

Unraveling the Synoptic puzzle: stylometric insights into Luke's potential use of Matthew

Sophie Robert Hayek^{1,2,*}, Jacques Istas² and Frédérique Michèle Rey¹

¹Laboratoire Ecritures, Université de Lorraine, France

²Laboratoire Jean Kuntzmann, Université Grenoble Alpes, France

Abstract

The literary sources behind the three canonical Synoptic Gospels, namely Luke, Matthew and Mark, have long intrigued scholars because of the Gospels striking similarities and notable differences in their accounts of Jesus's life. Various theories have been proposed to explain these textual relationships, including common oral witnesses, lost sources or communities possessing each other's works. However, a universally accepted solution remains elusive. Leveraging advancements in statistics, data analysis, and computing power, researchers have begun treating this as a statistical problem and quantitatively measuring the likelihood of the different theories based on verbal agreements and stylometric features. In this paper, we rely on a very recent Machine Learning based approach to solve the synoptic problem. We use Machine Learning classifiers two-sample tests, a novel approach relying on the analysis of the success rate of binary classifiers to identify whether two samples are drawn from the same distribution, to detect differences in sources within Luke's Gospel and variations in the edition patterns of Markan material between Matthew and Luke. This analysis is done on a pericope-per-pericope basis, defined as thematic units encompassing teachings or narrative episodes. The results suggest significant dissimilarities in style and edit distance, indicating that the double and triple material within the Gospel of Luke likely originate from different sources. This suggests that Luke derived his triple tradition from Mark and not from Matthew. Despite the necessity of cautious interpretation due to the size of the dataset, our study thus offers substantial evidence supporting the theory of Luke's dependency on Mark's material for his triple tradition and makes the two-source hypothesis, which suggests that Luke did not have access to Matthew's work, the most likely explanation based on our methodology.

Keywords

synoptic problem, New Testament, stylometry, digital humanities;

1. Introduction

The striking similarities, along with the notable differences, between the different accounts of Jesus' life related in the three canonical Synoptic Gospels, usually referred to as the Gospel of Luke (Lk), Matthew (Mt) and Mark (Mk), have led scholars since Antiquity to speculate on the order of their composition and the literary sources available to each community the Gospels came from. Over the years, several theories have been proposed to explain this phenomenon,

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France


*Corresponding author.


†These authors contributed equally.

✉ sophie.robert@univ-lorraine.fr (S. Robert Hayek); jacques.istas@univ-grenoble-alpes.fr (J. Istas);

frederique.rey@univ-lorraine.fr (F. M. Rey)

ORCID [0000-0003-4359-9124](https://orcid.org/0000-0003-4359-9124) (S. Robert Hayek); [0000-0003-0028-4931](https://orcid.org/0000-0003-0028-4931) (F. M. Rey)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

including the idea of shared oral witnesses [8], the existence of now-lost common sources [35], and the possibility of the community having access to each other's works [9, 10]. The debate remains heated and no solution has been universally accepted as the most likely one.

With the recent advances in statistics and data analysis, as well as the exponential advances in computing speed, several researchers have suggested treating this problem as "a kind of problem in Synoptic arithmetic" [34, p. 202] and have attempted to provide some quantitative measurements regarding the likelihood of each theory, either by studying the verbal agreements, *i.e.* occurrences of the same words in the same context in two or more Gospels, or the stylometric features of the text.

In this paper, we suggest a novel approach based on the use of a Machine Learning classifier two-sample tests to detect different sources within the Gospel of Luke and different edition patterns of the Markan material between Matthew and Luke. We compare statistically the distribution of the stylometric features and show that the theory of Luke and Matthew being independent, also known as the two-source hypothesis, is the most likely given our methodology.

The major contributions of this study are:

- The introduction of a novel paradigm by conducting stylometric analysis at the pericope level rather than focusing on individual verses when comparing the Synoptic Gospels.
- The use of a very recent statistical approach, Machine Learning-based two-sample test, to identify stylistic differences within double and triple material of Luke's Gospel.
- The application of this Machine-Learning based approach to characterize the edits made by Luke and Matthew when it comes to the Markan material.

The results of our study allow us to lean towards the two-source hypothesis as being the most likely one given the stylometric data of Luke's Gospel.

This paper is structured as follow. We give in section 2 more insights regarding the synoptic problem and the major solutions suggested over the history of the Gospels. Section 3 presents works related to ours which rely on statistical methods to propose solutions to the synoptic problem. Section 4 presents our selected stylometric approach and its results when applied to the *SBL Greek New Testament* (SBLGNT) text [36] are presented in section 5. We then conclude our work and give some insights into our further works in section 6.

