Correspondence Analysis of the New Testament

Harry Erwin, Michael Oakes
University of Sunderland
DCET, DGIC, St. Peter's Campus, St. Peter's Way, Sunderland SR6 0DD, England
E-mail: michael.oakes@sunderland.ac.uk

Abstract

In this paper we describe the multivariate statistical technique of correspondence analysis, and its use in the stylometric analysis of the New Testament. We confirm Mealand's finding that texts from Q are distinct from the remainder of Luke, and find that the first 12 chapters of Acts are more similar to each other than to either Luke or the rest of Acts. We describe initial work in showing that a possible "Signs Gospel", describing Jesus' seven public miracles, is indeed distinct from the remainder of John's Gospel, but that the differences are slight and possibly due to differences in genre.

1. Introduction

Correspondence analysis is a multivariate statistical technique, originally developed by Benzécri (1980). When it is used for the comparison of texts, we start with matrix of whole numbers where the rows correspond to the text samples, and each column corresponds to a countable linguistic feature of that text. In the studies described in this paper, the text samples are 500-word samples of the New Testament in the original Greek, and the columns are word counts for each of the 75 most common words[1] in the Johannine corpus (John's Gospel and Epistles, Revelation). The technique of correspondence analysis takes into account the fact that many linguistic features vary together – texts containing many occurrences of one feature such as the word ημων also contain many occurrences of γαρ and υμων. By considering such similarly distributed groups of linguistic features as one, texts originally described by a large number of features can be plotted on a graph determined by just two factors or groups of co-occurring features. The individual words constituting each factor can also be plotted on the same graph, as shown in Figure 1. At the top of the figure, the text samples from Revelation (labelled "r") score most highly on the second factor (y axis), which s characterised by high occurrence of such words as επι, επτα and γησ. They also have small positive scores on the first factor (x axis), which is most characterised by such words as αυτον and ειπεν. Commands in the R statistical programming language for performing correspondence analysis on texts are given by Baayen (2008:128-136). We used the Westcott and Hort Greek texts[2], stripped out the apparatus and the verse numbers, and then broke each test up into sequential samples of a fixed length of 500 Greek words (except for the final sample from each book, which is shorter). Correspondence analysis cannot be performed if any of the values in the input matrix are 0, so we increased all the frequency counts in the entire matrix by 1.

As seen in Figure 1, our correspondence analysis for the entire New Testament grouped the text samples into four main clusters: Revelation ("r"), the Synoptic Gospels ("Mt", "Mk", "Lk"), the Epistles or letters ("l") and the Gospel of John ("jg" and "s", see section 5). These patterns suggest that the technique is trustworthy, since texts that we would expect to be similar do indeed group together, away from groups of dissimilar texts. Secondly, we see that genre plays a major role in grouping texts. For example, the letters all group together although written by various authors; similarly the Synoptic Gospels group together despite having been written by three different authors. Finally, we see that the texts of Revelation are quite distinct from those of John's Gospel, supporting the general opinion that it has a separate author. Having used correspondence analysis to gain an overview of the main text groupings in the New Testament, in the remainder of this paper we will look at some of the main questions of authorship that have been considered by New Testament scholars. In section 2 we will examine evidence for Q, a proposed source of the Synoptic Gospels. In section 3 we will attempt to answer the question "Did Luke write Acts?". In Section 4 we will discuss the extent of the Pauline corpus, and in Section 5 we will examine whether correspondence analysis throws any light on the question of whether the Gospel of John draws on an earlier source called the "Signs Gospel". We will conclude with some thoughts on how to control for genre effects which can swamp the effects of individual writing styles.

