Automatic source extraction of the Synoptic Gospels

Statistics and the synoptics

Dr Sophie Robert-Hayek - Pr Frédérique Rey

University of Lorraine
Écritures - *Maison des Sciences de l'Homme – Lorraine*

2023 Society of Biblical Literature

Introduction

Introduction

- Huge daily transformation because of the application of **statistical tools to address various challenges and problems**.
- Has led to their application to **biblical studies to solve various issues**: computational stemmatology, scribal detection, authorship attribution...

The required rigor of the synoptic problem lends itself to using mathematical tools to provide answers.

Solages 1959:8

"The [synoptic] problem lends itself perfectly to a mathematical calculus ...which, if it succeeds, will have ...the advantage of a great objectivity."

State of the art: the synoptic problem and statistics

The possible approaches to using statistics to understand the relationship between the gospels

Ever since the recent advances in computer science and statistical tools, **studies give insights regarding the possible interrelation of the Gospel using statistics**.

Roughly divided into:

- Study of **the verbal agreements** and their distribution across the gospels.
- Study of the distribution of **lexicometric** and **stylistic features** of the text.

Verbal agreements

Verbal agreements

Verbal agreements are *"the use in two (or three) of the synoptic gospels of the same grammatical form of the same word"* (Honoré 1968).

- **Pre-suppose the Q hypothesis** and confirm/infirm (Rosché 1960; Mattila 2004; O'Rourke 1974...);
- **Do not suppose the Q hypothesis** (Honoré 1968; Carlston and Norlin 1971; Bergemann 1993..);

Severe disagreements on the agreements..:

- Should the synonyms be taken into account ?
- Should the words be inflicted/conjugated ?
- Identical in forms and/or sequence ?

Back and forth controversy regarding verbal annotation agreements: Carlston and Norlin 1971; Mattila 2004...

Verbal agreements and Q

Survey of some studies on Q and their conclusion:

Study	**Conclusion**
Honoré 1968	Markan priority; Extra saying source.
Rosché 1960	Q as a saying source exists but was not written
Morgenthaler 1971	Luke knew Q and Matthew
Carlston and Norlin 1971	Q as a saying and written source exists
Bergemann 1993	The principal source for the Sermon on the Mount/Plain is not Q, but an aramaic tradition
Ronning 1989	Mark is the middleman of a linear stemma (Farrer's, Augustinian)
...	..

Verbal agreements

No clear-cut interpretation of the obtained data... in spite of the great hopes of the 50s.

The results too strongly depend on the definition of verbal agreements and what constitutes double and triple tradition.

Poirier 2008

Landmark study gloomily concludes: "*The prospect that the use of* ***word statistics*** *would provide* ***an objective measure for the study of gospel interrelations*** *has often been held out with* ***an unrealistic hope*** *[...] having too often amounted* ***to coded expressions*** *of their user's commitments.*"

Stylometric analysis

Other possible approaches: take into account the **stylometry** of the gospel as:

- Stylometric changes can indicate **different sources**;
- Stylometric similarities across the different gospels can indicate **relationships between the gospels**.

Mealand 2011

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.*

Stylometric analysis

Roughly, separated into two possible approaches:

- **Supervised approach**: 2ST is pre-supposed and treated as such using statistical tests (Mealand 2011);
- **Unsupervised approach**: No model pre-supposed, data is analyzed and/or visualized, and then compared to the tagged data (Mealand 1997; Mealand 1995; Linmans 1998; Mealand 2011).

Mealand 2011 tries both and concludes regarding the existence of Q using stylistic analysis by **analyzing Matthew only**.

Limit of the studies

Limits of verbal agreements: Poirier seems right, relying on verbal agreements encodes pre-supposed theories.

Limits of existing stylometric analysis:

- Comprehensive analysis of Matthew's gospel only;
- Stop words analysis instead of whole range of possible speech features;

Our contribution

Contribution of this study

- Stylometric analysis of **Luke's Jesus logia** using 325 features (instead of most frequent words)
- Working at **discourse level** (instead of a gliding window).