2. Motivation: the synoptic problem and the two-source hypothesis

Solving the synoptic problem consists in providing a theory that describes the relationships between the three canonical Synoptic Gospels (Matthew, Mark, and Luke) by analyzing their similarities and differences. If the potential influence of oral traditions on early Christian teachings cannot be overlooked [8] to explain the parallels between the gospels, the often *verbatim* wordings and the overall alignment of pericopes indicate a common written source. The pericopes can be categorized as *single tradition* (found in only one Gospel, such as the different accounts of Jesus's birth), *double tradition* (present in Matthew and Luke, such as the teachings in the Sermon on the Mount and the Sermon in the Plain in Mt 5-7 // Lk 6:17-49), or *triple*

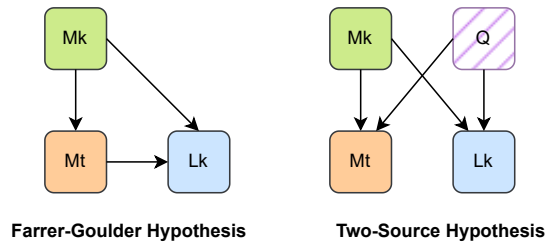


Figure 1: Graphical representation of the Two-Source Hypothesis and the Farrer-Goulder Model. Square in dashed lines indicate an hypothetical source that has never been found.

tradition (found in all three Gospels, often with variations in order and context, such as the healing of the paralytic found in Mt 9:1–8, Mk 2:1–12, and Lk 5:17–26). The distribution of the different material per verse number and per pericope is available in table 1.

Traditionally, the Augustinian model [3] proposed that Matthew was written first, followed by Mark as a summary of Matthew, and finally Luke¹. However, the past two centuries of textual criticism have challenged this model, starting with Griesbach’s hypothesis [29] that Matthew was the initial Gospel, which was later used by Luke, while Mark combined elements from Luke and Matthew [9, 22]. The dominant theories in the 20th century presuppose Markan priority, suggesting that Mark’s Gospel was written first and served as a source for the two other canonical evangelists. When assuming Markan priority, two competing theories currently co-exist: the *Farrer-Goulder hypothesis* (FGH) [10, 12] proposes that Luke had access to both Mark and Matthew, drawing material from both and the *Two-Source Hypothesis* (2SH or Q hypothesis) [35, 37], now widely accepted, which posits that Luke and Matthew utilized a source composed of sayings of Jesus without any narrative material, called Q (abbreviation of *Quelle*, German for source), along with Mark’s Gospel. Matthew and Luke consequently derived their double tradition material from Q and their triple tradition material from Mark. Figure 1 provides a schematic representation of these two theories. As Farrer underlines in [10], the Q theory “*wholly depends on the incredibility of St Luke having read St Matthew’s book*”, as this theory relies on the assumption that Luke deliberately altered key sayings of Jesus, such as the Lord’s prayer (Mt 6:9–13, Lk 11:2–4). In this study, we presuppose Markan priority and leverage stylometric arguments to determine the likelihood of Luke having read Matthew’s Gospels and determine which theory is better supported by the evidence gathered from the stylometric features.

3. Related works

The use of statistical data in the analysis of the synoptic problem goes back to at least the end of the nineteenth century, when Hawkins’ *Horae Synopticae* was published in 1899 [15]. Ever since the recent advances in computing power and statistical algorithms, the Synoptic Gospels have received wide attention from statisticians.

The research can be separated into two wide categories: the study of *verbal agreements* [27,

¹Followed by John. Due to the complexity and ongoing research in the field of the relationship between Johannine literature and the Synoptic Gospels, this article does not delve at all into this subject matter.

Table 1

Distribution of pericopes across genres and traditions. Our results vary from the now widely accepted statistics provided by Honoré [17] as we are working on a per pericope basis instead of on a per verse basis and do not take into account the number of verses per pericope.

(a) Pericope repartition across tradition			(b) Honoré’s statistics on a verse basis		
Book	Tradition	%	Book	Tradition	%
Lk	Lk-Mk	3	Lk	Lk-Mk	1
	Lukan	22		Lukan	35
	Double	22		Double	23
	Triple	54		Triple	41
Mk	Lk-Mk	5	Mk	Lk-Mk	3
	Markan	5		Markan	3
	Mt-Mk	8		Mt-Mk	18
	Triple	82		Triple	76
Mt	Matthean	13.96	Mt	Matthean	20
	Mt-Mk	6		Mt-Mk	10
	Double	23		Double	24
	Triple	57		Triple	46

17, 7, 1, 21, 32], consisting in counting the number of occurrences of the same word being present in two Gospels within the same narrative framework, either in its inflected or lemmatized forms, and stylistic analysis of the Gospels for source detection. Much of the last fifty years of research has relied on the study of these agreements and J.C. Poirier has provided a very thorough survey of the main verbal agreement work in [31]. He however concludes his study with little hope regarding the feasibility of relying on verbal agreements to “find a final solution to the synoptic problem”, as the “statistical studies have too often amounted to coded expressions of their user’s commitments”.

