) from which it produces new axes or coordinates, usually one less than the original feature count. The new axes are mutually orthogonal and are automatically presented by CA in order of decreasing variance. Thus a bivariate graph of the first two CA axes will yield more information about the dataset than a graph of any other two CA axes, or any two of the original variables. In this case, the CA graph represents 46% (29.5+15.4,

[17] O. Nenadic and M. Greenacre (2007) "Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package," Journal of Statistical Software, 20(3):1–13.

as noted in the axis labels) of the system variance, much more than the 5-6% that might be expected from randomly combining any two of the 35 original variables.

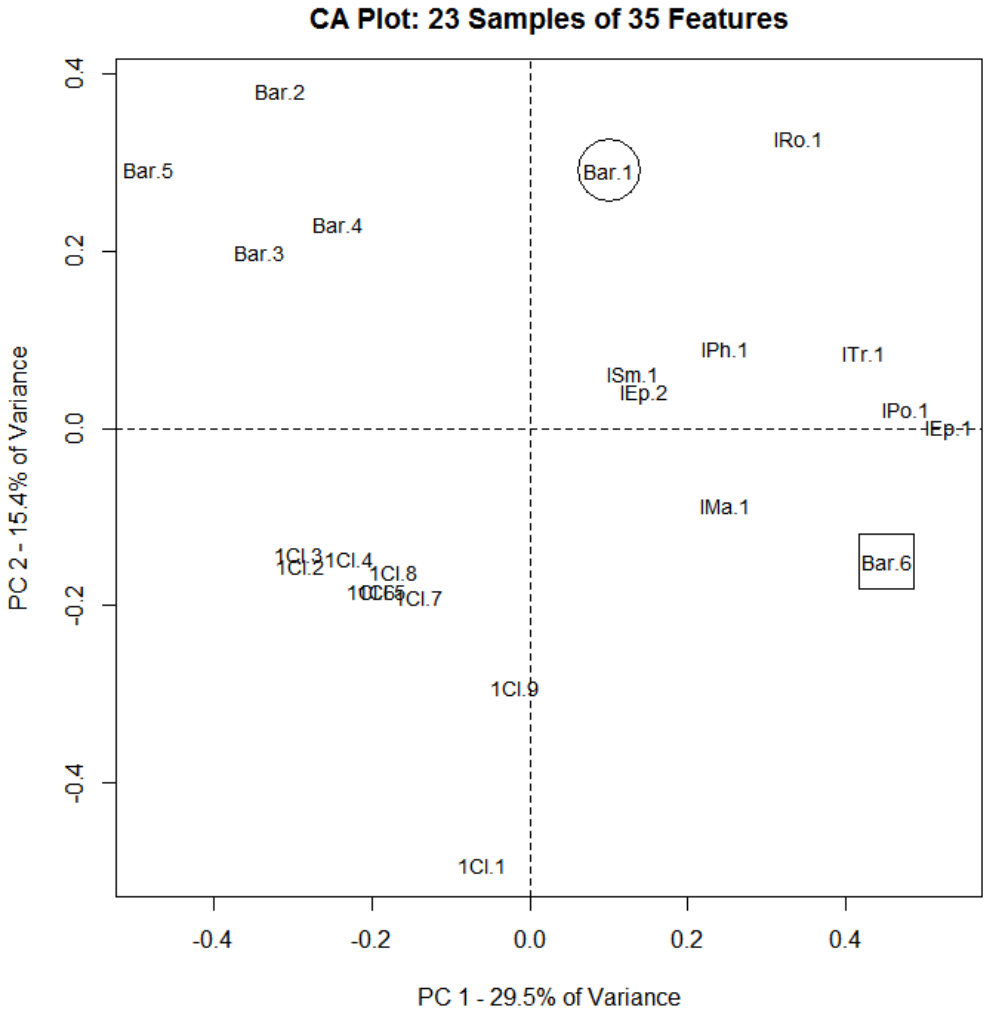

Figure 2 Correspondence Analysis Plot of First Two Principal Coordinates
23 Partitions of 35 Features from Selected Apostolic Fathers
"Bar" = Barnabas, "1Cl" = 1 Clement, "I--" = Ignatius
Data Analysis by R package {ca}, function "ca"

In CA graphs, the relative nearness between sample points gives an accurate visual representation of the relative similarity between the sample data, usually with the points oriented around the data centroid. The small sample count here allows use of full partition labels on this and the next graph in Figure 3 to better view the relationship between the two analysis methods. A pattern of separation between the three groups, and nearness within the three groups, is visually obvious: all 1 Clement segments are bunched in the lower left quadrant, all Ignatius segments are in the right two quadrants in the arc between 1:00 and 4:00 from the center, and all but one of the Barnabas segments are in the top two quadrants in the arc between 10:00 and 1:00. The circled Barnabas segment "Bar.1" is slightly out of place closer to the Ignatius group. The squared Barnabas segment "Bar.6" is quite out of place; that text begins at Barnabas 18:1 and contains the "Two Ways" exposition.

Moving on to Cluster Analysis, that tool uses a distance measurement, here the "Euclidian" method to join together two "leaves" (samples) out of all 23 that are closest together across the entire 35 dimensional feature set. Those two are averaged together into one so there are now 22 data points. Cluster Analysis then finds and joins the next closest two of the new set of 22, and continues this process until all the leaves and branches are clustered together into a single tree. The tree graph or "dendrogram" in Figure 3 was produced by applying the Hierarchical Cluster Analysis tool "hclust" in R package "stats"[18] to the same Apostolic Fathers dataset. Hclust represents this process graphically showing the leaves at the bottom as they are joined together at different heights working upwards until the final union at the top of the (upside down) tree. Note that this hclust graph takes account of 100% of the data variance.