[1] www.mrl.nott.ac.uk/~axc/DReSS_Outputs/LFAS_2008.pdf
[2] http://www-user.uni-bremen.de/%7Ewie/GNT/books.html

2. Q

A widely, but not universally, held view is that the Gospels of Matthew and Luke each draw both upon the Gospel of Mark and a second, probably oral, Greek source. This second original source is referred to as Q, which stands for "quelle" or "source" in German. A detailed recent review of the computational evidence both for and against the existence of Q is given by Poirier (2008). The authors were first alerted to the potential of correspondence analysis to determine the provenance of texts by Mealand's (1995) use of this technique to show that material thought to be from Q is fairly distinct from both those sections from Luke which also appear in Mark (labelled "m"), and those sections which are unique to Luke (labelled "L"). For comparison, he also included samples of Mark's Gospel (labelled "M"). Mealand presents his raw data at the end of his paper, where frequency counts for 25 parts of speech, common words, or other numeric linguistic data, are given for each of the text samples of Luke. The texts are labelled according to whether they constitute infancy narrative, genealogy, are common to Mark, are unique to Luke, or thought to come from Q. Figure 2 shows our own correspondence analysis using Mealand's data for the frequencies of the three Greek words "kai", "nou" and "aut" in each text sample. We were also able to achieve the broad separation of the Q samples from the other texts as described by Mealand. As previously found, one "m" sample was found far from the others, at the extreme left of the diagram. This sample is Luke 21:1-34, Jesus' apocalyptic speech – on its own, due to its being the sole member of that genre in that data set (Linmans, 1998:4). It is also more difficult to discriminate between the material in Luke common to Mark from that unique to Luke. A second source of numeric linguistic data obtained for a correspondence analysis on the "problem of Q" is given by Linmans (1995), where each text has counts for the 23 most common words and 20 parts of speech, and the text samples are classified into one of four genres: narrative, dialogue, aphorisms and parables.

3. Luke and Acts

Traditionally scholars have considered the book of Acts to have been written in its entirety by the author of the Gospel of Luke, since Luke was a companion of Paul, both prefaces are dedicated to "Theophilus", and the preface of Acts refers to a former book by its author. The narrative also follows on smoothly from the end of Luke to the start of Acts. However, Greenwood's (1995) computer study suggested that only the early chapters of Acts resemble Luke stylistically, while the later chapters, describing Paul's missionary journeys, are stylistically distinct. His technique was to use the hierarchical clustering algorithm of Hartigan and Wong on the frequencies of the most common words in the whole data set for each chapter of Acts. To investigate this question for ourselves, we performed a correspondence analysis on our set of 500-word samples for both Luke and Acts, taking into account the frequencies of the 75 most common Greek words in the Johannine corpus. Our results are shown in Figure 3, where the text samples numbered 1 to 38 are from the Gospel of Luke, those numbered 39 to 44 are from the first 12 chapters of Acts, and the remainder are from the latter part of Acts. It can be seen that while Luke and the latter part of Acts are largely distinct from each other, the early chapters of Acts are in an intermediate position. They cluster closely together, so clearly have much stylistically in common with each other, but appear to be distinct from both the Gospel of Luke and the later chapters of Acts.

4. The Pauline Epistles and Hebrews

Some of the Epistles of Paul are more widely thought to be authentic than others. The general consensus is that the four so-called "Hauptbriefe" (Romans, 1 and 2 Corinthians, and Galatians) are certainly authentic. Most scholars also accept 1 Thessalonians, Philippians and Philemon. The most disputed letters are Colossians, Ephesians, and 2 Thessalonians. The Pastorals (1 and 2 Timothy and Titus) and Hebrews are considered least likely to have been written by Paul, and indeed, the authorship of Hebrews is a complete mystery. Since the Reformation, Hebrews has been generally considered not to have been written by Paul, partly because unlike in his other letters, Paul does not introduce himself at the start. However, there is some kind of link with Paul, as Hebrews mentions Timothy as the author's companion. Computer studies of the writing style of the Pauline Epistles have been performed by Neumann (1990:191). Using the Mahalanobis distance of various texts from each other, based on a matrix of texts and the counts of a large number of linguistic features, he concluded that the Pastoral Epistles were not written in the style typical of St. Paul's more accepted writings. Using another multivariate technique called discriminant analysis, he compared the distances of disputed texts from the "Pauline centroid", the main cluster of accepted texts. This technique was thus a form of outlier analysis, and Neumann concluded that there was "little reason on the basis of style to deny

authenticity" to the disputed letters Ephesians, Colossians and 2 Thessalonians. Other multivariate techniques, namely principal component analysis and canonical discriminant analysis were performed by Ledger (1995). He found that 1 and 2 Corinthians, Galatians, Philemon, 2 Thessalonians and Romans seem to form a "core Pauline group", while Hebrews was a definite outlier. He also felt that the authorship of all the remaining letters was doubtful. Greenwood (1992), again using the hierarchical clustering technique of Hartigan and Wong, found distinct clusters corresponding to the Missionary, Captivity and Pastoral letters.