Another possible approach relies on the analysis of the style of the Gospels, either using *Correspondance Analysis* as suggested by A. Linmans in [19] and D. Mealand in [23, 25, 24]. Notably, Mealand provides several different approaches to leverage stylistic analysis to provide insights regarding the likelihood of the 2SH: he uses in [23] *Correspondance Analysis* and *Discriminant analysis* to separate Lukan material into three sources (Q, Lukan and Markan) and adds cluster analysis and *Generalized Linear Models*, while accounting for genre (which was not the case in his Lukan study), to perform a similar analysis on Matthew’s Gospel. The author finds that the results of his study mostly confirms the 2SH as he is able to correctly classify material for both Gospels. Linmans is even more prudent, and concludes that when accounting for genre, sources can no longer be clearly distinguished and that the whole case still hangs in the balance [19], even when considering parallels between the Gospels. In spite of Poirier’s doubts, we expect that the newest advances in data science, able to capture more efficiently complex patterns in data, will bring some new lights into the synoptic problem.

Our contribution to the current state of the art of statistics and the Synoptic Gospels is three-folds : (1) we leverage more recent advances in data science and move from Mealand’s and Linmans Linear Discriminant Analysis [24, 19] to classifier two sample tests using Random Forests; (2) we work at the pericope level instead of at the verse level or sliding pools of words;

(3) we focus not only on stylometric analysis, but take into account the editing patterns of Luke and Matthew, which takes a step further the approach suggested by Linmans when he uses Principal Correspondance Analysis for parallels.

4. Methods

4.1. Dataset

Data from the *SBL Greek New Testament* (SBLGNT) edition and its morphological parsing and lemmatization, provided by MorphGNT [36], were used for the study. Given its wide reception and its on-going critical work, we assumed the SBLGNT text to be the "correct" one and did not consider the different variants, which will be included into some of our further works, where we hope to show the impact of considering different base texts on stylometry analysis. The verses were grouped into pericopes based on synoptic parallels suggested by the landmark work of K. Aland [2]. These pericopes were labeled as triple (Mt-Lk-Mk), double (Mt-Lk), Matthean, Lukan, or Marcan. There were a total of 276 distinct pericopes, with 613 different variations of these pericopes: 237 from Luke, 154 from Mark, and 222 from Matthew. The triple tradition consisted of 127 pericopes, and the double tradition has 50 pericopes. Additionally, we differentiated between narrative and sayings pericopes, as their genre can have a significant impact on style as underlined by Linmans [19], Mealand [24] and Oaks [28]: we manually classified the pericopes across *sayings* (including prophecy and parable) and *narratives* (including Passion material, controversy and miracle stories). We acknowledge that some of this labeling can be considered arbitrary as some of the parable contain large narrative materials (such as the *Parable of the Prodigal Son*, Lk 15:11-32), and further consideration regarding the impact of this classification will be included in future works.

4.2. Defining the stylistic features

To perform the stylistic comparison between the different pericopes, we decide to focus our study using only non-significant language patterns, instead of focusing on content word frequency with metrics like *tf-idf* computed on the corpus as a whole: as the pericopes relate the same stories, they naturally tend to use the same vocabulary which could bias a distance metric. To focus solely on style, we only keep features that do not bring any meaning to the sentence other than its grammatical and logical structure.

In total, 103 metrics were designed, based on recent advances in stylometry [11, 14, 33], that can be roughly divided into three categories: (1) *grammatical and morphological features*, which consists in computing the ratio of inflects, *Part Of Speech* (POS), verb conjugations, sentences length, punctuation, capitalization ...; (2) *function word frequencies*, consisting in conjunctions, generic words such as temporal markers ...; (3) *dialog features*, which consists in counting the number of occurrences of words indicating dialog, basing ourselves on the classification provided in [4]. The exhaustive list of all the computed metrics are available in table 6 of the appendices. The 103 features are then computed for each pericope, each pericope being projected into a 103-dimensional space.

Table 2

Examples of computations of stylistic differences on a subset of features for a triple tradition pericope (*The Healing of Peter’s Mother In Law*, pericope 37 of Kurt Aland’s classification, Mt 8:14–15, Mk 1:29–31, and Lk 4:38–39)

(a) Input				(b) Output			
Book	# Words	# Proper Nouns	# καί	Book	# Words	# Proper Nouns	# καί
Mt	33	2	6	Mt-Mk	-15.00	-3.00	-1.00
Mk	48	5	7	Mt-Lk	-8.00	0.00	3.00
Lk	41	2	3	Lk-Mk	-7.00	-3.00	-4.00

4.3. Computing stylistic difference

The stylistic difference between two pericopes is then defined as the difference between two stylometric vectors. The closer the style transformation, the smaller the differences in each variable, and vice versa. For pericopes from the triple tradition, there are two distances per pericope, and from the double, there is one distance per pericope. Single tradition material does not have any distance, as it has no reference point within other Gospels by definition. Stylistic difference vectors for pericope number k and feature i will be denoted as $\Delta^k(A - B)_i$, for $A, B \in \{Mt, Lk, Mk\}$, $A \neq B$. The gospel of Mark is used as a reference as both tested theory posits Markan priority.