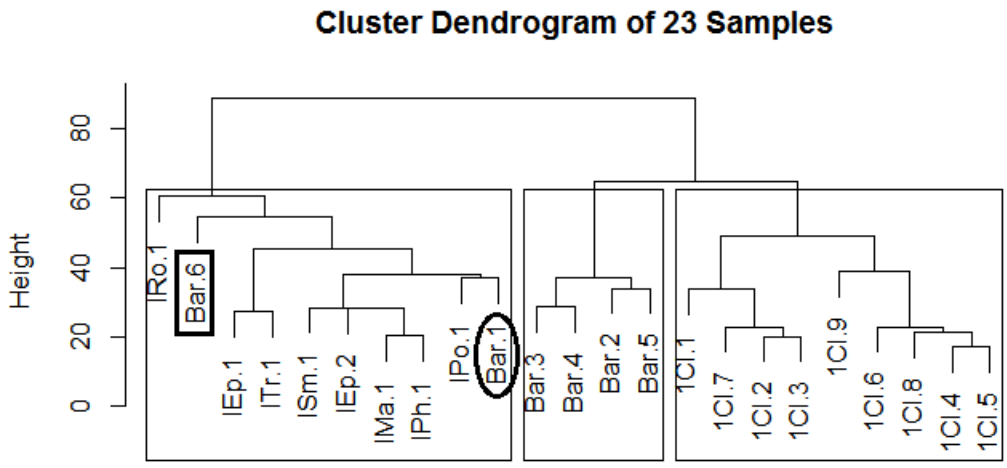

Figure 3: Hierarchical Cluster Dendrogram of 23 Partitions of Selected Apostolic Fathers
Euclidian distance and "complete" clustering methods of 35 Features
Data Analysis by R package {stats}, function "hclust"

The dendrogram is cut to show rectangular highlighting of the top three branches. The separation pattern shown by Correspondence Analysis Figure 2 is confirmed by its replication here: all 1 Clement segments are clustered together in the branch on the right side of the tree, all of Ignatius is clustered in the left-most branch, and four of the six Barnabas segments cluster in the middle branch. The circled Barnabas segment "Bar.1" in Figure 2 between Ignatius and the main Barnabas group is graphed here by hclust with Ignatius, enclosed in an ellipse, adjacent to the central Barnabas cluster. The squared "Two Ways" Barnabas segment "Bar.6" in Figure 2 on the far side of the Ignatius group is graphed here by hclust with the Ignatius leaves, enclosed in a box, and farthest away from central Barnabas cluster. Using a different mathematical approach, Cluster Analysis of the Apostolic Fathers control texts yields results that confirm the predictions of Correspondence Analysis. We may now apply these tools to the New Testament dataset of similar era and epistolary genre with a high degree of confidence in our methodology.

[18] R Core Team (2012), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

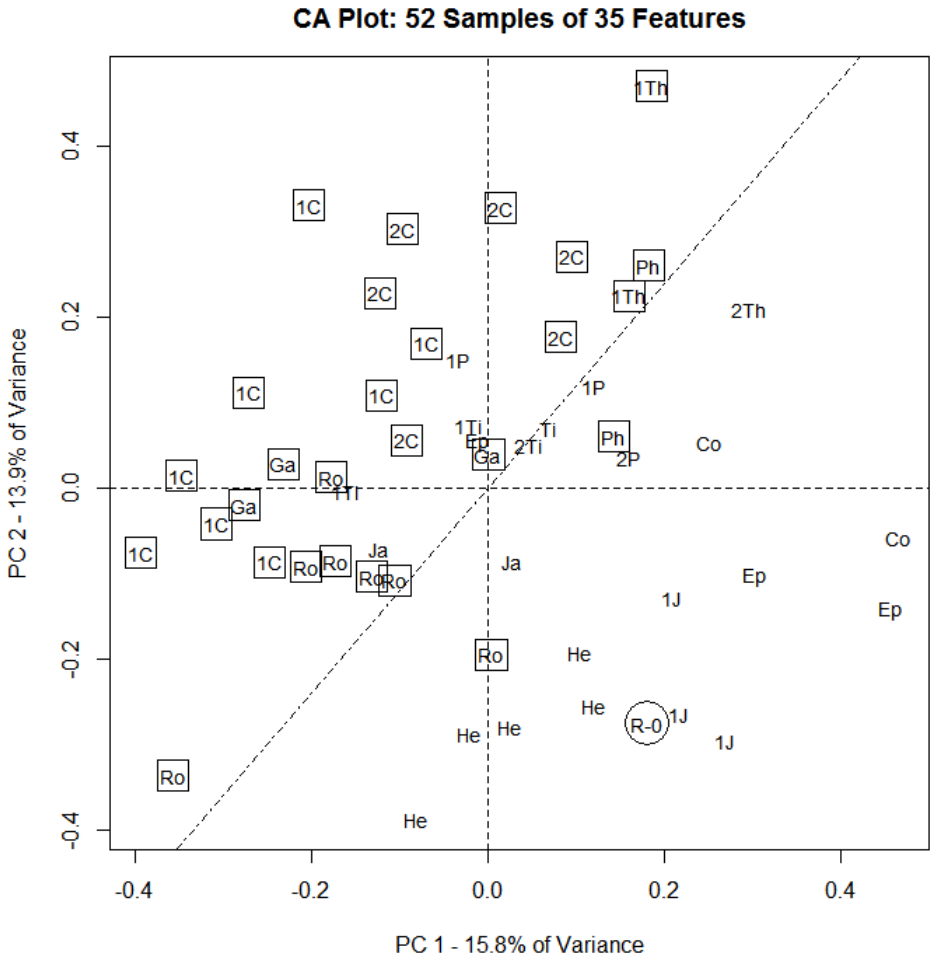

Figure 4: Correspondence Analysis Plot of First Two Principal Coordinates
52 Partitions of NT Epistles of 35 Most Frequent Lemmas
Romans 1:18-2:29 enclosed in circle, Undisputed Paulines enclosed in squares
Data Analysis by R package {ca}, function "ca"

Applying Correspondence Analysis to the New Testament dataset, Figure 4 above shows the graph of the first two principal coordinates produced by ca, comprising 29.7% of the system variance. The undisputed Pauline segments are enclosed in squares, while segment R-0 segment is circled. Due to higher sample count, the graphs in Figure 4 and Figure 5 use abbreviated labels that identify only the book name.