The results of our own correspondence analysis of the New Testament epistles are shown in Figure 4. As well as those letters traditionally attributed to Paul, for comparison we included the letters of John ("Jn1", "Jn2" and "Jn3"), James ("jam"), Jude ("jude") and Peter ("1Pet", "2Pet"). The first letter of John was most clearly distinct from all the other letters, with relatively high occurrences of "εσπν" and "οπ". The four Hauptbriefe, ("1cor", "2cor", "rom" and "gal", all in darker type) all group together on the left hand side of the graph, suggesting that they make a homogeneous group. The disputed Pauline Epistles are mainly found on the right hand side of the graph, with Ephesians ("eph") and Colossians ("col") the most distinct from the Hauptbriefe. Although Hebrews ("heb") is not thought to be written by Paul, the Hebrews samples form a close cluster on the borderline between Paul's Hauptbriefe and the disputed letters. Differences in the styles of Paul's letters might have arisen through dictation to various amanuenses. We do know that on at lest one occasion Paul used dictation, as Romans 16:22 contains the words "I Tertius, the writer of this letter"[3].

5. The Signs Gospel

The term "Signs Gospel" was first used by C.H. Dodd (1963) to refer to chapters 2 to 12 of the Gospel of John. He used this name because these chapters describe Jesus' seven public miracles, which were signs of his messianic identity (Thatcher, 2001). Later, this "Signs Gospel" was thought to also consist of a Passion narrative. There are four main theories regarding the use of early sources in John's Gospel. Firstly, we have the oral tradition theory of Dodd and Thatcher themselves, which is that many sayings of Jesus were drawn from an oral tradition, some of which was also used by the writers of the Synoptic Gospels. The written source theory is that John's Gospel was drawn from two written sources, a miracle source and a version of the Passion story which had been combined before the time of John. These postulated sources have since been lost. The third theory is the synoptic dependence theory, in which the Gospel of John was also based on written sources, most clearly the Synoptic Gospels. The problem with this theory is that the differences between John's Gospel and the other three are much greater than the similarities between them, but recently the Leuven school have come to believe that there are some key correspondences such as Luke 24:12 and John 20:3-30. Fourthly, the developmental theory is that the Gospel was based on repeated editing by a Johannine community (Thatcher, 1989). Felton and Thatcher (1990) performed a stylometric analysis where the t-test was used to compare texts thought to be from the Signs Gospel with those from the remainder of John, according to the frequencies of certain linguistic features such as the number of definite articles in each text block, verb-verb sequences, and the number of words containing from each of 1 to 10 characters. Their results were inconclusive.

We used a correspondence analysis to determine whether the text of the putative Signs Gospel differs stylometrically from the rest of the book of John. Once again we used the Westcott and Hort original Greek text, and this time followed the reconstruction of the Signs Gospel given by Fortna (2010:189). Some of our findings can be seen in Figure 1, the correspondence analysis of the New Testament as a whole. The 500-word samples from John's Gospel are labeled "s" for the Signs Gospel, and "jg" for the rest of John's Gospel except for the 3 samples labeled "jf" which come from a passage in John known as the "Farewell Discourse". For comparison, the letters of John are labeled "jl", in contrast to all the other letters in the New Testament which are simply labeled "l". Nearly all the samples from John's Gospel are low down in the south-east quadrant of the diagram, showing that the writing style in John's Gospel as a whole is quite distinct from the rest of the New Testament. The "s" samples are grouped close to each other, and so although it has been suggested that the Signs Gospel was originally composed by combining a life narrative with a passion narrative, there is no evidence for more than one author. The "s" samples are not far from the "jg" samples, but a t-test for matched pairs showed that there were significant differences between the sets of co-ordinates for both factor 1 ($p = 0.0005$) and factor 2 ($p = 0.0013$) for the two sets of samples. It remains for us to determine whether these small differences really

[3] http://ww2.ferrum.edu/dhowell/rel113/pauls_letters/pathway.htm

were due to distinct writing styles in the two sources, or were in some measure due to genre. Three samples from the Gospel of John did stand out, namely the "jf" samples of the "Farewell Discourse" (John 13:31 to 17:26). These samples had more in common stylometrically with the letters of John (which are widely thought to be authentically by John) than with the other parts of John's Gospel.