For example, focusing only on feature 1 corresponding to the number of words in the pericope of *The Healing of Peter’s Mother In Law*, which is within the triple tradition (Mt 8:14–15, Mk 1:29–31, and Lk 4:38–39, pericope 37 according to Aland’s classification), Mark uses 48 words, while Luke uses 41 and Matthew uses 33: this results in a stylistic difference for feature 1 of $\Delta^{37}(Mt - Mk)_1 = -15$, $\Delta^{37}(Lk - Mk)_1 = -7$ and $\Delta^{37}(Mt - Lk)_1 = -8$. This difference is then computed for every of the 103 stylistic features and we give an example of the computation of these vectors in table 2: for example, we can see that when editing the *Mother In Law Pericope*, Luke and Matthew both decided to reduce the number of proper nouns to 2 and to shorten the number of words in the pericope. If similar editing choices are observed regularly, then this would give evidence that editing was not performed independently and vice versa.

Because the reduced space can get sparse (as some pericopes have had little change), we choose to apply *Principal Component Analysis* [18], selecting the number of components required to explain at least 90% of the variance of the dataset.

4.4. Using Random Forests classifiers for two-samples testing

Working with such a large feature set poses a challenge due to the multivariate nature of the stylistic vectors, which would require multivariate response models if we were to work using a classic statistical framework: a potential solution, as suggested by Mealand in [24], is to employ Generalized Linear Models with a multivariate target variable. However, Mealand’s decided to limit its model to only five dimensions. Given the extensive input space in our study, such models would not provide reliable results as we would have to compute one statistical test per feature, leading to hundreds of statistical test that would increase the risk of family-wise error rate.

To tackle this issue and still work in a multi-dimensional space, we use the approach suggested

by Lopez-Paz et al. in [20], who suggest using the prediction of classification of Machine Learning (ML) models as a two-samples hypothesis test in order to assess if two samples are from different distributions, with the following heuristic: if a ML model can discriminate more accurately between the observed samples than between randomly drawn ones, then the two classes must have different distributions. More formally, we compare two samples, P and Q. If the null hypothesis ($P = Q$) is true, a binary classifier's accuracy on a held-out subset should be around chance-level. Conversely, if the alternative hypothesis ($P \neq Q$) is true, the classifier's accuracy will be higher than chance-level, leading to rejection of H_0 . As our classifier, we use Random Forests [5], an ensemble-based learning method. Random Forests help mitigate overfitting in low-sample scenarios by training CART [6] trees independently on bootstrapped samples. This introduces randomness, preventing a single tree from dominating the ensemble. Final predictions are then based on the most frequent labels among the individual trees' predictions.

4.5. Tested hypotheses

To enhance the comprehensiveness of our study, we adopt a dual approach. We begin by examining the stylistic variations within the Lukan text between the double and triple traditions, aiming to identify distinct sources (as discussed in the next paragraph 4.5.1). We then propose a novel approach by investigating the interdependence among the edits made by different authors (as outlined in paragraph 4.5.2).

4.5.1. Source detection for stylometry

To decide which scenario is the most likely, one can inquire if the styles vary differently across the triple and the double material: if so, one could assume that Luke has taken his double and triple material from two different sources, like the 2SH posits. It would however not completely invalidate the FGH, as one could argue that Luke could have taken all of his triple tradition material from Mark, but would mean that we cannot observe any *editorial fatigue*² as argued by Goodacre in [12], where Luke style would match Matthew's instead of following Mark's.

To assess the stylometric difference across the sources, we design three different stylometric hypotheses to be evaluated using the test methodology described in subsection 4.4. These tests are named using the convention **Stylo.n**, $n \in 1, 2, 3$ and their alternative hypotheses and implications in the synoptic problem are summarized in table 3.

Stylo.0 measures style differences between sayings and narrative in the triple and double tradition of Luke's gospel to evaluate the relevance of our selected features. If we cannot detect significant stylometric changes, it implies our features may not be reliable for detecting subtle variations like distinct sources. **Stylo.1** tests the style difference between the double and triple traditions using a ML model. Successful classification would indicate that the style varies differently across the two traditions, supporting the idea that Luke sourced his double and triple material from separate origins, thus validating the 2SH. **Stylo.2** acts as a sanity check, considering the genre of the classified contents. The goal is to build a model that accurately classifies

²As defined by Goodacre in [13], "Editorial fatigue is a phenomenon that will inevitably occur when a writer is heavily dependent on another's work. In telling the same story as his predecessor, a writer makes changes in the early stages which he is unable to sustain throughout."

