1. In his I.C.C. commentary on the Fourth Gospel, Bernard discusses the relation of the Gospel of John to the Gospel of Mark and points out – The Gospel According to St. John, T. and T. Clark, Edinburgh, 1942, introduction page 97 para. 2 – that John repeats the wording of Mark 2.9 in his own Gospel at 5.8 modifying the sequence of eight Markan words only by omitting the first occurrence of 'kai'. This verbal coincidence Bernard takes as one reason for supposing that John is here dependent on the Markan text for his account.

This is one illustration of a method generally employed by scholars. When two texts have a series of words in common, a series distinctive enough to be uniquely identified, then the verbal coincidence is taken as proof of the existence of some relation between the two texts.

In practice the method depends for its validity on a two-part argument. First there must be a credible historical background which makes it possible to accept that some relation between the texts is possible; if this background is lacking, then verbal coincidences are dismissed as the product of pure chance. The second part of the argument is that the words involved in the series of coincidences must be numerous enough, and distinctive enough, to rule out any chance coincidence of wording, such as can be expected to occur in any two texts from the same cultural milieu, as being the true explanation of the coincidence.

It is with this second proposition, that chance could not have created the verbal coincidence between the two texts, that this analysis is concerned and the sole aim of the analysis is to set out the basic condition which must be met before it can be said, with any degree of confidence, that chance can be excluded as the cause of the verbal coincidences between texts.

2. Suppose we have a text which contains n words. The only stipulation to be made is that n is a number large enough to be substantially unchanged by the addition to it or substraction from it of 1; so that $(1 - \frac{1}{n})$ can be taken as 1. If n is 1,000, then $(1 - \frac{1}{n}) = 0.999$; so n need not, in literary terms, be a large number of words.

Now if we take from another group of words, a single word which occurs in the group of n and compare our single word with each word in the group of n in turn, then once we will get a coincidence and n - 1 times we will not. So the probability of getting a coincidence in a single comparison is $\frac{1}{n}$ and of not getting a coincidence in a single comparison is $(\frac{n - 1}{n})$. This last expression can be written as $(1 - \frac{1}{n})$. If we repeat the comparison n times, the number of cases which show no coincidence will be given by $(1 - \frac{1}{n})^n$ and this equals e - 1 where e is the natural base of logarithms and a number which plays an important role in mathematical theory. Tables showing the values of e to different powers are to be found in almost any set of mathematical tables and from such a table e^{-1} is given as 0.3678794.

If the proportion of comparisons which show no coincidence is e^{-1}, then the proportion which show a coincidence must be 1 - e^{-1}, which is 0.6321206.

If we then add the same word to our comparison word and to each of the words in the group of n words, then all the cases of no coincidence will become cases of two coincidences, all case of two words coinciding will become cases of three words coinciding and so on. The proportion of one word coincidences will be e^{-1}, the same as the former cases of no coincidence.

If we then add another word to each comparison, the cases wich were of no coincidence for the original comparison will now become cases of a double comparison but each such coincidence will be counted twice as no allowance has been made for the order of the ., ords in the coincidence and as we are interested in sequential coincidences the number of double coincidences, e^{-1}, must be divided by 2 to eliminate the effect of word order. The argument can be continued indefinately, for treble coincidences we have a proportion $\frac{e^{-1}}{3!}$, where 3! is shorthand for 1 x 2 x 3 = 6 and is again the correction needed to eliminate the duplication due to counting the words in all possible orders.

Similarly for 4 coincidences the proportion will be $\frac{e^{-1}}{4!}$ and for x coincidences it will be $\frac{e^{-1}}{x!}$.

Table 1 can now be drawn up to show the proportion of coincidences between two groups of words. As it stands Table 1 is not directly useful and it might be objected that in the construction of the Table no attention had been paid to the common repetition of some word forms and the rarity of others. This objection is irrelevant for two reasons. First in a comparison of this kind an occurrence of 'kai' in a supposed verbal coincidence is not being compared with any occurrence of 'kai' in the text but only with the unique occurrence which comes at the right place in the series of words involved in the verbal coincidence.

Secondly the effects of grammar will be to make some kinds of verbal coincidence occur more often than would be the case if language was a random mixture of separate words, but the fact that these kinds occur more often means that other coincidences of the same dimension occur less often than pure chance would generate them.

In the New Testament, for example, the three words 'the kingdom of' are likely to be followed by the word 'heaven'. But this diminishes the chance of finding the kingdom of hell and leaves the number of four word sequences unaltered.

The aim of this analysis is to lay a foundation, to calculate the minimum influence of chance on verbal coincidences regardless of language, literary form or any other factors and it is therefore important not to overestimate the effect of chance but to underestimate it in any doubtful case.