6. Conclusions

In this paper we have described the use of correspondence analysis to first map out the main stylometric groupings (letters, Synoptic Gospels, Revelation and John) among the New Testament texts as a whole. We have shown that John in both his Gospel and letters is distinct in his use of common words from the other books of the New Testament. Revelation and the John samples are at opposite poles of the correspondence analysis plot, showing that as is commonly supposed, they authors are unlikely to be one and the same. We then examined specific controversies about New Testament authorship. We have reproduced Mealand's finding that Q is stylometrically distinct from the rest of Luke, and shown that the first 12 chapters of Acts form a homogeneous group, intermediate in style between Luke and the rest of Acts. In our on-going study of the Signs Gospel, a possible precursor of John, we found that the Signs samples clustered very close together, suggesting single authorship of those samples. However, the positions on the plot of the Signs samples were only slightly different from the rest of John, and we have not yet controlled for the fact that these differences in position might be due to differences in genre. There are at least three suggestions in the literature for factoring out genre. One, by Mealand (2011), is that the first factor in a correspondence analysis, accounting for most of the variation between text samples, might be the one most due to genre differences, and thus later factors might be less affected. A plot where the axes are the second and third factors might then show more clearly differences in individual style than the more commonly displayed plots where the axes are the first and second factors. A second suggestion is simply to compare "like with like". Thus for example, only samples in the same genre, such as narrative texts, should be compared in the same analysis. Thirdly, Linmans (1995, 1998) has proposed using correspondence analysis in conjunction with another multivariate technique, log linear analysis (LLA), to factor out genre differences.

The work described here stands in contrast with the work of Jockers et al. (2008) and Sadeghi (2011), who performed authorship studies on the Book of Mormon and the Quran respectively. These differed from our starting point in that the accepted assumption is that these texts have a single author, and stylometrics were used to help test this assumption.

7. References

Baayen, R. H. (2008). *Analysing Lingusitic Data*. 2008. Cambridge University Press.

Benzecri, J-P. (1980). *L'analyse des données*. Tome 2: L'analyse des correspondances. Paris: Bordas.

Felton, T. and Thatcher, T. (2001). Stylometry and the Signs Gospel. In R. T. Fortna and T. Thatcher (editors), *Jesus in Johannine Tradition*. Louisville-London: Westminster John Knox Press, pp. 209-218.

Fortna, R. T. (2010). The Signs Gospel. In: Robert J. Miller (editor), *The Complete Gospels*, Fourth Edition, Polebridge Press, Salem, Oregon.

Greenwood, H. H. (1995). Common word frequencies and authorship in Luke's Gospel and Acts. *Literary and Linguistic Computing,* 10(3), pp. 183-187.

Greenwood, H.H. (1992). St. Paul revisited – A computational result, *Literary and Linguistic Computing*, 7(1), pp 43-47

Jockers, M.L., Witten, D.M. and Criddle, C.S. (2008). Reassessing Authorship of the Book of Mormon using Delta and Nearest Shrunken Centroid clustering. *Literary and Linguistic Computing* 23(4), pp. 465-492.

Ledger, G. (1995). An exploration of differences in the Pauline Epistles using multivariate statistical analysis. *Literary and Linguistic Computing* 10(2), pp. 85-96.

Linmans, A.J.M. (1995). *Onderschikking in de Synoptische Eevangelien.* Nederlandse organisatie voor wetenschappelijk onderzoek.

Linmans, A. J. M. (1998). Correspondence analysis of the Synoptic Gospels. *Literary and Linguistic Computing*, Vol. 13(1), pp. 1-13.

Mealand, D.L. (1995). Correspondence analysis of Luke, *Literary and Linguistic Computing* 10(3), pp. 171-182.

Mealand, D.L. (2011). Is there stylometric evidence for Q? *New Testament Studies*, 57, pp. 483-507.

Poirier. J. C. (2008). Statistical studies of the verbal agreements. *Currents in Biblical Research* 7(1), pp. 68-123.