Our approach: source extraction using clustering

Research question

Research question

Does the style used by Luke's Jesus differs throughout Luke's Gospel ?

Methodology: dataset

- The used dataset is the eclectic SBLGNT text (**a limit of our work which will be addressed into further work**)
- The text has been manually annotated:
  - Speech extraction through discourse delimators;
  - Genre annotation (aphorism, parable, narrative, controversy, prophecy), with only **aphorism** and **parables** retained out, using Linmans' classification;
  - Corresponding Q reference using RHK's critical edition.

Methodology: dataset

Stylometry takes into account:

- **Part of Speech**
- **Gender**: Masculine, Feminine, Neuter;
- **Case**: Nominative, Accusative, Genitive, Dative;
- **Number**: Singular, Plural;
- **Mood**: Indicative, Imperative, Subjunctive, Optative, Infinitive, Participle;
- **Tense**: Present, Imperfect, Future, Aorist, Perfect, Pluperfect
- **Voice**: Active, Middle, Passive;
- **Stop words** (and their followed inflection)

Methodology: embedding the dataset into a numerical space

To use **numerical methods**, the dataset needs to be projected into a **numerical space**.
To do so, we compute the *frequency* of each **grammatical occurrence**, in a 1, 2, 3 gram fashion:

$$\mathrm{F}(t, d) = \frac{\text{Number of occurrences } d}{\text{Total number of terms} d}$$

Methodology: embedding the dataset into a numerical space

Example:

- **Original sentence**: τί ὅτι ἐζητεῖτε με οὐ ἤδειτε ὅτι ἐν τοῖς τοῦ πατρός μου δεῖ εἶναι.
- **Part Of Speech**: Interrogative-Pronoun Conjunction Verb Personal-pronoun Adverb Verb Conjunction Preposition Definite-article Definite-article Noun Personal-pronoun Verb Verb

	Count	**POS frequency**
Conjunction	2	0.15
Verb	4	0.30
Conjunction-Preposition	1	0.08
Conjunction-Verb	1	0.08
...		...

Methodology: embedding the dataset into a numerical space

To account for rarer occurrences, multiply each *Frequency* by its *Inverse Document Frequency*:

$$IDF(t, D) = \log\left(\frac{\text{Total number of sayings}}{\text{Number of sayings containing the term } t}\right)$$

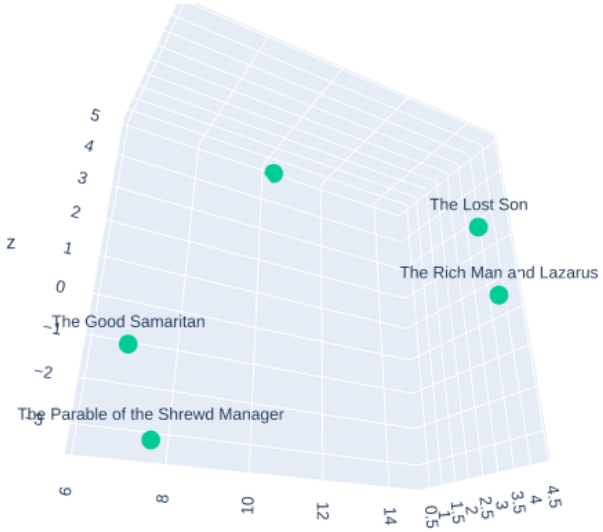
Rare n-grams will have a high TF-IDF, while common words will have low TF-IDF.

Total of 325 stylometric features.

Methodology: embedding the dataset into a numerical space

Dataset can now be projected into a numerical space, **each saying corresponding to a vector**, then **dimensions are reduced using Principal Component Analysis**.