Table 3

Summary of performed tests for comparison of style

Test ID	Compared samples and size	Alternative hypothesis	Has Luke read Matthew ?
Stylo.0	Narrative pericopes in Luke (120 samples) vs Sayings pericopes in Luke (115 samples)	There is a strong difference in style between narrative and sayings in Luke.	N/A
Stylo.1	Double tradition material in Luke (50 samples) vs Triple tradition material in Luke (127 samples)	There is a strong difference in style between double and triple material in Luke.	No
Stylo.2	Double tradition sayings in Luke (45 samples) vs Triple tradition sayings in Luke (49 samples)	There is a strong difference in style between sayings in double and triple material in Luke.	No
Edit.1	Stylistic distance between Luke and Mark (127 samples) vs Stylistic distance between Matthew and Luke (127 samples)	Luke editing of Mark differs significantly from Matthew's for the triple tradition.	No
Edit.2.1	Stylistic distance between Luke and Mark (127 samples) vs Stylistic distance between Matthew and Mark (127 samples)	Luke is significantly closer in style to Mark than to Matthew on the triple tradition.	No
Edit.2.2	Stylistic distance between Luke and Mark on narrative pericopes (49 samples) vs Stylistic distance between Matthew and Luke on sayings pericopes (49 samples)	Luke editing of Mark differs significantly from Matthew's for the triple tradition (sayings)	No
Edit.2.3	Stylistic distance between Luke and Mark on narrative pericopes (78 samples) vs Stylistic distance between Matthew and Luke on narrative pericopes (78 samples)	Luke editing of Mark differs significantly from Matthew's for the triple tradition (narrative)	No

sayings common to Matthew and Luke, as well as those found in Mark, to ensure that the model does not oversimplify classification by labeling double tradition material as "sayings" and triple tradition material as "narratives."

4.5.2. Comparison of edits

The other complementary approach that we suggest in this study is the comparison of the editorial choices made by Luke and Matthew of Markan material: if these patterns are different, then one could assume that they were made independently. We design two different statistical tests relying on a two-samples classifier, **Edit.1** and **Edit.2**, that we summarize in table 3.

Edit.1 aims at comparing the distance between the triple tradition between the three Gospels in order to understand where Luke's triple tradition comes from. If we make the hypothesis that a smaller variation in style means a similar source, then the distance between Luke-Matthew

should be significantly closer than the distance between Luke-Mark over the triple tradition if Luke had access to Matthew. Otherwise, one can assume that Luke took his triple tradition from Mark, which would lead to weakening the FGH and be consistent with the 2SH, as Luke and Matthew are expected to perform their editing independently. **Edit.2** aims at measuring the independence of Matthew’s and Luke’s use of Mark: the idea is to see if it is possible to distinguish an editing pattern between sources. If we can find no significant difference from the way Luke edits Mark than the way Matthew edits Mark, one could assume that the edition process did not happen independently and Luke had access to Matthew’s triple tradition. These tests are then extended to narrative and sayings data, as sayings of Jesus could be expected to be less edited and are more likely to be taken *verbatim*.

4.6. Implementation

The Random Forest classifiers, as well as the PCA, is taken from sklearn [30]. We use 100 trees in the Random Forest and use the Gini criterion to measure the homogeneity of each split. We re-implemented the method described in [20] for the classifier two-samples tests. All our data and our pipeline to obtain these results are available at https://github.com/metz-theolab/chr_2023.

5. Results

The results of the stylometric tests (**Stylo.n**) and edit tests (**Edit.n**) are available in table 4.

Table 4
Result of stylometric source detection in Luke

ID	Alternative hypothesis	p-value
Stylo.0	Lukan material differs in style significantly between sayings and narrative contents	0.0
Stylo.1	Double and triple tradition differs in style significantly	0.03
Stylo.2	Double and triple tradition sayings differ significantly	0.05
Edit.1	Luke is significantly closer in style to Mark than to Matthew on the triple tradition.	0.02
Edit.2.1	Luke editing of Mark differs significantly from Matthew’s for the triple tradition.	0.07
Edit.2.2	Luke editing of Mark differs significantly from Matthew’s for the triple tradition (sayings).	0.45
Edit.2.3	Luke editing of Mark differs significantly from Matthew’s for the triple tradition (narrative).	0.19