Sadeghi, B. (2011). The chronology of the Quran: A stylometric research program. *Arabica* 58 (3-4), pp. 210-299.

Thatcher, T. (2001) Introduction to *Jesus in Johannine Tradition*, In Robert T. Fortna and Tom Thatcher (eds), Louisville and London: Westminster John Knox Press.

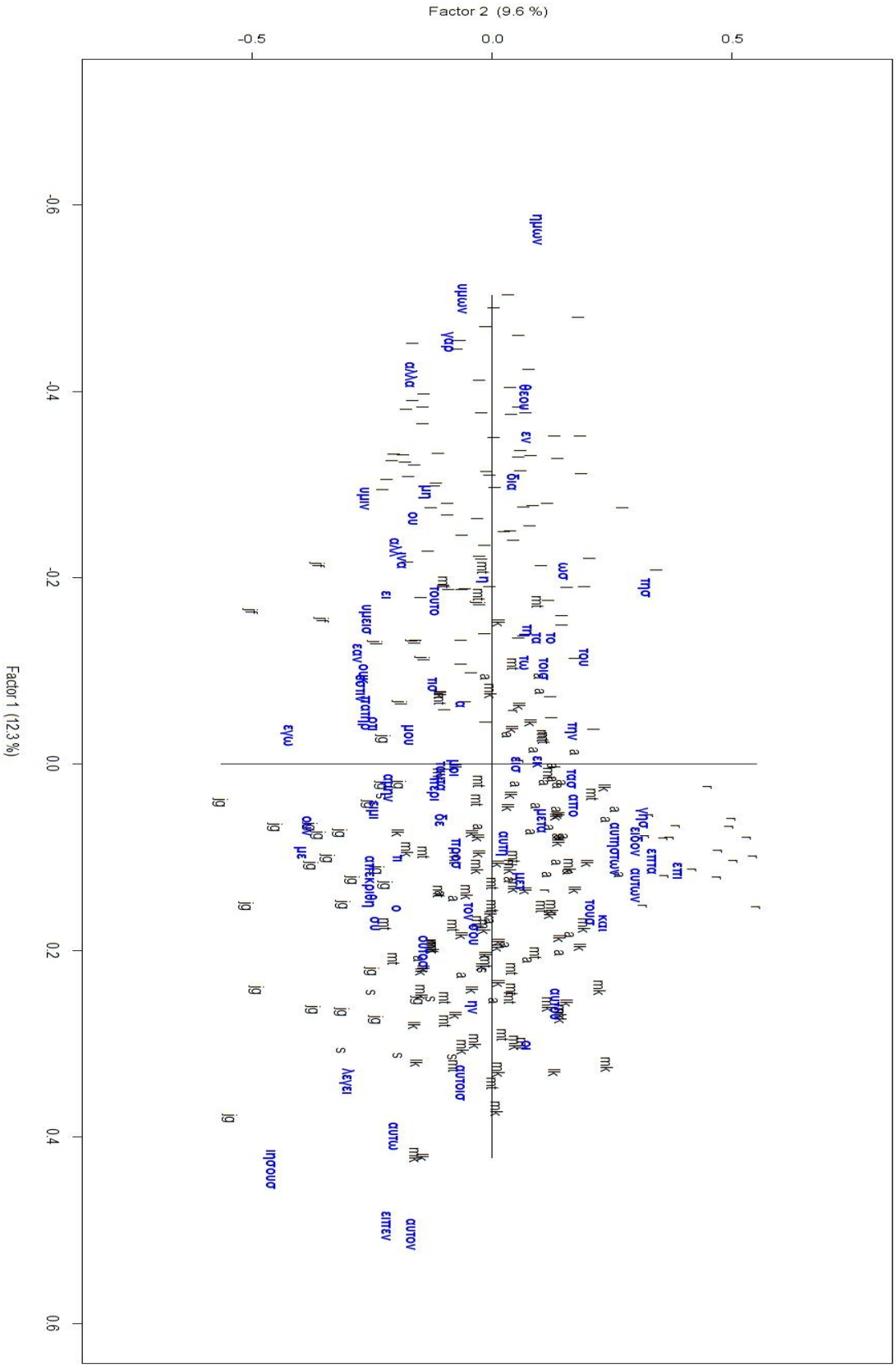

Figure 1. Overview of the New Testament by Correspondence Analysis. The four main clusters are Revelation ("r"), the Synoptic Gospels ("mt", "mk", "lk"), Epistles or Letters ("l") and the Gospel of John ("jg").