For repeated events, such as the occurrence of word coincidences, the probability of an occurrence can be defined as the ratio of the actual occurrences of the event to the total number of possible occurrences of the event. Thus the probability of an occurrence of each coincidence is the proportion which has been set out in table 1. In the scale of probability an event certain to take place at every trial would have $p = 1$. Any event certain not to take place no matter how many trials are made, has a probability $p = Q$. For all other events p is more than 0 but less than 1.

From Table 1 we can calculate the number of verbal coincidences of any dimension, likely to be generated by chance, and go on to compare this chance expectation with what is found. To work out the chance expectation we need to know three things. First is the number of words in the text from which the verbal coincidence is suspected to have been derived. In our example this is the Greek text of the Gospel of Mark which contains 11,242 words. Next we need to know the number of words in verbal coincidence and the number of words in the sequence in which the coincidences occur. These two numbers should be much the same and indeed, for the theory it has been assumed that they are equal. But in practise one finds that a few words are often omitted from a sequence as it is copied into a new document. But in theory the figures apply only to identical sequences of words. In the example, if the two Greek texts are laid side by side and the common words are underlined – it seems to make no appreciable difference if accidence is ignored – it is seen that there are seven coincidences in a sequence of eight words. The probability of a seven-word coincidence is, from Table 1, 0.0000730. The number of eight-word sequences in the Gospel of Mark we can calculate as follows : a sequence of eight words can begin with the first word in the text, another with the second word and so on up to the eighth last word. So the number of sequences in the text is the number of words in the text, less the number of words in the sequence. In the example : 11,242 less 8 = 11,234. The expected number of seven-word coincidences is then 11,234 x 0.0000730 = 0.82. The agreement of what is found and what is expected to be generated by chance alone is so near that it might suggest a lack of confidence that a coincidence of this size is any proof of dependence of the two texts.

In fairness to Bernard it must be said that he furnished further verbal coincidences and, in the instance which has provided the illustration, he

For our example the likelihood ratio is that of the number of actual occurrences, to the chance expectation; 0.82.

In general statistical work it is usual to exclude chance as the explanation of events when pure chance would account for them only once in twenty cases and the equivalent likelihood ratio is 20 : 1. For this example, if we follow accepted practice and make the smaller figure unity, the ratio is 1.2 : 1.

Likelihood ratios are useful for evaluating verbal coincidences which are large enough by themselves to be proof of dependence of two texts, usually coincidences of six words or more.

To assess smaller coincidences the chi squared test should be used. This test applies only where a minimum of five occurrences is to be expected and the test is fully explained in almost any textbook of statistics, for example, M. J. Moroney, Facts from Figures, Penguin Books, London, 1962, chapter 15.

There is one difficulty in dealing with short coincidences. Scholars comment on pairs of words shared between texts, they even comment on single words, but from Table 1 it is clear that there are very large numbers of such coincidences. Most of them involve the more frequent word-forms and pass unnoticed. However a few of them will involve rare words and scholars are impressed by these striking coincidences, quite unaware that they are making an unconscious selection from a large number of such coincidences and unaware that these coincidences seem significant because they are the only coincidences which their training prepares them to detect.

limited his argument to the dependence of the single narrative in which the sequence of words is found rather than to the Gospel as a whole. Against this it should be noted that his other verbal agreements seem no more out of line with chance expectation and that limiting the argument to a section of the text can defeat its own purpose. This is most easily seen in an example. The words "this too too solid flesh" are a five-word coincidence and, from Table 1 this alone is not good evidence of an acquaintance with the play Hamlet. But if the argument is narrowed to the five words in isolation then we compare the one occurrence of the five-word coincidence. This is convincing proof of a knowledge of the five-word quotation, but knowledge of the play need supporting examples from other parts of the text.

There are two simple methods of making precise comparisons of the chance expectation of coincidences and the actual occurrence of such coincidences. The first is to use the appropriate likelihood ratio. In ordinary language the likelihood ratio expresses the relative odds for two hypotheses which can explain the same data. For instance if one piece of Greek prose had one hundred 'kais' in it and we knew that Peter wrote such a piece once in twenty times, probability 0.05, while Paul wrote such a piece only once in two hundred times, probability 0.005, then the likelihood ratio for this situation is the ratio of the two probabilities, 0.05 : 0.005, which is 10 : 1 for Peter against Paul. The scholar who chooses Peter as the author has ten times more chance of being correct in his decision than a scholar who selects Paul.

Obviously likelihood ratios of around 1 : 1 indicate little or nothing to choose between the alternative explanations of the evidence while ratios of 20 : 1 or 1 : 20 show a weight of evidence for one or other hypothesis.

For our example the likelihood ratio is that of the number of actual occurrences, to the chance expectation; 0.82.

In general statistical work it is usual to exclude chance as the explanation of events when pure chance would account for them only once in twenty cases and the equivalent likelihood ratio is 20 : 1. For this example, if we follow accepted practice and make the smaller figure unity, the ratio is 1.2 : 1.