5.1. Stylometric source detection in Luke

Stylo.0: Difference in style between narrative and sayings material in Luke The test results demonstrate a significant ability of our stylistic features to differentiate between narrative and sayings content within the entirety of Luke’s pericopes (p-value < 0.001). When employing a Random Forest classifier using 75% of the data as train and 25% as test, and stratification based on observed genre, the classifier achieves an accuracy score of 81% on unseen pericopes (95% over the complete dataset). These results indicate that our stylistic features effectively enable the classification of pericopes based on their style. When analyzing the

misclassified pericopes, they tend to fall along the genre borderlines, with the 6 misclassified sayings containing many narrative elements (*The Parable of the Rich Man and Lazarus, the Day of the Son of Man, On Riches and the Rewards of Discipleship, The Parable of the Good Samaritan, The Parable of the Wicked Husbandmen, The Pharisee and the Publican*) and the 7 misclassified narratives featuring discourses (*The Beelzebul Controversy, The Rich Young Man, John the Baptists's Messianic Preaching, The Pharisees Seek a Sign, First Preaching Tour of Galilee, Peter's Denial Prediction and The Third Prediction of the Passion*). This high accuracy reaffirms the validity and relevance of the devised features for the remainder of our study.

Stylo.1: Difference in style between double and triple tradition material in Luke The two-samples test conducted on the classifier reveals a highly significant distinction (p-value of 0.03) between the styles of the double and triple traditions. Utilizing a Random Forest classifier with a test-train split ratio of 0.25, we achieve an accuracy of 91% for unseen pericopes and 97% across the entire dataset, successfully identifying material from both the double and triple traditions. This disparity in style strongly supports the hypothesis that Luke drew his material from two distinct sources for the double and triple traditions, aligning with the 2SH.

While proponents of the Farrer's model may argue that Luke derived the double tradition material from Matthew and the triple tradition material from Mark, this interpretation does not align well with the concept of editorial fatigue proposed by Goodacre, as our tests indicate that Luke disregarded Matthew's style in his handling of the Matthean edition. However, the study of the confusion matrix available in table 5b shows that the unbalanced dataset affects the accuracy per class (there are 127 triple tradition pericopes and 50 double tradition). Further works of ours will include adding some techniques for small unbalanced dataset [26], to enhance the quality of our prediction model.

Stylo.2: Difference in style between double and triple tradition sayings in Luke As shown in the previous paragraph, the stylometric features differ strongly depending on the genre of the pericope, and we need to make sure that the results of the test **stylo.1** are not only a measure of the difference of style, as the double tradition is mostly composed of sayings. However, the results of test **Stylo.2**, with a p-value of 0.05, shows a significant style difference between sayings from the double and triple tradition, allowing us to confirm the stylistic difference between material, even when accounting for genre. A random forest using a test-train split ratio of 0.25 yields an accuracy score of 75% on unseen data and 92% of the whole dataset. The confusion matrix is available in table 5c and the small sample size once again affects the power of the evaluation of the classifier on test data: the error is however balanced across the double and the triple tradition.

5.2. Difference in edition behavior

Edit.1: Where did Luke take his triple tradition from ? The two-sample classifier tests reveal a statistically significant distinction (p-value of 0.02) between the stylistic similarity of Luke and Mark compared to Luke and Matthew. Luke and Mark exhibit a closer style (with a combined total distance of 50.46) than Luke and Matthew (with a combined total distance of 72.81). When a Random Forest classifier is trained using standard test/train fitting (with a 0.25

Table 5
Confusion matrix of **Stylo.n** tests

(a) Confusion matrix of Stylo.0			(b) Confusion matrix of Stylo.1			(c) Confusion matrix of Stylo.2		
Reality Predicted	Double	Triple	Reality Predicted	Double	Triple	Reality Predicted	Double	Triple
Double	21	6	Double	8	2	Double	9	3
Triple	5	27	Triple	2	33	Triple	3	9

ratio of test and train split), it achieves an accuracy of 79% when classifying unseen samples and a 95% accuracy across the whole dataset. The significant proximity observed between Luke and Mark, and the distance from Matthew to Luke, suggests that Luke most likely drew his triple material from Mark rather than Matthew. This finding once again contradicts the notion of editorial fatigue, as proposed by Goodacre, which would have resulted in a less pronounced difference in stylistic distances. We believe it is thus unlikely that Luke relied on Matthew’s text to copy his triple tradition material, which supports the plausibility of the 2SH.