Note that R-0 lies in the lower right quadrant separate from the rest of Romans and nested among segments of Hebrews and 1John. The diagonal line through the data centroid from the lower right to the upper right quadrant shows that Rom.2 and Php.2 are the only two undisputed Pauline partitions that intrude into the lower right half (180 degree arc around the centroid) of the graph, which is otherwise occupied only by R-0, all of Hebrews, 1John, Colossians, and 2 Thessalonians, most of Ephesians, and half of Peter, James, and the Pastorals. Above the diagonal line, moving counter clockwise from the upper right corner, we see the undisputed Paulines blend from Philippians and 1 Thessalonians into 2 Corinthians, into 1 Corinthians and

Galatians, and then Romans. Thus, ca firmly suggests that Romans 1:18-2:29 is stylometrically separate from the undisputed Paulines and closer to Hebrews, 1John, and Ephesians-Colossians.

Using hclust on the same 52 sample epistolary dataset produces the cluster dendrogram shown below in Figure 5. Higher sample volume requires this tree to be graphed horizontally.

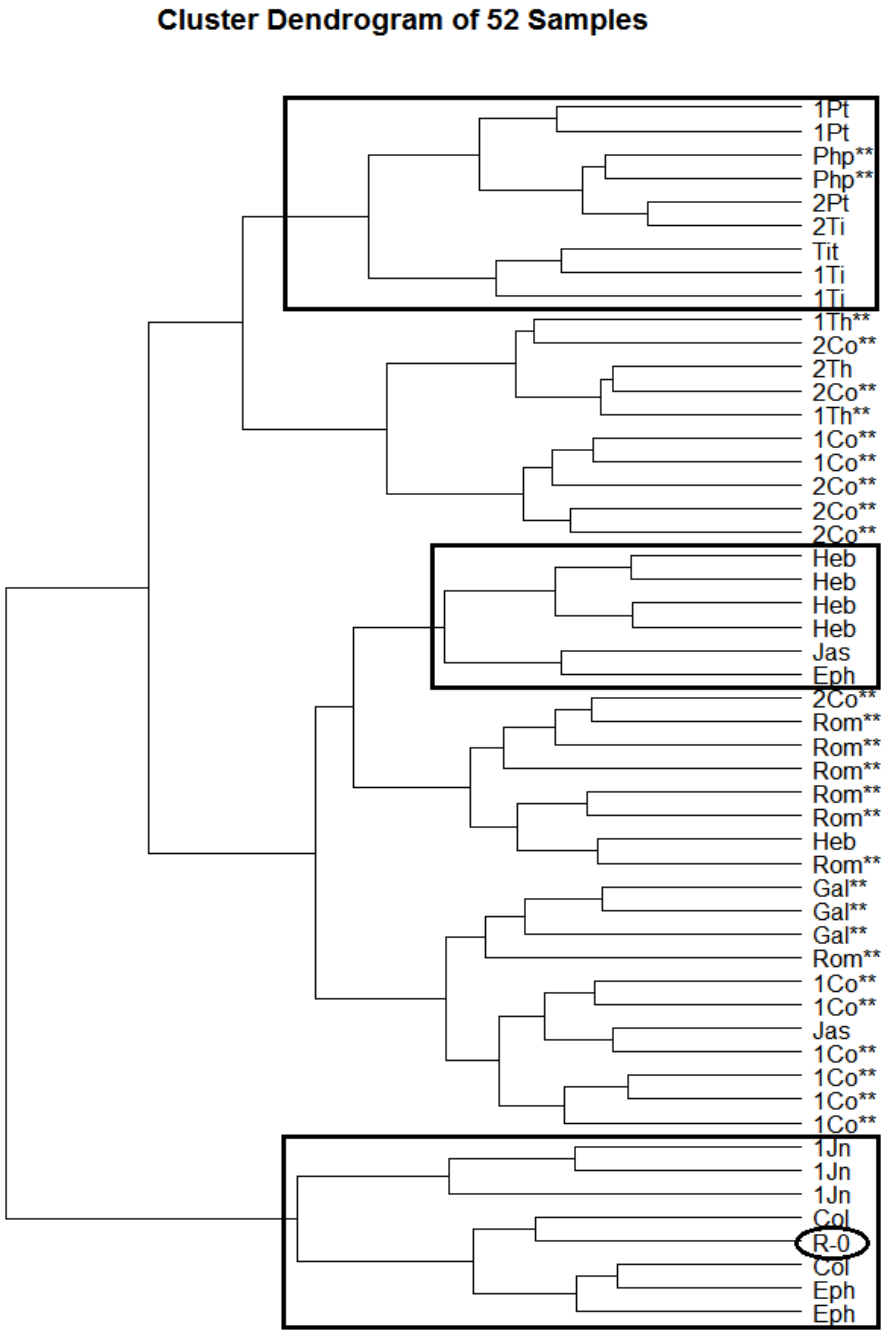

Figure 5: Hierarchical Cluster Dendrogram of 52 Partitions of New Testament Epistles Euclidian distance and "complete" clustering methods of 35 Most Frequent Lemmas Romans 1:18-2:29 ("R-0") enclosed in ellipse, Undisputed Paulines have double asterisk Data Analysis by R package {stats}, functions "hclust" and "dendrogram"

For ease of reference, all the undisputed Pauline leaf labels in Figure 5 are marked with double asterisks, and R-0 itself is enclosed in an ellipse. Three sections of interest are blocked off in bold rectangles. The bottom rectangle outlines the branch containing the R-0 leaf clustered between two leaves of Colossians, and those three are in turn clustered on either side with all of 1John and two of the three Ephesians leaves. Obviously, none of the undisputed Paulines are in this branch, and this was the last branch to be clustered with the rest of the tree. The middle rectangle outlines a minor branch with two nodes in which are clustered four of the five Hebrews leaves along with a clustering of the third leaf of Ephesians and one of James. This branch, too, contains no undisputed Paulines. The top rectangle surrounds another major branch containing the clusters of the two 1Peter leaves, the two Philippians leaves, 2 Peter with 2 Timothy, and finally the other three Pastorals.

