New Test. Stud. 57, pp. 483–507. © Cambridge University Press, 2011
doi:10.1017/S002868851100018X

Is there Stylometric Evidence for Q?

DAVID L. MEALAND
Fellows Room, New College, University of Edinburgh, Mound Place, Edinburgh EH1 2LX, UK.
email: D.Mealand@ed.ac.uk

Stylometric tests were run to assess whether, in Matthew, Q material differs in style from that of M. Correspondence Analysis was used on larger samples. Then counts of the five most frequent words in smaller samples were tested using three further methods: GLM, Discriminant Analysis and Cluster Analysis. These tests assigned about 80% of the samples to the expected source. This result permits a cautious preference for the Two Source Theory against the theory upheld by Farrer, Goulder and Goodacre.

Keywords: Greek, Gospels, source, stylometry, statistics, Synoptic

1. Introduction

This paper offers a stylometric perspective on a topic which has provoked much recent debate. That debate has seen literary 'Excavations' by Kloppenborg Verbin,[1] a vigorous Case against Q by Goodacre,[2] a 581 page Critical Edition of Q[3] and much more. There has been an Aramaic Approach by Casey,[4] replies to Goodacre by Foster and Kloppenborg,[5] proposals from Watson[6] for a redactional

1 J. S. Kloppenborg Verbin, *Excavating Q: The History and Setting of the Sayings Gospel* (Minneapolis: Fortress, 2000).
2 M. S. Goodacre, *The Case Against Q: Studies in Markan Priority and the Synoptic Problem* (Harrisburg, PA: Trinity, 2002).
3 J. M. Robinson et al., eds., *The Critical Edition of Q: Synopsis including the Gospels of Matthew and Luke, Mark and Thomas with English, German, and French Translations of Q and Thomas* (Minneapolis: Fortress; Leuven: Peeters, 2000).
4 P. M. Casey, *An Aramaic Approach to Q* (Cambridge: Cambridge University, 2002).
5 P. Foster, 'Is it Possible to Dispense with Q?', *NovT* 45 (2003) 313–37; J. S. Kloppenborg, 'On Dispensing with Q? Goodacre on the Relation of Luke to Matthew', *NTS* 49.2 (2003) 210–36; see also J. S. Kloppenborg, 'Variation in the Reproduction of the Double Tradition and an Oral Q?' *ETL* 83 (2007) 53–80.
6 F. Watson, 'Q as Hypothesis: A Study in Methodology', *NTS* 55.4 (2009) 397–415.

approach and further analysis of Q by Burkett.[7] The aim in this paper is to use stylometric methods to see if there are, or are not, significant stylistic differences between material attributed to M and to Q.

Statistical work on Q has been carried out for several decades, most of it focused on the patterns of verbal agreement between the texts, and the extent to which they are identical in form and sequence. A thorough and impressive article on these statistical studies was published recently by John Poirier.[8] The present study is proposing a rather different approach by testing whether there are *stylometric* differences between Q and M material in Matthew. Some earlier stylometric work by Linmans[9] and by Sewell[10] suggested that the answer may be negative or doubtful. Their work deserves attention and respect for the care with which it was conducted, but it may still be the case that stylistic differences between the sources could be located from a fresh analysis of the data. Statistical work has also been done by D. Gentile[11] using data in which the Synoptic material is classified into 19 categories.

7 D. Burkett, *Rethinking the Gospel Sources*. Vol. 2. *The Unity and Plurality of Q* (Atlanta: SBL, 2009).

8 J. C. Poirier, 'Statistical Studies of the Verbal Agreements and their Impact on the Synoptic Problem', *CBR* 7 (2008) 68–123.

9 A. J. M. Linmans, *Onderschikking in de Synoptische Evangeliën* (Leiden: FSW, 1995) esp. 96–9, 319. Linmans made careful and extensive use of Log-Linear Analysis and Correspondence Analysis. He divided the text of the Synoptics by classifying very short units as narrative, dialogue or sayings, and compared large blocks of these, but with less emphasis on partitioning the data to determine within-group as against between-group differences or similarities. He covered the whole of the three Synoptic gospels, and focused on text type (discourse type) preferences and gospel preferences. He found little evidence of source preferences. The present paper focuses more on possible source differences within Matthew, using more data partitioning, in order to try more specifically to pursue that particular aspect of the data.

10 P. Sewell, 'The Synoptic Problem: A Stylometric Contribution Regarding Q', *Colloquium (The Australian and New Zealand Theological Review)* 33 (2001) 59–74, 153–68, esp. 73. Sewell made careful use of a method which compared sections of Matthew against a set of passages made up from different parts of the NT; it also compared sections of Matthew directly against each other. It required a p value of 1%, which means that it may have rejected some differences which were significant but not highly significant. He concluded there was some variability within M, but that Q passages are not in general markedly different from the rest of Matthew. But it would seem that Q sayings and M sayings were not directly compared like for like.

11 D. Gentile, http://www.davegentile.com/synoptics/main.html (26.01.2010). The study reported on this website was done by the statistician D. Gentile with some collaboration by another statistician David Inglis. The source gives information about the origin and processing of the data. Large numbers of the more frequent words in 19 Synoptic categories were included in the counts. The statistics checked correlations based on comparisons adjusted for the varying size of the texts for each category. Though the original source data did not include some of the most frequent words, did use many content words and did not sub-classify by genre, or allow further partitioning, the careful analysis of correlations repays attention.

The experiments reported here look at two sets of samples drawn from material commonly attributed to different Synoptic sources. The first ones are based on counts of 30 words sufficiently frequent in 1000 word samples from all three Synoptic Gospels. The main series used similar counts, but only of the 5 most frequent words in a set of 60 out of 61 much smaller samples from Matthew.[12] (A sample containing the genealogy was normally excluded.) These two sets of data form the evidence on which this paper depends.

In this paper the terms 'Q' or 'Q material' will be used to denote the sets of verses from the double tradition common to Matthew and Luke, and attributed to the Q source by the Two Source Theory (2ST). The question at issue is whether the style of the Q material does or does not provide evidence to raise the probability that it comes from a distinct source. Similarly the term 'Markan material' denotes Matthean verses paralleled in Mark, while 'M material' or 'M' denotes verses not paralleled elsewhere in the Synoptics outside Matthew.

The question at issue here is whether, after allowing for genre differences within Matthew, there are stylistic differences between the Q material and passages found only in Matthew and attributed to M. The aim is to test the null hypothesis that there are no such significant differences. The result could then favour one of two main rival theories about the Synoptic problem over the other.

2. Samples and Criteria

The investigation used two sets of samples. There is a set of 35 larger samples. Each of these contains 1000 words. There are 12 such samples from Matthew, 10 from Mark and 13 from Luke. These 35 samples are divided by genre and by proposed source attribution.[13] The tests on the set of smaller samples used counts of the 5 most frequent words in the 60 samples of 250 words each drawn from Matthew. Tests using these explore more precisely whether or not there are differences within Matthew between samples attributed by the 2ST to Mark or to Q or to M, especially between the latter pair.

The aim here is not to attempt to test all or even most of the theories, but to look specifically at the choice between two of the most widely supported rivals: the 2ST and the Farrer–Goulder theory (FGT) that Luke made use of

Genre was considered at the interpretative stage. The main conclusions supported Markan priority and though initially the author was cautious about the merits of FGT, 2ST and 3ST (Three Source Theory), in recent discussion he has inclined more towards the latter.