Projection's of Jesus' parabola in 3D:



Performing clustering

Clustering consists in grouping together points that are close (according to a metric) into a numerical space:

Each saying of Jesus located into a single cluster (= group) corresponding to its **closest** sayings, in terms of **style** (*stylometric analysis*).

Performing clustering

We use:

- **Agglomerative Hierarchical Clustering**: recursively group together closest sayings.
- **Manhattan similarity**: measures the sum of the absolute difference between the vectors.

Experiment plan

Experiment 1: *On parables and aphorism*:

1. Two automatic clustering on:
   - Parables (see Libby 2015 on the importance of genre separation in authorship attribution);
   - Aphorisms;
2. Post-processing by comparison and discussion with RHK's Q (Robinson 2000).

Experiment 2: *On Q material*:

1. Clustering on Lukan Q material;
2. Post-processing by comparison and discussion with RHK's Q.

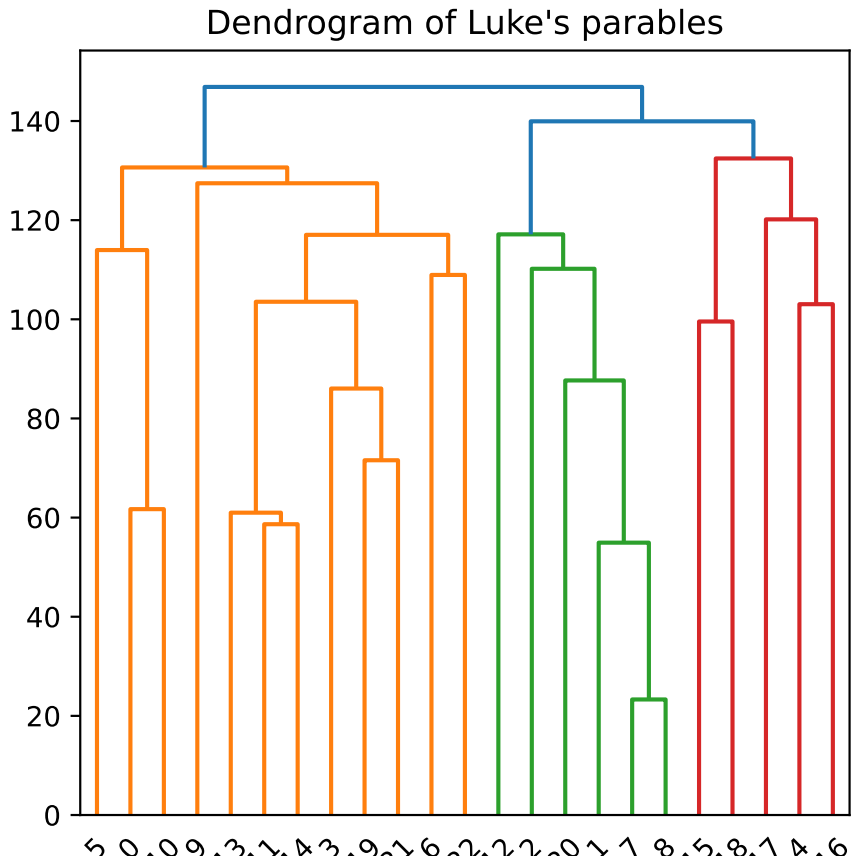
Comprehensive study on Luke's and Matthew's in full paper as this is already quite dense!

Results

Clustering parables

Clustering results

Study of clustering shows **3 different stylometric behavior of parable stylometry**:
Cluster 0 (red), cluster 1 (orange), cluster 2 (green).