Comparing the hclust graph of Figure 5 with the ca graph of Figure 4, we see that hclust agrees with the prior ca results in several ways: R-0 shows a definite separation from the rest of Romans and from the other undisputed Paulines segments; almost all of Hebrews is clustered together, separately from the undisputed Paulines; the Pastoral and Petrine segments cluster in the same area along with Philippians, and close to Thessalonians.

The third and final testing to be performed is Linear Discriminant Analysis (LDA). Unlike Correspondence Analysis and Cluster Analysis, LDA does not extract patterns from a dataset without predetermined parameters. Rather, LDA must first be given a set of "training" samples tagged with known group identifiers. After analyzing the training data, LDA can compare a test sample from an unknown group and assign it to the known group from the training set to which it is most similar. For example, we may give LDA a training set containing samples from two or more letters of known authorship, say Romans and Hebrews, and then ask LDA to assign the R-0 sample to which of Romans or Hebrews it is most similar in terms of the feature set we have been using.

A notable restriction in using LDA is that the number of variables in the feature set cannot exceed the number of samples in the smallest group being used, and ideally, that number should be one less than the smallest group sample count. Thus, in order to use, say the five most frequent lemmas as the feature set, there cannot be any group in the training set with less than five samples. LDA does have cross-validation tests to assist in assessing the validity of the results, regardless of the size of the feature set.

For LDA testing, the function "lda" in R package "MASS"[19] along with the Wilks Lambda discriminant validation test of the "summary.manova" function in R package "stats"[20] were both used. A Wilks Lambda p-value less than .05 (or 5%) indicates that the hypothesis of a valid multivariate discriminant function will **not** be rejected. Lda's cross-validation routine will also be used, employing a "leave one out" method whereby it assigns the group membership of each

[19] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[20] R Core Team (2012), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

sample in the training groups by calculating LDA without that sample, and then classifying the sample's membership on that calculation.

The column headed "LDA Label" in Table 1 shows the sample groupings set up for LDA. Preliminary testing revealed poor cross-validation for the "Pas" group of Pastorals samples, so that group was dropped from consideration. The "J_P" group of James and 1-2 Peter fared better on cross-validation, but was also dropped as it performed more poorly than either the Ephesians-Colossians or Hebrews group when tested against any or all of the undisputed Paulines.

Chart 1 below shows the results of using lda to run multiple testing of R-0 with three or two groups at a time. There are two major columns representing testing with two different feature sets. The left hand set of tables used the five most frequent lemmas, and the right hand tables used the first five coordinates of Correspondence Analysis (CA). Using CA coordinates lets lda take account of greater variance (54%) without introducing bias into the selection process itself.

Chart 1: LDA 3-Way and 2-Way Tables

	5 Most Frequent Lemmas				Top 5 CA Coordinates (54% of variance)			
Decision Tables								
	Rom	1Co	2Co	GPT	Rom	1Co	2Co	GPT
E_C, Heb	E_C	E_C	E_C	E_C	E_C	E_C	E_C	E_C
E_C	E_C	E_C	E_C	E_C	E_C	E_C	E_C	E_C
Heb	Heb	Heb	Heb	Heb	Heb	Heb	Heb	Heb
Cross-Validation Assignment Percentages								
	Rom	1Co	2Co	GPT	Rom	1Co	2Co	GPT
E_C, Heb	82.4%	88.9%	87.5%	94.1%	88.2%	94.4%	93.8%	88.2%
E_C	75.0%	100.0%	100.0%	100.0%	91.7%	100.0%	90.9%	75.0%
Heb	75.0%	92.3%	90.9%	100.0%	83.3%	100.0%	90.9%	100.0%
Wilks Lambda p-value								
	Rom	1Co	2Co	GPT	Rom	1Co	2Co	GPT
E_C, Heb	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
E_C	0.0172	0.0009	0.0028	0.0022	0.0280	0.0001	0.0017	0.0290
Heb	0.0340	0.0268	0.0036	0.0219	0.0062	0.0000	0.0011	0.0008

Within each major column are three minor tables, the "Decision Table" at the top showing the group to which lda assigned R-0 in a prediction test; the middle showing the percentages of correct group assignment by lda in cross-validation; and the bottom showing the Wilks Lambda p-values for the respective tests. The six minor tables have a common format. The four columns represent the four groups of Undisputed Pauline epistles, as designated in Table 1 at the end of

this paper, "Rom", "1Co", "2Co", and "GPT" ("GPT" = Galatians-Philippians-1 Thessalonians). The three rows represent the groups against which the four Pauline groups were tested, "E_C" designates Ephesians-Colossians and "Heb" designates Hebrews. The row labeled "E_C,Heb" indicates testing with both groups, and the other two rows indicate testing with one at a time.

Each intersecting cell in the minor tables displays the results of lda testing of the combination of groups represented by both the row and column labels for that cell. For example, the top left cell in the Decision Table on the left represents the test of Romans against both Ephesians-Colossians and Hebrews, and the "E_C" in that cell shows that lda's decision in this 3-way test was to assign R-0 to the Ephesians-Colossians group. The 82.5% in corresponding cell in the middle "Cross-Validation" table represents the percent of samples correctly assigned in that same test. Similarly, the corresponding cell in the bottom table shows a p-value of 0.0000 (rounded to 4 significant digits) for the Wilks Lambda multivariate discriminant test of the same combination .

Several observations about the LDA results are immediately evident:

- The results in the right column of tables track very closely with the corresponding results in the left column tables. The Decision Table cells are identical on both sides, while the cross-validation results in the middle tables are slightly better (higher) in the right side than the left. The Wilks Lambda p-values are also slightly better (lower) in the right side. The close similarity of results using either the five most frequent lemmas or the five first CA coordinates would seem to provide further internal confirmation to the test results.
- In the Decision Table, Ephesians-Colossians wins as the group to which lda assigns R-0, whenever tested with or without Hebrews against any of the four Pauline groups; Hebrews wins assignment for R-0 whenever tested by itself against any of the Pauline groups. These results are as consistent as they can be. Not once does any Pauline group win assignment
- Lda's cross-validation routine uses the "leave one out" method described above. The cross-validation results in the middle two tables are excellent: 15 of all 24 cells are over 90%, seven are a full 100%, and none are lower than 75%. Overall, cross-validation results in this table are about 90%, a very high accuracy rate for LDA[21].
- The Wilks Lambda numbers in the bottom set of tables are all well under .05, most are several magnitudes of order smaller. Thus the validity of a multivariate discriminant function of any of the test combinations will not be rejected.