12 See List A in the Appendix. Details of the numeric data can be obtained from the author by email, requesting file syndat1a.doc.

13 See Lists B, C and D in the Appendix.

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

Matthew.[14] If there are significant differences between the relevant sets of samples, then that favours the 2ST against one of its main rivals. If there are no such differences, then that will favour the rival theory that Luke used Matthew. The main task is to examine the stylistic evidence to see what results actually emerge from a fresh scrutiny of the evidence using different methods.

The size of the second set of samples is small. At only 250 words it is one quarter the size of the samples more commonly used. It is not the smallest size that some stylistic studies have used. Henri H. Somers[15] used samples as small as 80 words on Paul, Philo and the LXX. H. H. Greenwood[16] used samples the size of each chapter in the epistles. (For Romans the average is 445 words, for Titus 220 words.) Then Forsyth and Holmes[17] showed 200-word samples to be viable in benchmark tests. If it were the case that very rare words were being used as the criteria then huge amounts of text might be needed, but this is not the case. The tests do involve small samples of text but the criteria to be used will be the most frequent items in those samples. So the plan is to use items which are very frequent, and the 5 most frequent words in Matthew are the article, καί, αὐτός, δέ, and ἐν. In stylometric work in the last twenty years on a range of different literatures it is very common to use high-frequency function words such as these[18] and that is the preferred method here.

The choice of smaller samples, and the use of only the most frequent words, is also the result of other considerations. Many studies have shown texts to vary in style between speech and narrative.[19] It is therefore essential to allow for genre, and to partition the samples accordingly. Here they are classified either as samples of speech, or of narrative or of mixed genre. The latter includes parables and apophthegms or pronouncement stories. Obviously the texts also have to be

14 It would, in theory, be equally possible to test for the use of Luke by Matthew, with similar methods to those reported here on Matthew, to see if there is, or is not, a clear stylistic difference in Luke between double- and single-tradition material. In practice in addition to the paucity of narrative in the double tradition, there might be difficulty in finding enough sayings samples in the Lukan single tradition, as parables and apophthegms are more prominent in that material.

15 H. H. Somers, 'Statistical Methods in Literary Analysis', *The Computer and Literary Style* (ed. J. Leed; Kent, OH: Kent State, 1966) 128–40.

16 H. H. Greenwood, 'St. Paul Revisited: A Computational Result', *Literary and Linguistic Computing* 7 (1992) 43–7.

17 R. S. Forsyth and D. I. Holmes, 'Feature Finding for Text Classification', *Literary and Linguistic Computing* 11 (1996) 163–74, esp. 164, 170.

18 D. I. Holmes, 'The Evolution of Stylometry in Humanities Scholarship', *Literary and Linguistic Computing* 13 (1998) 111–17, esp. 113–14.

19 D. Biber, *Dimensions of Register Variation: A Cross-linguistic Comparison* (Cambridge: Cambridge University, 1995) esp. 153, 165, 237; J. F. Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Oxford: Clarendon, 1987) esp. 163–75.

divided by possible source. With three genres and three possible sources the 18,346 words of Matthew would in theory yield 9 samples of 2038 words each, but in practice rather less. Further, each of those 9 blocks has to be *partitioned* into several smaller samples in order to obtain a mean and standard deviation. Only in this way can proper statistical assessment be made both of differences *between* groups and also of the coherence or diversity of style *within* each group. This last point is a very crucial one.

3. The Statistical Methods Used

Most literary statistics are now performed using multivariate methods which have the considerable advantage of using several variables, or key word criteria, in combination in order to analyse and display the profile of sections of text. The specific methods used here were Correspondence Analysis (CA), the General Linear Model (GLM), two forms of Discriminant Analysis (Discrim and Candisc) and Cluster Analysis. Each method will be discussed in turn together with the results derived from it.

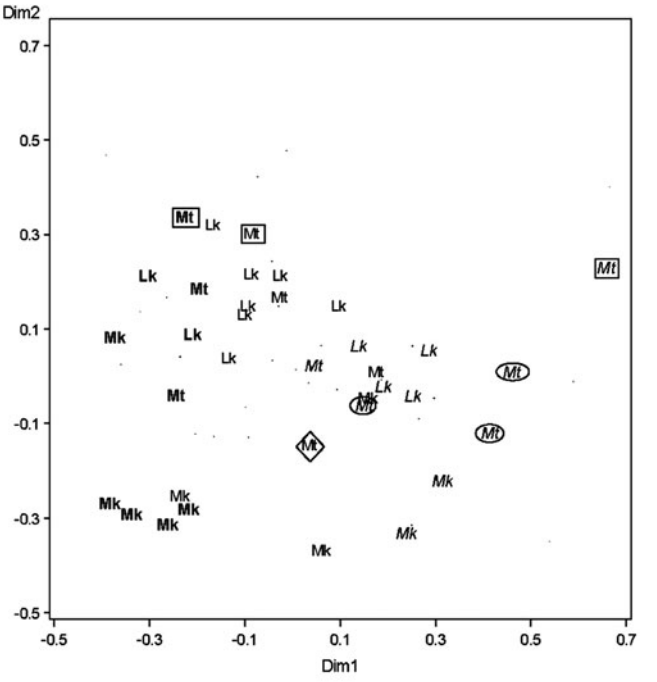
4. Using Correspondence Analysis

The first of these methods, CA, is neutral in that it does not know the groups to which the samples are attributed. It looks at all the samples, analyses the frequency of all the key words in these samples and then proceeds to plot the main differences between these sections of text in such a way that the samples can be located in two or three (or more) dimensional displays. (On the method see Greenacre.[20]) The extent to which the texts do, or do not, fall into groups is then sometimes very evident. So in Figure 1, using large samples of 1000 words from all the texts, it is very clear that the narrative samples are almost all on one side of the plot, and the samples of sayings are almost all on the other side of the plot. The difference of genre between speech and narrative is one of the most fundamental explanations of differences of style,[21] and appears on the main (horizontal) axis.

The output in Figure 1 shows the location of 35 large (1000 word) samples drawn from all three Synoptic Gospels. The criteria or variables used here are 30 words of fairly high or high frequency in Matthew. These 35 samples are shown with different fonts to identify three genre categories. Each Synoptic Gospel is identified as Mt or Mk or Lk and then narrative samples are in bold

20 M. J. Greenacre, *Theory and Applications of Correspondence Analysis* (London: Academic, 1984).

21 J. F. Burrows, 'Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information', *Literary and Linguistic Computing* 7 (1992) 91–109 esp. 96, 101–2.



Lk = Luke, Mt = Matthew, Mk = Mark. Narratives are in bold, speech samples in italic font. M samples are shown in square surrounds, Q samples from Matthew are shown in a diamond surround for mixed genre, or in oval surrounds for sayings samples.

Dim 1 & 2 = 40.24%
Dim 1-3 = 47.38%
File = f53a.sas
Dat = datn05d5

Figure 1. Plot showing samples from the Synoptic Gospels

face, speech samples in italic font and samples of mixed genre in normal font. Source attribution in Matthew is indicated by showing the 3 M samples in a square surround, Q mixed genre in a diamond, and the 3 samples of Q sayings in Matthew in oval surrounds.