Likelihood ratios are useful for evaluating verbal coincidences which are large enough by themselves to be proof of dependence of two texts, usually coincidences of six words or more.

To assess smaller coincidences the chi squared test should be used. This test applies only where a minimum of five occurrences is to be expected and the test is fully explained in almost any textbook of statistics, for example, M. J. Moroney, Facts from Figures, Penguin Books, London, 1962, chapter 15.

There is one difficulty in dealing with short coincidences. Scholars comment on pairs of words shared between texts, they even comment on single words, but from Table 1 it is clear that there are very large numbers of such coincidences. Most of them involve the more frequent word-forms and pass unnoticed. However a few of them will involve rare words and scholars are impressed by these striking coincidences, quite unaware that they are making an unconscious selection from a large number of such coincidences and unaware that these coincidences seem significant because they are the only coincidences which their training prepares them to detect.

The smaller coincidences are best read by a sliding fit program for a computer which literally compares a single word with every word in the text and then repeats the search for sets of two words, three words, four words, and so on.

3. This single example has shown that even such a reputable and cautious scholar as Bernard had considerably underestimated the effects of chance generation of verbal coincidences, Now we must turn to scholars in general, and see if they are all as unwittingly optimistic. This seems to be so. The point can be made by a single set of examples. In his book, The Formation of the Pauline Corpus of Letters, Epworth Press, London, 1955, Dr. C. L. Mitton refers, pages 20 and 21, to a review of the Apostolic Fathers carried out by a "combined research committee of distinguished Oxford scholars". Verbal coincidences with the New Testament books in the Apostolic Fathers were classified. "Conclusive proof" was given an A : "highly probable" was given a B : "less probable but quite likely" echoes were given C : "possible but not probable" acquaintance was given D.

Dr. Mitton was interested in the Pauline Corpus alone and so prints a table of the committee's findings on all the Pauline epistles. Only Romans and 1st Corinthians get A's and then only in 1st Clement, Ignatius and Polycarp. Romans being given A only in 1st Clement. The A class marking looks quite sound, for example 1st Clement 35.5 is parallel to Romans 1.29-30 and there is a coincidence of ten words in a sequence of seventeen. The number of words in Romans is 7,105, the number of sequences is therefore 7,105 less 17 which is 7,088. The chance of a coincidence of ten words is just about one in a million, so the likelihood ratio for this one coincidence is around 140 : 1.

But when we turn to the B list the results are less satisfactory. 2nd

Corinthians is taken as known to Polycarp. Reference to the Loeb text of Polycarp shows three references in the margin to 2nd Corinthians, Polycarp 2.2 and 2nd Corinthians 4.4; 6.1 and 2nd Corinthians 8.21; 6.2 and 2nd Corinthians 5.10. These are, respectively, a coincidence of five words in a sequence of ten words, of two words in two words, of four words in four words. 2nd Corinthians has 4,469 words in it so we can expect to find sixteen five-word coincidences, more than eighty four-word coincidences and several hundred two-word coincidences, generated by chance alone.

It is difficult to resist the conclusion that the committee's categories B, C and D, have little or no meaning and that only 1st Corinthians was generally known to the Apostolic Fathers though 1st Clement also knew Romans.

Such a conclusion would strengthen Dr. Mitton's argument about the collection of the letters.

4. Two immediate conclusions can be drawn from this investigation.

First it has shown that scholars, who are careful in their choice and use of words, tend to assume that some words, such as "chance" and "probability" have a commonsence use and an agreed meaning. The fact is that in the last few decades these words have acquired a highly specialised analytical background. It would be wise for scholars to treat these words with caution.

The second conclusion is that scholars have seriously underestimated the effects of chance in the generation of verbal coincidences and the work done in this field now needs revision.

Table One

The Chance Generation of Verbal Coincidences

Number of words in coincidence.	Probability of occurrence.	Number of words in coincidence.	Probability of occurrence.	Approximation for Probability of column 4.
0	0.3678794	1 or more	0.6321206	2/3
1	0.3678794	2 or more	0.2642412	1/4
2	0.1833397	3 or more	0.0803015	1/12
3	0.0613132	4 or more	0.0189883	1/53
4	0.0153283	5 or more	0.0036600	1/273
5	0.0030657	6 or more	0.0005943	1/1683
6	0.0005189	7 or more	0.0000834	1/12,000
7	0.0000730	8 or more	0.0000105	1/95,000
x	$\frac{e^{-1}}{x!}$			

NOTE : The figures apply exactly only to coincidences of successive words but they can be used for sequences where most of the words coincide with a quotation, even when a few words have been omitted from, or inserted into, the text being quoted.

A. Q. MORTON
Dept. of Computer Science,
University of Edinburgh,
8 Buccleuch Place,
Edinburgh, 8.