The only test left is to make a 2-way test of Ephesians-Colossians directly with Hebrews. Chart 2 below summarizes the results of that test, using both feature sets. The tests with the five most frequent lemmas are in the left column, while the right column has the same tests using the first five CA coordinates.

The result shows a split decision. The choice of assigning R-0 to Hebrews using CA coordinates has higher cross-validation percentage and a lower p-value, both ranking slightly better than the Ephesians-Colossians assignment in the left column based on lemma frequencies. Ephesians-Colossians, however, won all the previous 3-way assignments in Chart 1 against both Hebrews and the Pauline groups. Perhaps not coincidentally, the CA graph in Figure 4 above also mapped

[21] Mealand, "Q", 495-495, notes that cross-validation success rates over 90% would be considered quite high for LDA or any other text classification model.

R-0 closer to Hebrews than to the Ephesians and Colossians samples, although the latter were close by, thus giving support to the Hebrews choice. On the other hand, the Cluster Analysis graph in Figure 5 clustered R-0 immediately with Colossians and Ephesians samples, away from Hebrews, supporting the choice of Ephesians as the best fit for R-0.

Chart 2: LDA 2-Way Test, Eph-Col versus Hebrews
Two Feature Sets

	Top 5 Lemmas	Top 5 CA Coor
Eph-Col vs Hebrews	Eph-Col	Hebrews
Total Cross-Val %	80.00%	90.00%
Wilks-Lambda p.value	0.0153	0.0051

In the end all three analysis methods agree on the basis of the same 35 lemma feature set that Romans 1:18-29, is stylometrically separate from the rest of Romans and from all the other undisputed Pauline samples, and is better classified in the style of Ephesians-Colossians or Hebrews. That has been the consistent result of all the testing done in this study. The fact that Correspondence Analysis, Cluster Analysis, and Linear Discriminant Analysis all agree on this finding makes the conclusion well founded that we may attribute the authorship of Romans 1:18-2:29 to someone other than the author of the rest of Romans or of any other undisputed Pauline letter. Rather, we may with significant confidence attribute this text to someone writing in a style which, for whatever reason, lies closest to that of Ephesians-Colossians or Hebrews. Thus, sylometry firmly supports the proposals of Sanders and Walker.

Table 1: 52 Samples of Partitioned New Testament Epistolary Text

Partition Definitions and Word/Feature Counts

Excludes Old Testament Quotations and Wm O. Walker non-Pauline Interpolations

ID	Graph Label	LDA Label	Starting Word	Ending Word	Word Counts	Feature Counts	Feature/ Word %
Rom.0	R-0	R-0	06.Ro_01:18.01	06.Ro_02:29.23	709	334	47.1%
Rom.1	Ro(m)	Rom	06.Ro_01:01.01	06.Ro_04:11.01	709	358	50.5%
Rom.2	Ro(m)	Rom	06.Ro_04:11.02	06.Ro_06:03.07	709	374	52.8%
Rom.3	Ro(m)	Rom	06.Ro_06:03.08	06.Ro_07:20.11	709	377	53.2%
Rom.4	Ro(m)	Rom	06.Ro_07:20.12	06.Ro_08:38.10	709	378	53.3%
Rom.5	Ro(m)	Rom	06.Ro_08:38.11	06.Ro_11:06.06	709	386	54.4%
Rom.6	Ro(m)	Rom	06.Ro_11:06.07	06.Ro_13:09.21	709	362	51.1%
Rom.7	Ro(m)	Rom	06.Ro_13:09.22	06.Ro_15:19.05	709	354	49.9%
1Co.1	1C(o)	1Co	07.1Co_01:01.01	07.1Co_03:11.01	709	353	49.8%
1Co.2	1C(o)	1Co	07.1Co_03:11.02	07.1Co_05:11.17	709	346	48.8%
1Co.3	1C(o)	1Co	07.1Co_05:11.18	07.1Co_07:22.06	709	370	52.2%
1Co.4	1C(o)	1Co	07.1Co_07:22.07	07.1Co_09:10.18	709	382	53.9%
1Co.5	1C(o)	1Co	07.1Co_09:10.19	07.1Co_11:31.05	709	377	53.2%
1Co.6	1C(o)	1Co	07.1Co_11:31.06	07.1Co_14:14.01	709	321	45.3%
1Co.7	1C(o)	1Co	07.1Co_14:14.02	07.1Co_15:26.03	709	346	48.8%
1Co.8	1C(o)	1Co	07.1Co_15:26.04	07.1Co_16:24.09	705	328	46.5%
2Co.1	2C(o)	2Co	08.2Co_01:01.01	08.2Co_02:14.15	709	406	57.3%
2Co.2	2C(o)	2Co	08.2Co_02:14.16	08.2Co_05:02.09	709	376	53.0%
2Co.3	2C(o)	2Co	08.2Co_05:02.10	08.2Co_07:13.03	709	360	50.8%
2Co.4	2C(o)	2Co	08.2Co_07:13.04	08.2Co_09:13.16	709	381	53.7%
2Co.5	2C(o)	2Co	08.2Co_09:13.17	08.2Co_11:24.03	709	382	53.9%
2Co.6	2C(o)	2Co	08.2Co_11:24.04	08.2Co_13:10.10	709	356	50.2%
Gal.1	Ga(l)	GPT	09.Gal_01:01.01	09.Gal_02:20.26	709	345	48.7%
Gal.2	Ga(l)	GPT	09.Gal_02:20.27	09.Gal_04:22.02	709	337	47.5%
Gal.3*	Ga(l)	GPT	09.Gal_04:22.03	09.Gal_06:18.13	694	327	47.1%
Eph.1	Ep(h)	E_C	10.Eph_01:01.01	10.Eph_02:18.11	709	398	56.1%
Eph.2	Ep(h)	E_C	10.Eph_02:18.12	10.Eph_04:22.14	709	360	50.8%
Eph.3	Ep(h)	E_C	10.Eph_04:22.15	10.Eph_06:11.08	709	364	51.3%
Php.1	Ph(p)	GPT	11.Php_01:01.01	11.Php_02:15.14	709	349	49.2%
Php.2	Ph(p)	GPT	11.Php_02:15.15	11.Php_04:09.13	709	348	49.1%
Col.1	Co(l)	E_C	12.Col_01:01.01	12.Col_02:11.06	709	382	53.9%
Col.2	Co(l)	E_C	12.Col_02:11.07	12.Col_04:09.09	709	363	51.2%
1Th.1	1Th	GPT	13.1Th_01:01.01	13.1Th_03:12.13	709	380	53.6%
1Th.2*	1Th	GPT	13.1Th_03:12.14	13.1Th_05:28.09	663	326	49.2%
2Th.1	2Th	Pas	14.2Th_01:01.01	14.2Th_03:10.09	709	367	51.8%
1Ti.1	1Ti	Pas	15.1Ti_01:01.01	15.1Ti_04:01.10	709	282	39.8%
1Ti.2	1Ti	Pas	15.1Ti_04:01.11	15.1Ti_06:11.13	709	298	42.0%