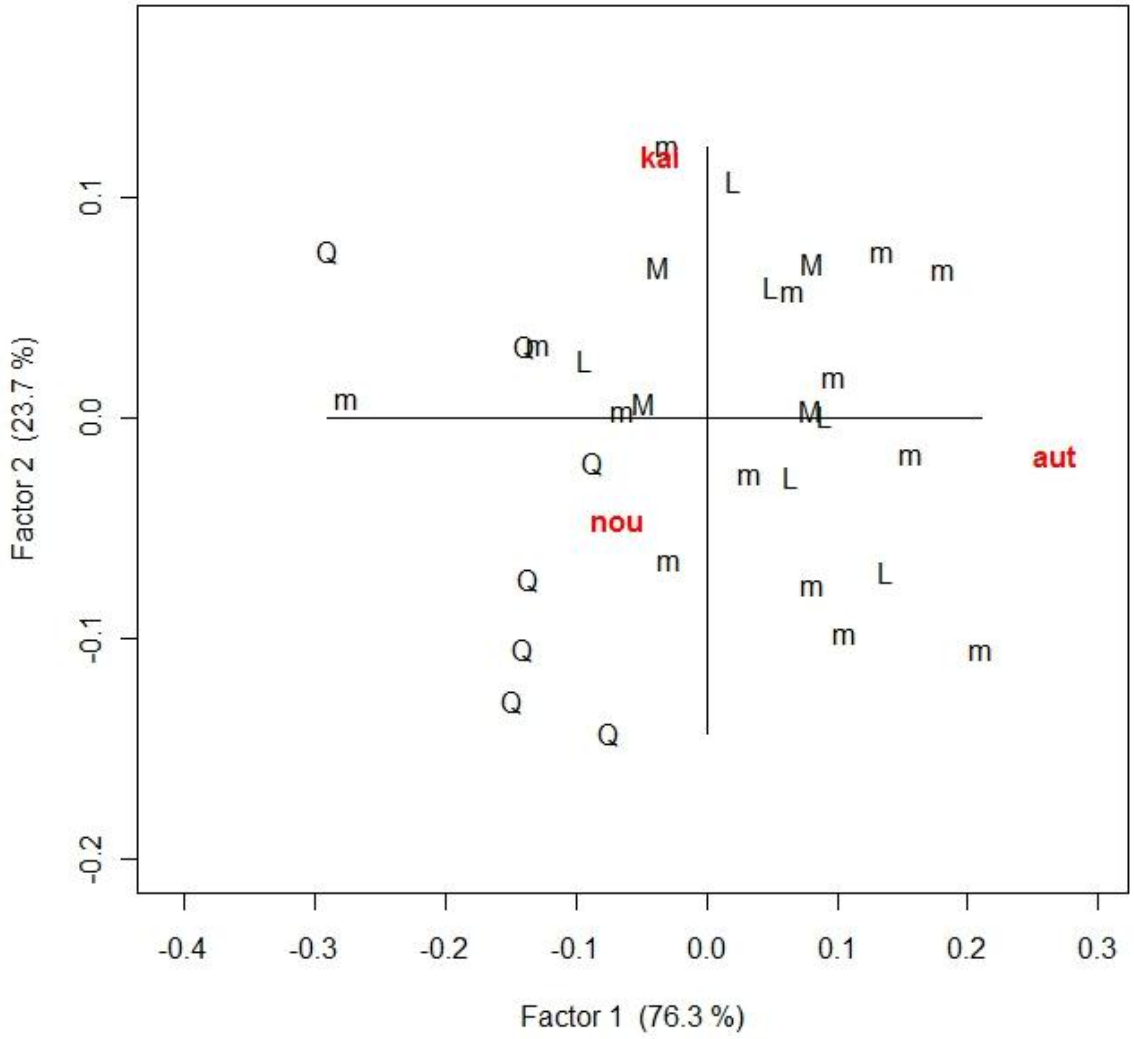

Figure 2. Evidence for Q from Mealand's (1995) Data. The samples of Q broadly stand out from the other material in Mark and Luke.