Edit.2: Independence of Matthew’s and Luke’s use of Mark All statistical tests conducted to evaluate the variation between Matthew’s and Luke’s edits of Mark have yielded inconclusive results, with p-values greater than 0.05. However, it is worth noting that the p-value of Edit2.1 is relatively low at 0.07, suggesting a tendency to reject the hypothesis that Luke and Matthew edit Mark in a similar manner. This inclination leads us to consider the hypothesis that Luke disregards Matthew’s edits and implies that Luke did not consult Matthew’s use of the Markan material. When examining sayings or narrative material, our study encounters limitations in drawing definitive conclusions due to the combination of high p-values and low statistical power, particularly when the effect size is small. The available dataset consists of only 156 triple tradition narrative pericopes and 96 limited triple tradition sayings pericopes. Consequently, our test subset comprises merely 39 and 24 samples, respectively, considering a 0.25 test-train splitting ratio. Given these constraints and until further works consisting on using methods for unbalanced dataset, we are unable to confirm or dismiss the notion of a difference in editing style between Luke and Matthew: we have to rely on the other performed tests, which all point to the 2SH.

5.3. Discussion

The two facets of our analysis seem to point towards the 2SH: the stylometric tests show that the double and triple tradition differ significantly in terms of style, even when considering sayings (**Stylo.1** and **Stylo.2**). Even though one must remain careful on results on smaller datasets, such a difference seems to confirm that the double and triple material come from different sources, and thus that Luke took his triple material from Mark and not from Matthew. This use of Mark’s material by Luke is further confirmed by the distance stylistic tests (**Edit.1**), which shows that Luke’s style is significantly closer to Mark’s than to Matthew’s. Whenever comparing the editing style, the test **Edit.2** tends to show that Matthew and Luke

use the Markan material differently, even though the test p-value and power is not sufficient to definitely conclude regarding the difference in material edition. We believe our features aggregated on the pericope level might be too global and we will investigate in further works more refined features, such as the verbal agreements, the addition and the deletion orders and use Deep Learning embedding models such as LSTM [16].

Whenever considering previous studies, our results generally align with results from other statistical studies of the synoptic Gospels: Mealand demonstrated (prudently) a difference in word usage on Matthew's Gospel into a Matthean, double, and triple tradition, and the Linear Discriminant Analysis he runs on Luke's Gospel also shows a division in double and triple tradition. As there is to our knowledge no stylometric study of the edition of patterns between Matthew and Luke, we cannot compare our results to previous studies, but future works will compare the feature importance of the Random Forests to the manual works performed by scholars to characterize Matthew and Luke's handling of Markan material.

6. Conclusion and further works

As a conclusion, our work adds yet another argument towards the likelihood of the 2SH model. We however want to emphasize that our intention is not to offer a definitive answer to the synoptic problem, but rather to introduce innovative analytical tools that can assist in presenting the most plausible solution to the problem through style analysis, even though numerous variables other than style, such as pericope orders, must be considered to provide a comprehensive answer.

Our further studies will include adding additional features better able to characterize source usage, such as verbal agreements, and see how it affects the results of the edition tests. We will also work on augmenting the number of samples, either by simulation through bootstrapping or by considering moving windows across the different verses, to increase the statistical power of our study. We will also take into account variants instead of taking the SBLGNT as the final text of the Gospels.

7. Acknowledgements

This work was funded by the SCRIBES (Biblissima+ Grant) and the ANR SHERBET (CE38-2023) projects.

References

- [1] A. Abakuks. *The Synoptic Problem and Statistics*. Chapman & Hall, 2015.
- [2] K. Aland. *Synopsis Quattuor Evangeliorum: Locis parallelis evangeliorum apocryphorum et patrum adhibitibus edidit*. Württembergische Bibelanstalt Stuttgart, 1964.
- [3] Augustine. *The Harmony of the Gospels (De Consensu evangelistarum)*. Kessinger Publishing, 2015.