Table 1 – Continued							
ID	Graph Label	LDA Label	Starting Word	Ending Word	Word Counts	Feature Counts	Feature/ Word %
2Ti.1	2Ti	Pas	16.2Ti_01:01.01	16.2Ti_03:06.01	709	321	45.3%
Tit.1*	Ti(t)	Pas	17.Tit_01:01.01	17.Tit_03:15.17	659	264	40.1%
Heb.1	He(b)	Heb	19.Heb_01:01.01	19.Heb_04:12.02	709	344	48.5%
Heb.2	He(b)	Heb	19.Heb_04:12.03	19.Heb_07:07.07	709	319	45.0%
Heb.3	He(b)	Heb	19.Heb_07:07.08	19.Heb_09:15.05	709	330	46.5%
Heb.4	He(b)	Heb	19.Heb_09:15.06	19.Heb_11:04.18	709	301	42.5%
Heb.5	He(b)	Heb	19.Heb_11:04.19	19.Heb_12:10.08	709	289	40.8%
Jam.1	Ja(s)	J_P	20.Jam_01:01.01	20.Jam_02:20.06	709	341	48.1%
Jam.2	Ja(s)	J_P	20.Jam_02:20.07	20.Jam_05:04.04	709	333	47.0%
1Pt.1	1P(t)	J_P	21.1Pt_01:01.01	21.1Pt_02:25.09	709	317	44.7%
1Pt.2	1P(t)	J_P	21.1Pt_02:25.10	21.1Pt_05:07.04	709	326	46.0%
2Pt.1	2P(t)	J_P	22.2Pt_01:01.01	22.2Pt_02:20.16	709	308	43.4%
1Jn.1	1J(n)		23.1Jn_01:01.01	23.1Jn_02:25.13	709	433	61.1%
1Jn.2	1J(n)		23.1Jn_02:26.01	23.1Jn_04:07.16	709	422	59.5%
1Jn.3	1J(n)		23.1Jn_04:07.17	23.1Jn_05:20.27	709	432	60.9%
				Column Totals	**36,753**	**18,313**	**49.8%**

* Gal.3, 1Th.2 and Tit.1 Feature vector sums adjusted in analysis calculations to 329, 346 and 282, respectively

Table 2: William O Walker's Non-Pauline Interpolations
Texts Other than Romans 1:18-2:29 Within Undisputed Paulines
Excluded from Data Analysis Computations

ID	Starting Word	Ending Word	Word Counts	Word Counts Excl OT Quotes
Rom.i.1	06.Ro_13:01.01	06.Ro_13:07.24	144	144
Rom.i.2	06.Ro_16:25.01	06.Ro_16:27.13	53	53
1Co.i.1	07.1Co_02:06.01	07.1Co_02:16.13	211	196
1Co.i.2	07.1Co_10:01.01	07.1Co_10:22.08	318	309
1Co.i.3	07.1Co_11:03.01	07.1Co_11:16.16	213	213
1Co.i.4	07.1Co_12:31.07	07.1Co_14:01.03	206	206
1Co.i.5	07.1Co_14:34.01	07.1Co_14:35.18	36	36
1Co.i.6	07.1Co_15:29.01	07.1Co_15:34.14	81	75
2Co.i.1	08.2Co_06:14.01	08.2Co_07:01.19	111	77
Gal.i.1	09.Gal_02:07.10	09.Gal_02:08.14	18	18
1Th.i.1	13.1Th_02:13.01	13.1Th_02:16.22	109	109
		Word Count Sums:	**1500**	**1436**

Table 3: Feature Selection

35 Most Frequent Lemmas in New Testament Epistles
Exclusive of Content Words

Lemma	Gloss	Freqency	Cumulative Sums	Cumulative %
ὁ	the	6130	6130	13.64%
καί	and	2219	8349	18.57%
ἐν	in	1300	9649	21.47%
σύ	you	1086	10735	23.88%
ἐγώ	I	989	11724	26.08%
αὐτός	he/she/it	916	12640	28.12%
δέ	but	819	13459	29.94%
εἰμί	to be	786	14245	31.69%
οὐ	not	662	14907	33.16%
γάρ	for	594	15501	34.48%
εἰς	into	586	16087	35.79%
ὅς	who/what	552	16639	37.02%
μή	not	466	17105	38.05%
ὅτι	because	422	17527	38.99%
οὗτος	this	408	17935	39.90%
διά	through	385	18320	40.76%
ἀλλά	but	376	18696	41.59%
ἵνα	in order to	312	19008	42.29%
ἐκ	from	294	19302	42.94%
κατά	according to	261	19563	43.52%
εἰ	if	259	19822	44.10%
ἔχω	to have	256	20078	44.67%
ὡς	as	226	20304	45.17%
τις	some	209	20513	45.63%
λέγω	to say	204	20717	46.09%
γίνομαι	to become	194	20911	46.52%
ἐπί	upon	186	21097	46.93%
πρός	to(ward)	184	21281	47.34%
ἀπό	from	166	21447	47.71%
ἑαυτοῦ	him/her-self	151	21598	48.05%
ἤ	or	145	21743	48.37%
ποιέω	to do/make	139	21882	48.68%
οὖν	therefore	137	22019	48.98%
τίς	who/what?	133	22152	49.28%
ἐάν	if	132	22284	49.57%