In CA the first major contrast in the data is shown on the main axis, normally the X or horizontal axis. Here it is very obvious that narrative samples are those in bold face on the left, and speech samples are those in italic font on the right. Samples of mixed genre are more central, range from −0.25 to +0.2 on this axis and overlap the others only slightly. Genre is evidently a strong factor in the style of the Synoptics. Other output (not printed here) shows that narratives use words such as καί and δέ (and, but) and genitive endings. The speech samples are characterized by words such as εἰ, γάρ, ἀλλά, μή, οὖν (if, for, but, not, so then).

The second contrast on the Y (vertical) axis captures the next most important feature of these texts. The lower third of the plot contains 8 of the 10 samples from the Gospel of Mark, whereas all three samples from M are in the top section of the

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

plot. The third dimension (not shown) captures the differences between Matthew and Luke. CA therefore picks out the most prominent differences one might expect to find: narrative v. speech, Mark v. M, Luke v. Matthew.

4.1. *Comparing Narrative Samples*

At first a sample at the very top included a short section containing the genealogy. This was exchanged for a similar sample but with other normal M narrative replacing the genealogy. Even so this sample is still higher than all the other narrative samples and so clearly distinct from Mark, and to some extent from other Synoptic narrative also.

4.2. *Comparing Samples of Speech (Sayings)*

As the Q material consists mainly of sayings the right-hand side of this plot is of crucial importance. Here it is clear that the large sample of M sayings is higher than all the other sayings samples, and separated from even the nearest of them by a very definite clear white space. This does seem to suggest a probable difference between M and Q sayings in Matthew. It should also be noted that the sample of mixed genre from M is higher than that of mixed genre from Q.

To conclude this section: when each genre is examined separately, each reveals a difference between M material and the rest. The lack of sufficient narrative in the Q material allows little further progress in examining the narratives. In the sayings and mixed samples M is distinct. The distinctiveness of the M material is compatible with the 2ST, and less favourable to the FGT, but further analysis is evidently desirable as it would be better for the data to be partitioned further.

5. Using High Frequency Words in 60 Smaller Samples

As the material had to be divided both by genre and by source it was quickly apparent that there were insufficient samples if the sample size is set at 1000 words. The amount of text cannot be increased, so each sample from Matthew was divided into four samples of 250 words, and some additional sections, previously too small, could now fit. A consequence of this strategy is that only the most frequent words in the text will now be used as variables. At the same time a change of method will permit samples of more than one genre to be tested simultaneously, as there is a method which allows for genre, before testing whether an overall source effect is unlikely to be due to random variation.

This also allows the statistical significance of the results to be assessed, something which is not the case with the simple numerical counts based on vocabulary in many commentaries on Synoptic texts. There are studies which do use statistics which include, or are amenable to, significance tests. Works by Lloyd Gaston[22]

22 L. Gaston, *Horae Synopticae Electronicae: Word Statistics of the Synoptic Gospels* (SBLSBS 3; Missoula: Scholars, 1973).

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

and Raymond A. Martin[23] and a number of studies of patterns of verbal agreements in the Synoptics cited by John Poirier[24] provide examples of this.

Matthew yields 60 samples of 250 words (when one containing the genealogy is omitted). The samples are classified by genre and by proposed source. The key word variables are now the 5 most frequent words in Matthew. These, with the totals for each in Matthew, are: the definite article (art) (2790), καί (1178), αὐτός (aut) (922), δέ (494) and ἐν (293).[25] Where subsidiary tests use only four criteria it is ἐν which is omitted. There is still insufficient narrative for tests on Q narrative material, but Markan and Q and M material can be tested on the sayings and mixed genre groups. The latter include parables and apophthegms. In order to eliminate possible bias from using variable selection methods, these were avoided. The 5 most frequent words were used as standard for all the main tests, though some subsidiary tests were used with the 4 most frequent words and/or the removal of samples suspected of being problematic. These subsidiary tests were strictly limited to the exploration of supplementary hypotheses, and only the standard tests were used for the main conclusions.

6. Using GLM to Test for Source Effect while Allowing for Genre

The method used in the next section is part of the SAS statistical suite and is known by the abbreviation GLM. It is their version of a more widely used method known as the General Linear Model. In the work reported here GLM was run on the variation in the use of the 5 most frequent words in the 250 word samples from Matthew. The null hypothesis to be tested is that, when allowance has been made for a genre effect, there is no overall source effect. If the resulting p value is very low, then the conclusion is that the difference indicates that a variation in the word counts correlates with a change of source, and is unlikely to be due to random variation. Normally a p value of 0.05 (5%) or less is regarded as significant, and one of 0.01 (1%) or less as very significant. (More loosely researchers tend to say that a low p value indicates that a given effect is significant, but that should be interpreted in the manner described above.)

The results from GLM normally indicated a significant genre effect, but also a significant source effect, even when allowing for the genre effect. There are two types of potential error here. Accepting a value of 5% may be too lenient, in which case the researcher might infer a significant source effect when that is not the case. Alternatively, however, insisting that only a p value of 1% or less is to be regarded as significant could lead to the opposite error, and cause the

23 R. A. Martin, *Syntax Criticism of the Synoptic Gospels* (SBEC 10; Lewiston: Mellen, 1987).

24 Poirier, *Statistical Studies*, 77–8.

25 F. Neirynck and F. van Segbroeck, *New Testament Vocabulary: A Companion Volume to the Concordance* (BETL 65; Leuven: Peeters, 1984) 290, 119, 221, 229, 246.

researcher to reject evidence of a significant source effect, when it does indeed exist. One cannot reduce the risk of one of these errors without increasing the risk of the other. That is the more serious calculation which lies behind the looser explanation that a 5% p value is significant and a 1% p value very significant. The decision that an effect is not random does not exclude an objector appealing to some further factor. Such an appeal would, however, need to present an auxiliary hypothesis, and itself need to be tested.

The tests carried out using GLM were on samples divided into 3 genre categories (narrative, sayings and mixed) and allow for 3 possible sources (Mark, Q and M). In some tests only some of these categories were available, or selected even if available. GLM can be asked to divide the material by source and genre, calculate the relevant means, analyse the variance, give a p value to the hypothesis that there is no overall genre effect and then do the same for the hypothesis of no overall source effect. It will also, on request, plot the actual and predicted (mean) values for each sample in each group.

For tests on Matthew GLM was run by a SAS statement providing a model of the following kind:

model:art kai aut de en = source genre source*genre;

The model is designed to test whether the frequencies of the items listed on the left vary in response to the effect of source, genre and the interaction between source and genre. In other words the model is testing the view that a change of genre in a text results in a change in the use of the most frequent words in the text. It is also testing the view that, when allowance is made for genre, a change of source can also result in a change of use in these very high frequency words. Sometimes there can be an interaction between such factors, and that should be considered as well. In most cases, in this study, the possibility of interaction can be checked and then excluded as not significant.

What this procedure does is to formalize what literary scholarship often does intuitively. The careful reader of a text might well expect to find that style changes, and that the use of particular words changes when there is a change of genre in a long text, or when there is thought to be a change of source. Technically the terms on the left are potential response variables. The model sets up a test of the hypothesis that the items on the left do *not* vary in response to changes of genre or source as specified on the right. This null hypothesis is rejected if a sufficiently low p value results, as discussed above.