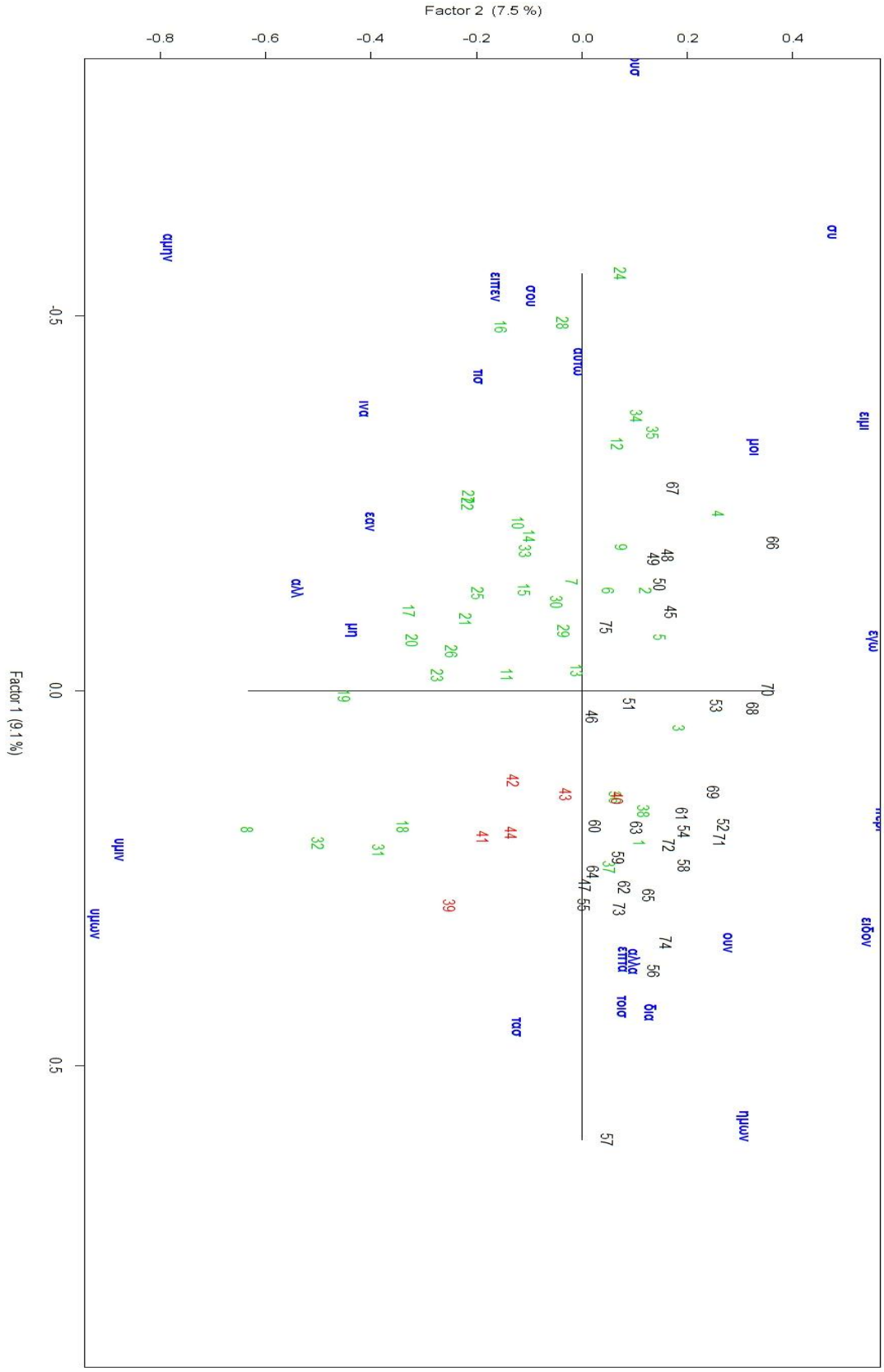

Figure 3. Comparison of Luke and Acts by Correspondence Analysis. Three main clusters are seen: Luke (samples 1-38), the first 12 chapters of Acts (samples 39-44), and the remainder of Acts (samples 45-75).