Word Count in NT Epistles, UBS3-4/NA26-27 Text, Romans-Jude: 44951

Features Word Count: 22284

Features Cumulative Percentage of NT Epistolary Text: 49.57%

Table 4: Frequencies of 35 Features in 52 Samples

ID	Sums	ὁ	καί	ἐν	σύ	ἐγώ	αὐτός	δέ	εἰμί	οὐ	γάρ	εἰς	ὅς	μή	ὅτι	οὗτος	διά	ἀλλά
Rom.0	334	122	39	25	6	1	24	8	12	10	13	9	6	5	4	3	5	5
Rom.1	358	95	25	25	15	11	14	12	11	8	15	10	8	7	6	2	13	5
Rom.2	374	124	28	13	0	15	12	13	14	12	12	18	11	5	7	3	27	9
Rom.3	377	105	13	12	11	22	9	22	14	14	21	14	14	7	10	8	11	9
Rom.4	378	117	20	22	7	24	13	23	12	14	18	6	11	1	10	8	8	9
Rom.5	386	94	21	9	5	14	15	17	12	21	18	13	13	9	11	7	2	13
Rom.6	362	119	25	13	19	6	18	16	6	5	14	10	3	17	2	7	5	9
Rom.7	354	108	34	15	18	11	5	16	7	8	17	12	9	18	4	4	7	5
1Co.1	353	91	30	27	23	24	8	18	22	13	12	3	4	10	7	3	6	8
1Co.2	346	73	29	26	28	20	2	17	16	10	14	3	7	8	4	10	5	7
1Co.3	370	95	29	16	16	8	7	21	16	22	7	1	5	13	7	10	5	14
1Co.4	382	96	34	13	9	14	11	23	22	18	7	5	7	20	5	11	6	4
1Co.5	377	94	27	13	18	22	4	14	12	17	15	7	7	11	5	13	5	10
1Co.6	321	82	23	16	11	5	8	32	23	15	8	5	3	16	4	4	2	5
1Co.7	346	81	31	18	14	12	7	26	18	16	10	3	11	4	12	1	3	8
1Co.8	328	93	33	21	22	9	11	26	8	10	11	4	4	3	6	7	2	6
2Co.1	406	95	40	19	31	41	10	10	11	10	12	15	10	8	11	7	11	9
2Co.2	376	115	17	27	7	23	8	14	11	14	13	9	6	10	7	7	9	18
2Co.3	360	77	30	39	20	28	9	12	5	10	12	6	3	6	8	4	9	16
2Co.4	381	102	37	18	34	20	12	13	7	10	8	18	3	6	9	13	7	9
2Co.5	382	76	26	23	27	18	10	12	11	21	16	18	13	6	7	4	5	8
2Co.6	356	65	30	29	29	29	7	10	11	25	14	7	6	12	8	7	4	8
Gal.1	345	80	32	14	10	30	14	17	13	12	8	13	12	5	10	2	8	7
Gal.2	337	82	17	11	23	15	5	20	22	12	10	7	7	4	12	3	7	8
Gal.3 *	327	89	18	13	16	13	3	21	18	8	16	7	4	14	6	7	4	8
Eph.1	398	137	37	49	14	20	33	2	16	4	3	13	14	2	3	4	8	1
Eph.2	360	139	32	36	12	12	15	8	12	2	0	15	12	2	2	4	9	1
Eph.3	364	106	47	22	14	4	14	9	16	4	8	5	4	11	7	6	2	10
Php.1	349	81	47	34	24	24	9	8	8	4	6	11	2	3	10	9	9	7
Php.2	348	87	44	22	19	25	21	12	6	7	6	8	11	2	6	6	5	7
Col.1	382	127	46	42	22	13	25	2	14	3	2	13	20	2	3	3	9	1
Col.2	363	111	42	39	23	8	16	3	11	5	3	5	16	8	2	4	5	2
1Th.1	380	93	50	26	49	32	6	4	3	7	11	11	2	3	6	4	5	7
1Th.2	326	83	42	26	32	14	14	10	7	10	11	13	1	11	5	5	4	5
2Th.1 *	367	100	48	22	35	24	15	7	6	6	4	14	11	8	10	3	8	3
1Ti.1	282	52	40	31	5	12	3	13	12	5	3	10	11	13	6	6	3	5
1Ti.2	298	77	47	11	7	1	2	17	15	3	10	6	6	10	6	12	3	6
2Ti.1	321	76	35	23	12	22	7	13	9	9	7	8	14	3	6	9	10	8
Tit.1 *	264	61	37	13	8	19	8	8	16	1	6	2	9	14	1	5	3	4
Heb.1	344	102	28	15	3	6	25	10	8	8	18	11	11	4	1	7	15	4
Heb.2	319	101	54	4	5	4	16	12	7	6	11	8	13	4	0	4	6	2
Heb.3	330	104	34	6	0	3	9	14	12	10	18	11	15	1	3	8	14	1
Heb.4	301	105	30	11	2	5	14	12	6	6	15	17	6	3	1	10	5	4
Heb.5	289	94	41	7	3	7	16	7	6	9	11	9	11	7	5	4	6	1
Jam.1	341	86	33	18	15	12	21	22	12	9	10	7	4	11	7	4	1	3
Jam.2	333	96	56	14	21	7	12	11	16	14	5	6	2	6	6	4	1	2
1Pt.1	317	73	31	21	25	3	14	12	5	9	4	21	16	3	6	5	11	7
1Pt.2	326	97	28	22	21	1	8	11	5	2	6	14	12	9	7	4	5	9
2Pt.1	308	78	38	31	13	17	14	14	8	11	12	6	11	2	2	14	4	4
1Jn.1	433	108	51	28	22	20	28	6	34	19	1	1	13	4	23	9	1	6
1Jn.2	422	109	43	23	11	15	42	2	34	15	0	3	9	9	24	14	2	3
1Jn.3	432	140	37	28	1	22	32	3	30	14	2	5	8	7	29	15	2	4

*Frequency counts for Gal.1, 1Th.2, and Tit.1 adjusted in analysis calculations to compensate for lack of full 709 word count.