The tests do not examine every possible solution to the Synoptic problem, but have the more limited aim of testing the null hypothesis that there is no overall source difference between Q and M material in Matthew. If the Q theory ought to be rejected, and the view that Luke used Matthew ought to be accepted, then there should not be significant differences in Matthew between

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

Q passages and M passages. If the null hypothesis stands, then that favours the FGT, if it falls, that favours the 2ST. Faced with a complex set of problems it is sometimes advisable to aim at a more modest 'crucial experiment' to test between two rival views, rather than to try to find a method to solve everything at once. (It is not unreasonable to hope that others might consider adapting the methods to resolve other aspects of the overall problem.)

Some of these tests do also include comparisons of Matthean material held to derive from Mark. By comparing Q and M with such supposed Markan material it is possible to put the comparison of Q and M in a slightly wider context. It also helps to establish that the method used is also identifying other differences which are to be expected on the theory under test.

6.1. *Using GLM on Markan, Q, and M Material*

A series of tests were run and these are summarized in Table 1 below. Each test is first given a brief account and then the table of results follows the last of these accounts. The first six tests all use the standard variables consisting of the 5 most frequent words in Matthew. The remaining tests examine a number of minor variations in order to look briefly at various alternative ways of viewing parts of the evidence.

The first test includes text attributed to Markan, Q and M material. Narrative cannot be tested as there is insufficient pure narrative in the Q material. There are 22 samples of mixed genre, and 22 samples of sayings. Allowing for genre the p value for no overall source effect is 2.73% and so significant at below 5%. If only the saying samples are tested, the p value is 0.69% and so very significant at below 1%. It is reasonable to infer that there is a difference between the sources, and that it is even more evident in the sayings material.

6.2. *Using GLM on Q against M Material*

Three-way comparisons are more complex than binary ones, and the crucial question at issue is whether there are differences of style between Q

Table 1. The Main tests with GLM

Sources	Genres	Samples	Var.	p value	Verdict
Mk Q M	m & s	44	5	**2.73%**	**significant**
Mk Q M	s only	22	5	**0.69%**	**v. significant**
Q vs. M	m & s	27	5	**2.91%**	**significant**
Q vs. M	m & s	13 (500 words)	5	**0.15%**	**v. significant**
Mk vs.Q	m & s	33	5	**4.40%**	**significant**
Mk vs.M	n & s	26	5	**4.54%**	**significant**

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

material and M material. The attribution to Q and M is, of course, questioned, but the aim here is to establish whether it can or cannot be disconfirmed. The third test therefore omits the material parallel to Mark, and compares the Q material with that found only in Matthew. There are still two genres involved (sayings and mixed genre), and there are now 27 samples: the overall p value is 2.91%, and so significant, as it is below the 5% threshold.

Some variations were run on this result, as samples of 250 words are small, though viable when words of the highest frequency are used.[26] It is, however, worth giving closer attention to the result of combining pairs of the 250 word samples so as to make up 13 samples of 500 words each. When the fourth test ran GLM on these 500 word samples, the p value relating to source effect was very significant at 0.15%, well below even the more stringent 1% level.

The results so far have all indicated significant p values pointing to rejection of the hypothesis that Q and M material is similar in style. The first and third tests gave a p value under 3%. The difference between the groups of samples was more evident in the second test on sayings samples only, with a p value below 1%. In the fourth test the difference between 500 word samples of mixed and sayings genre was well below 1%. These results range from significant to very significant.

To put the previous result in context, the fifth test now compared Mark and Q using 33 samples of mixed and sayings genres. As before, using the 5 most frequent words and allowing for genre, the evidence against there being no source effect was significant. The p value was 4.4%, under the 5% threshold, but less clear than the test of differences between Q and M material. It is mildly

26 Some of the variations were minor and considered the following objection. Someone might assert that samples of 250 words should only be used with the 4 most frequent rather than the 5 most frequent words. To counter this objection, two further tests were made using only the 4 most frequent words. When mixed and sayings samples from Matthew attributed to Markan, Q and M material were used the p value came in at 1.34% and so it was more, not less, significant than the original 2.73%. In a straight comparison of Q and M material, allowing for genre differences between mixed and sayings samples, and using the 4 most frequent words, the p value relating to source effect came in at 1.85%, again more significant than the original 2.91%. The main results should therefore stand, and any objector note that to press the objection would actually strengthen, not weaken, the case made here. Using only the 4 most frequent words would also produce a slightly more significant p value for the source effect between Markan and Q material, but not, however, for the comparison of Markan and M material.

Some further tests were made omitting the sample which contains the parable of the sheep and the goats (sample s56). The reason for this is that the style of this parable is very distinctive and different from the other parables. In the comparison of Markan, Q and M material this made the p value for source effect more significant, both when 5- and 4-word variables were used. Similar results occurred with the comparison of Q against M material. This does not mean that these more significant p values should be pressed into becoming main results. It may, however, tend to confirm the suspicion that there is something unusual about the style of this parable.

surprising to note that the stylometric evidence is more definite about the difference between Q and M material than that between Markan and Q material.[27]

To complete the main tests, Markan and M samples were compared using narrative and sayings genres. Again the result was significant and below the 5% level, but by a slightly smaller margin again, coming in at 4.54%. There is evidence that Matthew made fewer editorial changes to sayings material,[28] and if that is correct it might explain why source differences are more evident where sayings material is more prominent (i.e. in tests involving Q material) than in those where narratives are involved.

The lower the p value is the more significant the result becomes. The main tests involving Q and M used the 5 most frequent words, though some subsidiary tests used only the 4 most frequent words, and/or omitted one of the samples. The argument rests on the main tests, but the results of the subsidiary tests (shown in Tabel 2) could hardly be said to detract from those. The results of the main tests using GLM are all significant at the p value of 5% or less. There is some indication that differences between Q and M are more marked in the sayings genre, and this can be pursued using Discriminant Analysis.

7. Using Discriminant Analysis on Single Genre Comparisons

Discriminant Analysis is a further resource which can be deployed where only one genre is involved. SAS provides two forms of this method. Discrim

Table 2. Subsidiary Variations on the Main Tests

Mk Q M	m & s		4	(1.34%)	significant
Q vs. M	m & s		4	(1.85%)	significant
Mk vs. Q	m & s		4	(2.21%)	significant
Mk Q M	m & s	om. M56	5	(0.42%)	v. significant
Mk Q M	m & s	om. M56	4	(0.15%)	v. significant
Q vs. M	m & s	om. M56	5	(0.26%)	v. significant
Q vs. M	m & s	om. M56	4	(0.09%)	v. significant

27 The difference between a 2.73% and a 4.4% likelihood that the result is due to chance is not massive. Any surprise arises from a higher expectation for Marcan priority due to the additional possibility of making comparisons with Mark itself, whereas discussion of Q passages is largely restricted to comparison of Matthew with Luke. The tests reported here are almost all based just on samples from Matthew attributed to Q, M and Mark, in order to focus attention mainly on one crucial aspect of the problem.

28 Poirier, *Statistical Studies*, 77–78.

supplies not only details of p values, but also Mahalanobis distances[29] between groups, and error rates where cross-validation assigns a sample to a group other than that expected. In this paper the words attribute and attribution are normally used to indicate the allocation of texts to sources by conventional literary scholarship. The terms assign and assignment are used to denote the outcome of cross-validation tests that test each sample in turn against all the others in two (or more) groups.