- [4] F. Baskevitch. *Le vocabulaire du son en grec ancien - étude sémantique et étymologique*. 2016.
- [5] L. Breiman. "Random Forests". In: *Machine Learning* 45(1) (2001), pp. 5–32.
- [6] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.
- [7] C. Carlston and D. Norlin. "Once More – Statistics and Q". In: *Harvard Theological Review* 64(01) (1971), pp. 59–78.
- [8] J. D. Dunn. *The Oral Gospel Tradition*. Wm. B. Eerdmans Publishing Co., 2013.
- [9] W. Farmer. *The Synoptic Problem: A Critical Analysis*. Mercer University Press, 1976.
- [10] A. Farrer. *On Dispensing With Q*. Blackwell, 1955, pp. 55–88.
- [11] R. Forsyth. "Feature-Finding for Text Classification". In: *Literary and Linguistic Computing* 11 (1996), pp. 163–174.
- [12] M. Goodacre. *The Case Against Q: Studies in Markan Priority and the Synoptic Problem*. Trinity Press International, 2002.
- [13] M. Goodacre. "Fatigue in the Synoptics". In: *New Testament Studies* 44 (1998), pp. 45–58.
- [14] J. Grieve. "Quantitative Authorship Attribution: An Evaluation of Techniques". In: *Literary and Linguistic Computing* 22(3) (2007), pp. 251–270.
- [15] J. Hawkins. *Horae Synopticae: Contributions to the Study of the Synoptic Problem*. Clarendon Press, 1899.
- [16] S. Hochreiter and J. Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (1997), pp. 1735–80.
- [17] A. Honoré. "A Statistical Study of the Synoptic Problem". In: *Novum Testamentum* 10 (2/3) (1968), pp. 5–147.
- [18] I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, 2002.
- [19] A. Linmans. "Correspondence Analysis of the Synoptic Gospels". In: *Literary and Linguistic Computing* 13(1) (1998), pp. 1–13.
- [20] D. Lopez-Paz and M. Oquab. "Revisiting classifier two-sample tests". In: *International Conference on Learning Representations (ICLR)* (2017).
- [21] S. Mattila. "Negotiating the Clouds around Statistics and "Q": A Rejoinder and Independent Analysis". In: *Novum Testamentum* 16(2) (2004), pp. 105–131.
- [22] A. McNicol and D. Dungan. *Beyond the Q Impasse: Luke's Use of Matthew : A Demonstration by the Research Team of the International Institute for Gospel Studies*. Bloomsbury, 1996.
- [23] D. Mealand. "Correspondence Analysis of Luke". In: *Literary and Linguistic Computing* 10(3) (1995), pp. 171–182.
- [24] D. Mealand. "Is there Stylometric Evidence for Q?" In: *New Testament Studies* 57(4) (2011), pp. 483–507.

- [25] D. Mealand. "Measuring Genre Differences in Mark with Correspondence Analysis". In: *Literary and Linguistic Computing* 12(4) (1997), pp. 227–245.
- [26] S. Narwane and S. Sawarkar. "Machine Learning and Class Imbalance: A Literature Survey". In: *Industrial Engineering Journal* 12 (2019).
- [27] J. O'Rourke. "Some Observations on the Synoptic Problem and the Use of Statistical Procedures". In: *Novum Testamentum* 16(4) (1974), pp. 272–277.
- [28] M. Oakes. *Literary Detective Work on the Computer*. John Benjamin Publishing Company, 2014.
- [29] B. Orchard and T. Longstaff. *J. J. Griesbach: Synoptic and Text - Critical Studies 1776--1976*. Vol. 34. SNTS Monograph Series (Cambridge University Press), 1978.
- [30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [31] J. Poirier. "Statistical Studies of the Verbal Agreements and their Impact on the Synoptic Problem". In: *Currents in Biblical Research* 7(1) (2008), pp. 68–123.
- [32] T. R. Rosché. "The Words of Jesus and the Future of the 'Q' Hypothesis". In: *Journal of Biblical Literature* 79.3 (1960), pp. 210–220.
- [33] J. Savoy. *Machine Learning Methods for Stylometry*. Springer, 2020.
- [34] A. Schweitzer. *The Quest of the Historical Jesus*. Dover Publications, 2005. 416 pp.
- [35] B. H. Streeter. *The Four Gospels: A Study of Origins, Treating of the Manuscript Tradition, Sources, Authorship & Dates*. Macmillan and Co., 1930.
- [36] J. K. Tauber. *MorphGNT: SBLGNT Edition. Version 6.12 [Data set]*. 2017. URL: <https://github.com/morphgnt/sblgnt>.
- [37] C. Weisse. *Die evangelische Geschichte kritisch und philosophisch bearbeitet*. Breitkopf und Hartel, 1838.

Table 6
103 computed features

(a) Grammatical features	(b) Words with computed frequencies
pos N	πατήρ
number S	κύριος
case G	ἀμήν
case N	Χριστός
gender F	θεός
gender M	ἑαυτοῦ
avg word length	ἄν
unique word ratio	ἀλλά
stop words ratio	ἀπό
is dialog	ἄρα
capitalized words	δέ
substring comma	δή
substring dot	διά
substring median	ἔτι
substring kai	ἐγώ
avg len comma	ἐκ
avg len dot	ἐν
avg len median	ἐπί
nbr words	εἰ
pos RA	γάρ
gender N	ἦ
pos D	καί
pos P	κατά
pos C	μέν
pos RP	μετά
pos RR	μή
pos V	ὁ
pos A	ὅδε
person 3	ὅτι
person 2	οὔτε
tense X	οὖν
tense A	οὐ
active M	οὐδέ
active P	περί
active A	σύ
mode P	σύν
mode I	ὑπέρ
mode S	ὑπό
mode N	ᾧ
number P	ὥστε
case D	ἐάν
case V	παρά
case A	πάλιν
pos RD	αὐτός
person 1	υἱός
tense P	ἡμέρα
tense I	σήμερον
pos RI	αὔριον
pos X	Ἰησοῦς
tense F	
mode D	
mode O	
tense Y	
pos I	