Table 4: Frequencies of 35 Features in 52 Samples – Continued

ID	ἵνα	ἐκ	κατά	εἰ	ἔχω	ὡς	τις	λέγω	γίνομαι	ἐπί	πρός	ἀπό	ἑαυτοῦ	ἤ	ποιέω	οὖν	τίς	ἐάν
Rom.0	0	5	5	1	4	1	0	1	1	3	0	2	2	3	4	2	0	3
Rom.1	4	7	8	5	2	2	4	8	6	4	4	1	0	4	2	7	8	0
Rom.2	4	9	3	4	2	4	2	2	2	6	1	2	2	2	1	4	1	0
Rom.3	5	5	1	6	2	0	0	3	9	2	0	6	3	2	4	7	3	3
Rom.4	2	3	11	8	2	0	1	1	0	1	2	3	2	6	1	3	9	0
Rom.5	2	18	6	2	3	3	2	20	3	2	2	2	0	4	2	9	10	2
Rom.6	5	5	8	13	3	2	3	6	5	3	0	1	3	0	0	6	3	2
Rom.7	4	3	3	2	4	2	2	3	2	0	2	1	7	4	1	8	3	5
1Co.1	8	1	5	3	0	5	5	5	2	1	2	2	0	2	0	1	2	0
1Co.2	9	2	1	6	3	8	10	0	7	1	3	1	1	7	0	1	4	3
1Co.3	2	2	2	10	9	4	7	4	2	4	3	2	3	10	2	3	5	4
1Co.4	4	2	2	6	12	8	6	4	3	4	2	1	2	4	4	2	3	8
1Co.5	14	5	0	8	3	5	6	5	9	2	0	1	5	3	5	4	3	5
1Co.6	8	5	1	7	6	2	1	5	0	0	4	0	1	7	0	1	2	9
1Co.7	2	3	6	10	5	1	6	5	6	3	1	1	1	8	0	4	4	5
1Co.8	8	2	2	3	0	1	6	3	6	1	6	0	2	2	1	2	0	7
2Co.1	10	4	2	5	5	1	3	0	3	5	6	5	2	3	0	1	1	0
2Co.2	4	10	2	5	6	5	2	1	1	3	7	3	6	2	0	1	1	2
2Co.3	7	4	5	7	3	11	1	3	2	3	8	2	6	0	1	3	0	0
2Co.4	10	5	3	2	4	4	2	3	3	6	2	3	1	1	2	2	0	2
2Co.5	5	0	11	10	2	8	12	3	0	4	3	2	10	4	3	1	2	1
2Co.6	8	6	2	7	1	2	4	2	1	2	5	1	3	2	3	1	3	2
Gal.1	7	10	6	7	1	1	4	2	1	0	5	5	2	4	1	0	0	2
Gal.2	7	16	4	7	0	6	0	6	9	4	2	1	1	3	0	4	2	0
Gal.3	3	8	7	6	3	1	4	3	2	3	2	2	4	0	3	2	4	5
Eph.1	5	3	9	0	3	1	1	2	1	5	0	1	0	0	4	0	3	0
Eph.2	6	2	10	3	1	1	0	2	1	3	5	1	2	1	3	3	3	0
Eph.3	6	3	4	1	4	13	3	3	5	2	3	1	12	6	3	3	2	1
Php.1	7	3	4	5	4	4	6	0	4	2	1	3	5	0	2	1	2	0
Php.2	5	6	5	7	6	3	4	3	2	4	3	0	1	1	0	4	0	0
Col.1	5	2	6	2	3	1	1	1	3	2	0	7	0	0	0	1	1	0
Col.2	4	5	7	2	3	6	5	0	1	4	5	2	2	4	2	4	1	3
1Th.1	0	6	0	0	3	6	2	0	11	4	11	6	4	3	1	0	2	2
1Th.2	6	0	0	1	5	3	2	2	0	1	2	3	2	0	3	2	1	0
2Th.1	5	1	4	0	1	2	3	2	1	4	4	8	3	1	2	1	0	1
1Ti.1	8	1	3	3	5	0	7	3	1	2	2	2	3	3	2	3	1	3
1Ti.2	6	1	3	10	8	4	10	1	3	3	3	1	1	2	2	1	0	0
2Ti.1	3	5	5	4	6	4	3	2	2	3	1	3	2	1	0	3	0	3
Tit.1	13	4	8	1	2	2	1	2	1	2	5	2	2	2	1	0	1	0
Heb.1	4	4	4	3	2	4	8	11	5	2	4	4	1	0	3	4	5	0
Heb.2	4	6	7	1	12	1	2	3	7	3	6	4	3	0	1	2	0	0
Heb.3	0	2	12	5	9	1	1	10	8	6	2	3	3	0	1	3	1	0
Heb.4	3	2	8	0	6	0	3	6	4	6	1	2	3	1	1	3	0	0
Heb.5	4	2	3	2	5	6	1	3	5	3	3	3	1	1	1	0	2	0
Jam.1	1	2	2	6	7	3	8	8	6	2	0	3	5	3	4	0	2	4
Jam.2	1	9	2	4	3	1	1	5	4	2	2	1	0	5	7	3	2	2
1Pt.1	5	7	5	6	2	12	1	0	1	4	1	1	1	2	0	2	1	0
1Pt.2	8	1	5	8	3	11	4	0	4	1	2	2	3	6	0	4	2	1
2Pt.1	1	3	1	2	5	4	1	0	4	1	1	1	1	0	4	0	0	0
1Jn.1	5	7	0	2	9	1	2	6	1	0	2	10	1	0	3	0	1	9
1Jn.2	7	20	0	2	7	1	1	0	0	1	1	6	1	1	8	0	2	6
1Jn.3	7	7	1	3	12	0	3	2	0	0	5	2	1	0	2	0	1	7