Candisc gives some of the same figures, but will also plot the resulting locations of the samples and so the groups. This allows a visual check on the extent to which the samples do, or do not, form groups whose separation is clear and distinct. The 5 most frequent words were once again used as variables, and the variable selection method (Stepdisc) was not used as this can introduce bias. The aim of the next series of tests is to look at the grouping of sayings attributed to Q material and sayings attributed to M material. The hypothesis to be tested is that these groups are not distinct, and that samples cannot be correctly assigned by this method to the group to which they are supposed to belong. This hypothesis should be rejected if the p values are low, if the Mahalanobis distances between groups are significant and if there are few classification errors. Again rejection of the null hypothesis would favour the 2ST over the FGT; acceptance of it would favour the FGT.

The first test was to see whether there was clear discrimination between the sayings samples from the three proposed sources in Matthew: Mark, Q and M. These provide 22 samples of sayings in all. There are four different ways of calculating the multivariate p value which results. The simplest way to report this is that using the 5 variables on three groups p was very significant at 0.69% or less. (The lowest value was 0.07%.) The Mahalanobis distances measure the separation of the three groups and the p values for these were between 1.11% and 1% in all three cases (Mk/M, Mk/Q, Q/M), signalling a significant distance between all three groups. The resulting plot (see Figure 2) sent M samples to the upper left quadrant, Q samples mainly to the lower left quadrant, and Markan samples to the right quadrants. (The plot is extended to the right.) Most of the samples are in three separate groups, but with three samples overlapping in the centre of the plot.

The error rate from cross-validation was 5/22 (22.7%) with 77.3% assigned to the expected group. Each sample in turn was stripped of its source identifier, and classified into whichever of the three groups it matched on the basis of the

29 This is an adjusted measure for calculating the distance between two clusters of data, or between outlying samples and a cluster. Because clusters may well not be perfect spheres, but bulge in one direction or another, a method needs to be used which calibrates the distance from the centre of the cluster allowing for the presence or absence of a bulge in the direction in question: see Figure 5 at the end of the Appendix.

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

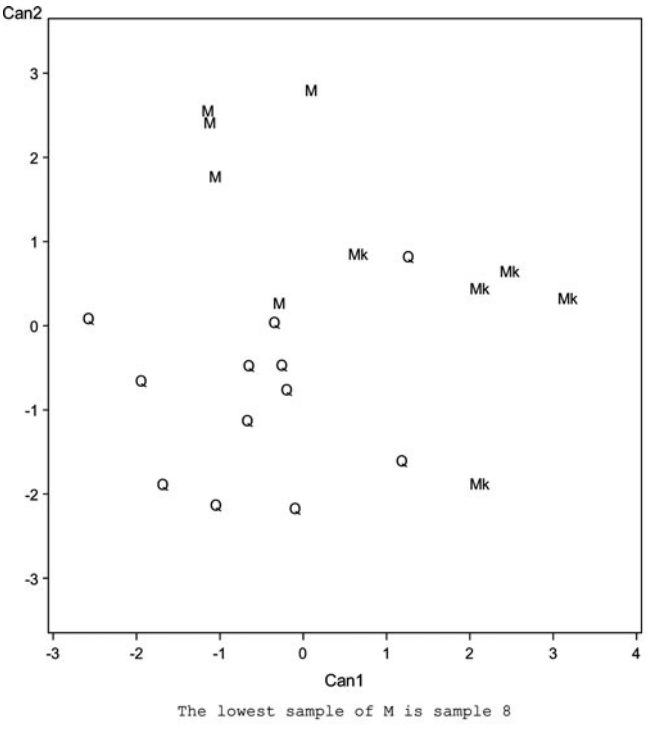


Figure 2. Plot showing Candisc on sayings samples from Matthew (Mk, Q, M)

statistics. When each sample had been separately assessed against all the others, the number of incorrect assignments was tallied. Success rates of around 73% for function words and 79% for strings were regarded as quite normal in stylometric studies of other literatures performed by Forsyth and Holmes.[30] Results for word frequencies by Grieve[31] in 2007 reached 80–90%, but when testing only two authors with a large function word profile. It requires very voluminous texts, such as lengthy novels,[32] to obtain success rates above 90%.

30 Forsyth and Holmes, 'Feature Finding', 169.

31 J. Grieve, 'Quantitative Author Attribution: An Evaluation of Techniques' *Literary and Linguistic Computing* 22.3 (2007) 251–70, esp. 261.

32 D. L. Hoover, 'Multivariate Analysis and the Study of Style Variation', *Literary and Linguistic Computing* 18 (2003) 341–60, esp. 343.

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

The overall multivariate p value of 0.69% (or less) did produce an outcome in the very significant category, though the error rates suggest slightly more caution. These results indicate that there probably is a stylistic difference between the three groups of samples attributed to the three different proposed sources. It is not possible, however, to assign every single sample to its expected group; this can be done only with about 77% of them.

Three-way contrasts are always more problematic than binary comparisons. So the next test focused on the relation between samples attributed to Q and those attributed to M. Whether or not there is a difference here is central to the purpose of this investigation. There were 12 samples of Q sayings and 5 of M sayings, making 17 in all. The 5 most frequent words were again used as variables. This time the p value was 1.58% and so well below 5% and clearly significant. Cross-validation gave 3 errors out of 17 samples. With 14 samples out of 17 correctly assigned the success rate was 82.3%. Candisc produced the plot which follows (see Figure 3).

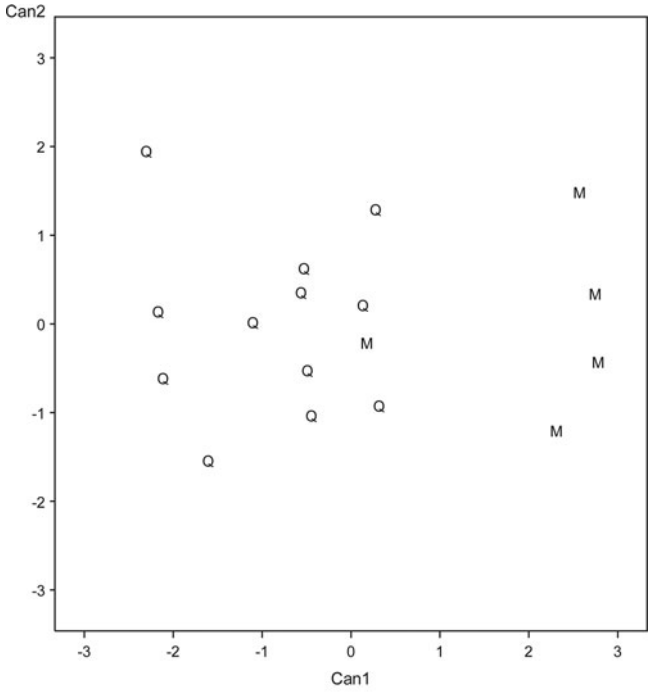


Figure 3. Candisc on Sayings samples from Q and M material in Matthew

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

The plot shows 12 samples of Q material towards the left, four samples of M material to the right, but one M sample nearer the centre of the plot on the right-hand edge of the Q group. The three errors involve three samples in the centre of the plot (S, R and 8). Cross-validation assigned S and R to M, and sample 8 to Q. This clarifies the issue in that it can be seen that, in the main, the two groups are distinct,[33] but there is a small area of overlap between them. The method can correctly distinguish Q sayings samples from M sayings samples in just over 4 instances out of 5.