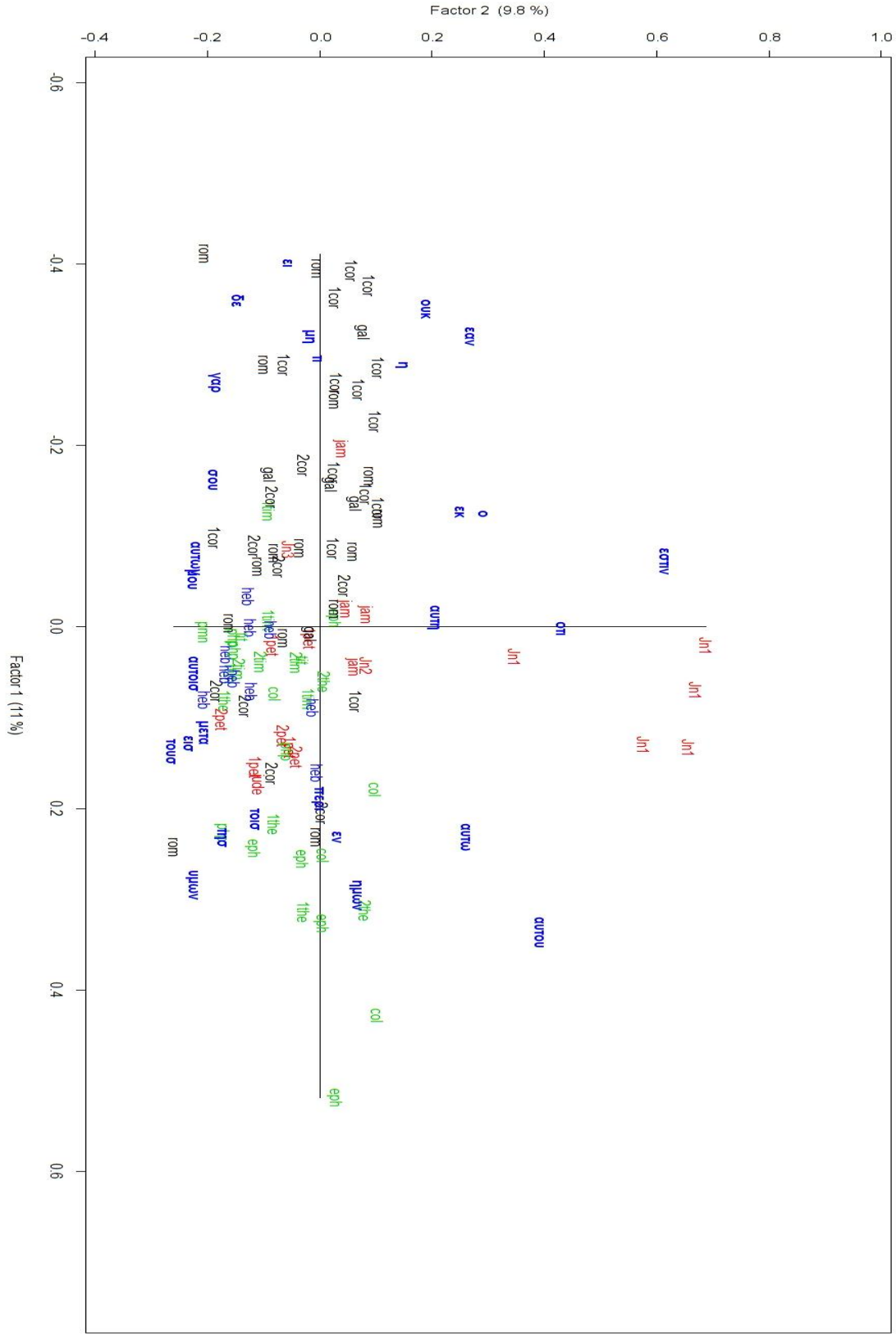

Figure 4. Comparison of the New Testament Epistles by Correspondence Analysis. The four *Hauptbriefe* ("1 cor", "2 cor", "rom", "gal") all group together on the left hand side of the graph, while Ephesians ("eph") and Colossians ("col") are the most distinct from the core Pauline group.