Some additional tests were carried out in order to explore just which samples were most prone to being assigned to a group other than that to which they are attributed by the 2ST. These were not intended to amend the findings of the main tests, so they are omitted at this point but reported in note 34 below. As the evidence of the main tests is sufficient on balance to reject the null hypothesis of no significant difference in Matthew between Q material and M material, then that favours the 2ST as against the FGT. This should make it clear that the purpose of these additional tests is not to make any special pleas in relation to the main argument, but rather to explore, elsewhere, what factors might be at issue in those samples which differ from the main trend of the results, and whether fluctuations observed in the main results might match one or another of the variations adopted by proponents of that theory.

Table 3 provides an overview of the main tests involving Discriminant Analysis.

Table 3. Results from Discriminant Analysis

Sources	Genre	Samples	Variables	p value	Error free	p (AxBxC)
Mk Q M	sayings	22	5	**<0.69%**	**77.3%**	<1.1%
Q M	sayings	17	5	**1.58%**	**82.3%**	-
Q M	sayings	17	4	(0.62%)	(88.23%)	-

The second result should be regarded as the main one and the supplementary explorations as interesting additional information which is not to be given the same weight. The main result was a p value of 1.58% for Q versus M sayings, three errors out of 17 and a discrimination rate of 82.3%.

33 Again any objector insisting on using only the 4 most frequent variables would be confronted with p comfortably under the 1% significance level at 0.62%, and only two samples of 17 cross-classified (sample R to M and 8 to Q), giving 88.23% correctly assigned and an even stronger conclusion.

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

There were, however, two further tests. These paired up 250 word samples to form samples of 500 words. Sample 9 had no partner so was first doubled and included, then omitted. In the first case 8 of 9 samples were assigned to the expected source (88.9%), in the second 7 out of 8 were assigned to the expected source (87.5%). In each case sample N + O was wrongly assigned.

8. Using Cluster Analysis

Another method which can be used here is Cluster Analysis. This method is not given prior knowledge of the groups to which samples are attributed. It simply joins the two samples closest to each other into a cluster, then joins the next two nearest items together, and so forth until eventually all the samples are in a single cluster. It is then possible to see if the resulting pattern matches expected groupings in the data, or reveals unexpected patterns deserving further exploration. Earlier tests reported in this paper had first used samples of 1000 words, then much shorter ones of 250 words, and briefly a set which combined pairs of 250 word samples into samples of 500 words. These latter provide a suitable set for Cluster Analysis, with one drawback that sample 9 had no partner and so was simply doubled. This solution might accentuate any wayward tendencies in sample 9, so as a safeguard Cluster Analysis was also run without this sample.

When the 9 sayings samples of 500 words were run with Cluster Analysis, using counts of the 5 most frequent words, Ward linkage and Euclidean distance, the 6 samples from Q clustered together and the 3 samples from M clustered together. The resulting chart shows the Q samples on the left and the M samples on the right. It is worth noting once again that Cluster is not told whether samples are attributed to Q or to M—it links samples by assessing the similarity of their usage of the 5 most frequent words.

The plot (see Figure 4) is in the form of a tree, or dendrogram, showing clustering starting near the left edge of the plot (where LM joins PQ). More pairs of Q samples join this cluster until all 6 are there. Meanwhile the M samples 56 and 78 join up, eventually followed by sample 99 (the doubled 9). The end result is remarkably close to the expectations of the 2ST. (The dendrogram of the run without sample 9 is very similar, apart from the absence of that sample.) This produces an initially puzzling result, that Cluster Analysis should not reveal variations which Discriminant Analysis identified. It is probably correct to regard cross-validation as more rigorous, but it is also worth noting that even this did give a higher assignment of samples to expected source when the sample size was 500 words. It is also worth noting that sample N + O was late joining the Q cluster here, though it did do so. When Cluster Analysis was run on the 250 word samples it assigned 14/17 to the expected source (82.3%), putting 6 & 8 with Q, and S with M.

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

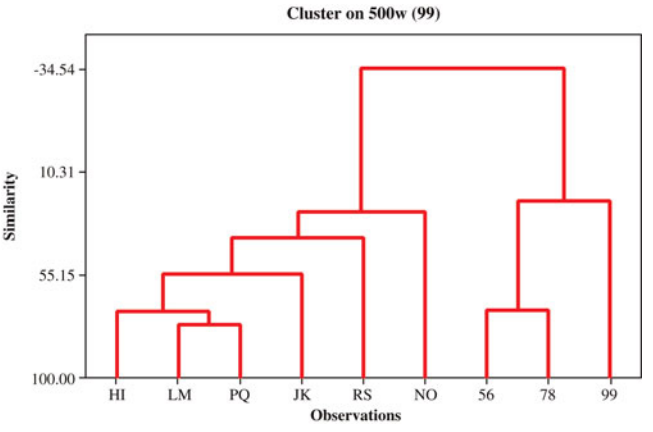


Figure 4. Six Q samples cluster on the left and three M samples cluster on the right

The overall result from Cluster Analysis supports the main conclusion that these tests can assign about 80% of the sayings samples attributed to Q or to M to the expected source. When the sample size is increased to 500 words this figure rises, but we are using fewer samples. GLM, Discriminant Analysis and Cluster Analysis allow a cautious judgment that Q material and M material are not indistinguishable, even if it is not possible to effect a clear separation of every last sample in every test.[34]

34 At this point it is worth considering the implications of the results of this study for the delimitation of Q. If the main results above are accepted, it is still possible that some aspects of the results would suggest that minor modifications should be made to the 2ST. Some of the subsidiary tests explored which samples were more often assigned to a source other than the one posited by the 2ST. Samples R, S and 8 strayed most often, but if either R and 8 were omitted, or given a changed attribution, then S did not stray. This suggests that the main problem lies with the attribution of R to Q and 8 to M, rather than with S.

Sample R contains a set of verses from Matt 23 (the woes against the scribes), and the fact that cross-validation assigns this sample to M does, in fact, match the view of some conventional literary scholars that these woes stem at least in part from M. Scholars have long-since noted low levels of agreement in the double tradition here, indications of the use of another source (so Burkett, *Rethinking*, 2.163–5, 236) and some evidence of divergent translation from Aramaic (Casey, *Q*, 82). Matthew is known to practice conflation, so it would not be unreasonable to suspect that at least some of these verses do not derive from Q (or not from Q alone). Sample 8 contains Matt 10.41; 11.14-15, 28-30; 12.36-37; 16.17-19; 18.10, 16-17. Some of these 7 short passages are preceded or followed by verses from the double tradition which are attributed to Q. In two cases they are both preceded and followed by such verses. The context can be summarized as Q Context: a) no, b) both sides, c) before, d) both sides, e) no, f) after, g) before. Long ago Schürmann argued on stylistic grounds that a range of passages found only in Luke, but in a Q context, should be attributed to Q, and made the equivalent proposal for a few passages in Matthew. On this see H. Schürmann, 'Sprachliche Reminiszenzen an abgeänderte

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

Stylometric work can find clear and distinct differences between major known authors of lengthy texts, but when dealing with texts much shorter than modern novels it is unreasonable to expect results with accuracy close to 100%. Forsyth & Holmes[35] give a figure close to 73% for success rates using function words on small samples in this kind of work. In the case of the proposed Synoptic sources the difficulty of achieving clear discrimination is greater. If there are sources it is likely that they have been adapted to a greater or lesser extent by the authors who included them. Where a proposed source still exists, or where there are two adaptations of a proposed source, estimates of the extent of that adaptation can be attempted. The aim of the current work is slightly different from that. Here the aim was to investigate the theory that if samples of text are taken from two proposed sources, and allowance made for genre there is still sufficient stylometric evidence to reject the view that there is no significant difference between them.

9. Conclusions

Genre effects are in almost every case more significant than source effects in these samples. Correspondence Analysis on large samples did seem to show some source effect even when genre was taken into account. M samples were distinct from those attributed to Mark or Q, and the latter from each other. This visual impression from plots of large samples was strengthened when GLM was used on smaller samples. Here the p values for the source effect, after allowing for genre, showed a distinction between Mark, Q and M samples of sayings and of mixed genre ($p = 2.73\%$). Omitting the Markan samples showed that this source effect still obtained when only Q and M were used (with the p value significant at 2.91%).

It is reasonable to conclude from this that there probably is a fairly consistent difference between Q and M passages. That difference seemed to be slightly stronger in the sayings. These were examined further with Discriminant Analysis. Here the main test of the three-way comparison gave a multivariate p value of 0.69% or less, and so below 1%, and very significant. It also showed 77.3% of samples correctly assigned to the groups to which they were attributed. The main comparison of Q and M sayings gave a p value of 1.58%, and this test correctly assigned 82.3% of the samples to the groups to which they were attributed.

---

oder ausgelassene Bestandteile der Spruchsammlung im Lukas- und Matthäusevangelium', *NTS* 6 (1960) 193–210 (reprinted in *Traditionsgeschichtliche Untersuchungen zu den synoptischen Evangelien* [Düsselforf: Patmos, 1968]). Though J. S. Kloppenborg, *The Formation of Q: Trajectories in Ancient Wisdom Collections* (SAC; Philadelphia: Fortress, 1987) 83 and n. 147 is more cautious about this view, it might be worth reconsidering the attribution of some of these verses.

35 Forsyth and Holmes, *Feature Finding*, 169.

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

The case must rest on the main results rather than on exploratory variations and permutations. These main results give a clear indication of a probability value below 5%, and in some cases below 3% for the differences of style being due to random error. These results are the ones given in bold in the tables. When discriminant analysis is used with cross-validation it does not assign every single sample to the source to which is attributed, but it does do this in around 80% of the cases. Note 34 explored some further implications of these results relating to the extent of Q. These results permit a cautious preference for the 2ST, against its main rival FGT, which rejects Q. They do not solve every aspect of the Synoptic problem, but they do give an indication as to which of the more strongly favoured theories has the support of the stylometric evidence offered here.

Appendix

List A shows 61 smaller samples in all from Matthew, but the genealogy was normally omitted, leaving 60 samples from Matthew for normal tests. Of those 60 samples 27 were more heavily used in the main tests, and these are those attributed to Q and to M, and were of mixed genre or of sayings. These samples are clearly indicated with a bold character in the second column.

The samples were obtained by dividing the text first by proposed source, then by genre. Apophthegms and parables were grouped into a 'mixed genre' set of samples to produce three main groupings: narratives, mixed genre and sayings. Quotations were omitted from the samples.

In the genre column n = narratives, s = sayings and m = mixed. In the reference column the numbers in brackets indicate words within verses.

sample	**name**	**genre**	**source**	**Matthew (samples of 250 words)**
s1	a	m	Mk	9.9-17; 12.1-4(2)
s2	b	m	Mk	12.4(rest), 8, 46-50; 13.54-58; 15.1-2
s3	c	m	Mk	15.3-11, 15-20; 18.1-5; 19.1(1-3)
s4	d	m	Mk	19.1(4)-9, 13-22(4)
s5	e	m	Mk	19.22; 20.20-28; 21.18-19, 23-24(2)
s6	f	m	Mk	21.24(3)-27; 22.15-26
s7	g	m	Mk	22.27-46; 24.1-2; 26.6
s8	h	n	Mk	8.1-4, 14-16, 23-31(9)
s9	i	n	Mk	8.32-34; 9.18-29(10)
s10	j	n	Mk	14.13-21; 15.29-32(30)
s11	k	m	Mk	9.1-15(15)

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

s12	l	m	Mk	9.15(16)-17; 12.9-10, 13-16; 14.22-28(6)
s13	m	m	Mk	14.28(7)-36; 15.26-28; 20.29-33(4)
s14	n	n	Mk	3.1-6, 13, 16-17; 4.18-23(4)
s15	o	n	Mk	4.23(5)-25; 9.35-6; 10.1-4; 15.21; 21.17; 26.1-3(4)
s16	p	n	Mk	26.3(5)-4, 13-26(17)
s17	q	n	Mk	26.26(18)-41(2)
s18	r	n	Mk	26.41(3)-51, 55-57(15)
s19	s	n	Mk	26.57(16)-74(2)
s20	t	n	Mk	26.74(3)-75; 27.1-2, 11-18, 20-23, 26-27(9)
s21	u	n	Mk	27.27(10)-45(6)
s22	v	n	Mk	27.45(7)-51, 54-61; 28.1, 5(1-3)
s23	w	m	Mk	13.1-9; 21.33-41(6)
s24	x	m	Mk	13.10-13, 18-23, 34; 21.40-42(9), 45(1)
s25	y	s	Mk	4.12-17; 10.17-18, 21-22, 42; 12.31; 16.1-6
s26	z	s	Mk	16.7-12, 25-28; 17.9-12(22)
s27	A	s	Mk	17.12(23)-13, 22-23; 18.6-9; 19.23-27, 29(1-2)
s28	B	s	Mk	19.29(3)-30; 21.20-22; 24.3-9(13)
s29	C	s	Mk	24.19-25, 29-36, 42(1-5)
s30	**D**	m	Q	8.19-22; 11.2-6; 12.22-28(11)
s31	**E**	m	Q	4.1-11; 8.5-11(15)
s32	**F**	m	Q	13.33; 18.12-13; 22.1-10; 25.14-15(4)
s33	**G**	m	Q	25.16-30(1)
s34	**H**	s	Q	3.7-12; 5.3, 4, 6, 11-13, 15, 18(7)
s35	**I**	s	Q	5.25-26, 32(4-end), 39(5)-42, 44-48; 6.9-13(1)

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

s36	**J**	s	Q	6.13(2-end), 19-30(14)
s37	**K**	s	Q	6.30(15)-33; 7.1-5, 7-13(18)
s38	**L**	s	Q	7.14, 16-27; 9.37(1-8)
s39	**M**	s	Q	9.38; 10.7-16, 19-20, 24-26(2)
s40	**N**	s	Q	10.26(3)-40; 11.7-8(6)
s41	**O**	s	Q	11.9-13, 16-19, 21-23, 25-27(14)
s42	**P**	s	Q	12.32-35, 38-45(11)
s43	**Q**	s	Q	12.45(12-end); 13.16-17; 17.20; 18.15, 21-22; 19.28; 23.35(15)-39(4)
s44	**R**	s	Q	23.4, 6, 7, 12-13, 23, 25-27, 29, 32, 34-35(14)
s45	**S**	s	Q	24.26-28, 37-41, 43-50(13)
s46	**T**	m	M	15.12-13; 17.24-27; 19.10-12; 21.14-16; 15.22(1-3)
s47	U*	n	M	1.1-17(4) (*genealogy, normally excluded)
s48	V	n	M	1.17(5)-25; 2.1-6(2)
s49	W	n	M	2.6(6), 7-17, 18(8), 19-20(1)
s50	X	n	M	27.3-9, 19, 24-25, 52-53, 62-64(29)
s51	Y	n	M	27.65-66; 28.2-4, 9-20(10)
s52	**Z**	m	M	13.24-30, 44-50(3)
s53	**1**	m	M	13.51-52; 18.23-35; 20.1(1-4)
s54	**2**	m	M	20.1(6)-16; 21.28(1-12)
s55	**3**	m	M	21.28(13)-32; 22.11-14; 25.1-8(13)
s56	**4**	m	M	25.31-44(end)
s57	**5**	s	M	5.7-10, 16-17, 19-24, 27-29(1)
s58	**6**	s	M	5.29(2)-39(4), 43; 6.1-4(16)
s59	**7**	s	M	6.5-8, 16-18, 34; 7.6, 15; 10.5-6, 23 (1-2)
s60	**8**	s	M	10.41; 11.14-15, 28-30; 12.36-37; 16.17-19; 18.10, 16-17(18)
s61	**9**	s	M	23.1-3, 5, 8-11, 15-22(12)

Two partial overlaps in Markan sections of Matthew (9 verses in sample a with k & l, and 18 words in w with x) might have a slight effect on tests on Markan material, but would not affect tests on Q and M.

List B Twelve larger samples of 1000 words from Matthew were compiled by joining up 48 of the smaller 250 word samples. Column 5 shows the identity of the samples using the symbols from column 2 in List A. (The unused samples were e f g h i j m v C T 4 9 and U which was replaced with Y.) The references for the component samples are given in List A.

sample	**name**	**genre**	**source**	**components**	**Matthew**
M1	l	m	Mk	a b c d	apophthegms
M2	m	m	Mk	k l w x	mixed genre & parables
M3	n	n	Mk	n o p q	narratives
M4	o	n	Mk	r s t u	narratives
M5	p	s	Mk	y z A B	sayings
M6	q	m	Q	D E F G	apophthegms & parables
M7	r	s	Q	H I J K	sayings
M8	s	s	Q	L M N O	sayings
M9	t	s	Q	P Q R S	sayings
M10	y	n	M	[U]V W X Y (Y replaced U)	narratives
M11	v	m	M	Z 1 2 3	parables
M12	w	s	M	5 6 7 8	sayings

List C gives details of 1000 word samples from Mark

In the reference column the numbers in brackets indicate words within verses unless otherwise indicated.

sample	**name**	**genre**	**Mark**
m1n	a	n	1.16-20, 35-39; 6.14-29; 11.15-17(1-8 & 18-21); 14.10-40(10)
m2n	b	n	14.41-62, 63-72(19); 15.1-39(10) (not v. 28)
m3nh	c	n	1.21-31, 41-45; 4.35-41; 5.1-7(20), 8-39(5)

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

m4nh	d	n	5.40-43; 6.30-52; 7.32-35(3), 37; 8.1-10, 22-26; 9.14-29(2).
m5ap	e	m	2.13-28; 3.20-29(10), 30-35; 6.1-6; 7.1-20(3) (not v. 16)
m6ap	f	m	9.33-40; 10.2-9, 13-22, 35-39; 11.12-14(13), 27-33; 12.13-34, 41-43(14).
m7mp	g	m	4.1-15(3), 20, 26-32; 12.1-10; 2.1-12; 3.1-6; 7.24-30; 10.46-50(6)
m8s	h	s	1.14-15; 4.21-25; 6.8-11; 8.11-12, 14-21; 8.34-38; 9.1, 9-11(10), 12-13, 30-32, 41-50 (not vv. 44,46); 10.10-12, 23-33(4).
m9s	i	s	13.6-36(1); 10.34; 10.42-45; 11.20-25; 12.38-40; 13.1-16(2) (some overlap)
mxa	k	n	1.4-13; 9.2-8; 11.8-11; 16.1-8 (doubled)

There are overlaps in samples i and k from Mark, used in other tests on the consistency of the style of Mark not at issue here. (These related to Mark 13, and to special narratives in Mark.)

List D gives details of 1000 word samples from Luke

sample	**name**	**genre**	**source**	**Luke**
L1	x	s	Mk	5.36-39; 8.9-18; 9.3-5, 23-27; 18.25-33; 21.5-33(10)
L2	y	m	Mk	21.33(11)-34; 22.15-23; 4.42-44; 5.27-35; 6.1-5; 8.19-21; 9.18-22, 28–36, 43(16)-50; 18.15-19(7)
L3	z	m	Mk	18.19(8)-24; 20.1-8, 19-35; 21.1-4; 22.24-34; 23.27-31; 8.4-8; 20.9-18; 4.31-33(8).
L4	A	n	Mk	4.33(9)-41; 5.12-26; 6.6-10; 8.27-47 (3)
L5	B	n	Mk	8.47(4)-56; 9.37-43(15); 18.34-43; 8.22-25; 9.6-17; 19.28-33; 22.7-14, 35-39(11)
L6	C	n	Mk	22.39(12)-66; 23.1-16, 18-26, 32-33, 35-46(4)
Lq1	D	m	Q	7.1-10; 14.15; 9.57-62; 11.1, 14-17, 37-38; 17.5-6; 12.41-46; 14.16-24; 13.18-30; 6.20-22(2)

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press

Lq2	E	s	Q	6.22(3)-49; 7.18-32; 10.1-13(7)
Lq3	F	s	Q	10.13(8)-16, 21-24; 11.2-4, 9-13, 17-26, 29-36, 39-54; 12.1-4(12)
Lq4	G	s	Q	12.4(13)-12, 22-40, 49-59; 13.34-35; 16.16-18; 17.1-4, 22-32
LL1	H	m	L	10.17-20; 11.5-7; 12.15, 47-48; 14.11-14, 25-35; 16.9-15; 17.7-10; 9.51-56; 10.25-28, 38-42; 11.27-28; 12.13-14; 13.1-4(2)
LL2	I	m	L	13.4(3)-5, 16-17, 31-33; 14.1-6; 17.11-21; 10.29-37; 12.16-21; 13.6-9; 14.7-10; 15.1-13(12)
LL3	J	m	L	15.13(13)-32; 16.1-8, 19-31; 18.1-9; 13.10-15(8)

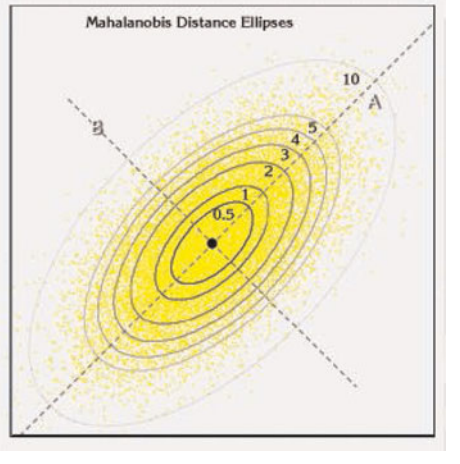


Figure 5. In Euclidian terms B is nearer the centre of the cluster than A, and seems to be so, but the adjusted calculation tallies with the correct statistical judgment that A is more likely to belong to the cluster than B, if account is taken of the shape of the cluster. (The diagram is adapted from one by J. Jenness.)

https://doi.org/10.1017/S002868851100018X Published online by Cambridge University Press