Comparison to Robinson's Q

Cluster	In Robinson's Q	Percent
0	No	100.0
1	No	58.3
	?	25.0
	Yes	16.7
2	No	33.3
	Yes	66.7

- **Cluster 0**: 100% cluster of longer Lukan material (*Good Samaritan, Prodigal Son, The Shrewd Manager, The Rich Man and Lazarus, The Unfair Judge*);
- **Cluster 1**: Complex mix of traditions that requires more investigation, mostly belonging to Q and the Triple tradition.
- **Cluster 2**: Mostly Q's parable, with two additional Lukan traditions, limit case of parable classification (*Forgiving debt*, Lk 7:41-42, *Building a tower*, Lk 14:28-32)

Conclusion regarding parables

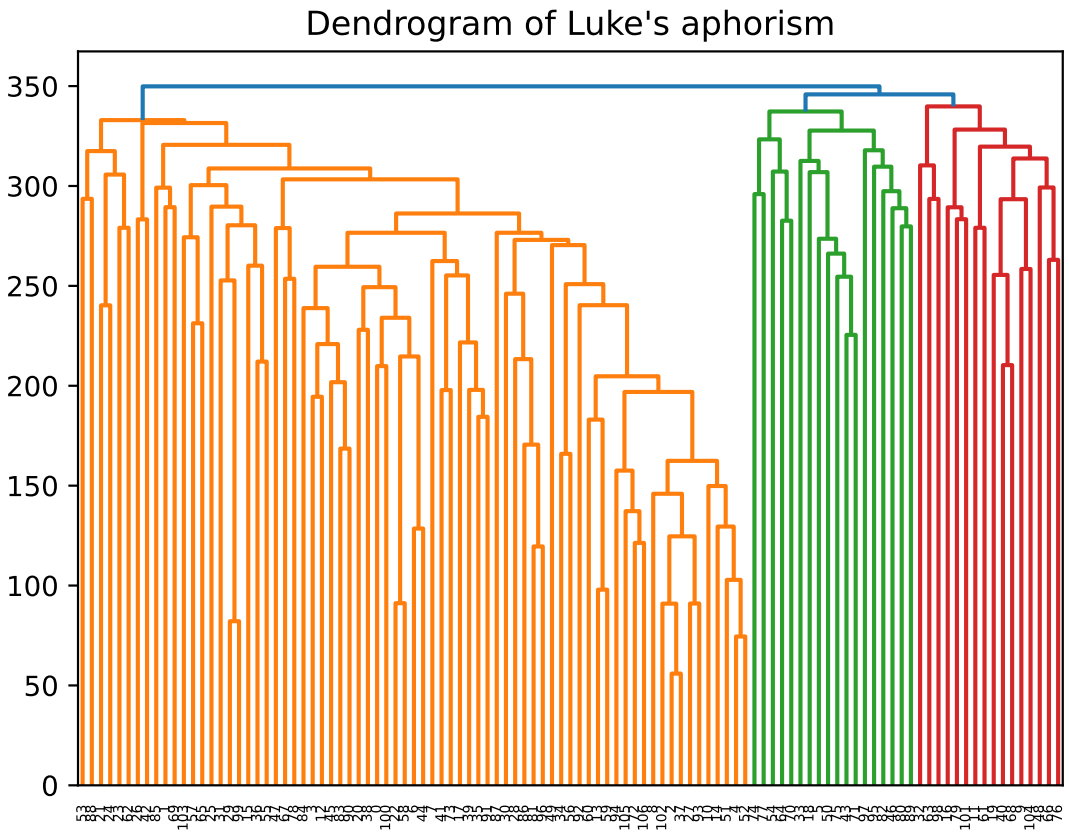
Interpretation:

- Separate Lukan style for longer passages compared to shorter Lukan parable;
- Statistical distribution seems to show a slight Q style;
- Very similar style/genre affects cluster's separability: similar style because of redactional processing even if 2ST.

Clustering aphorisms

Aphorism clustering results

No clear clusters... 3 ?:



Aphorism clustering results

No easy distinction between Q and Triple tradition when it comes to aphorism styles...

Cluster	In Q	Percent
0	False	0.8
	True	0.2
1	False	0.7
	True	0.3
2	False	0.8
	True	0.2

Clustering such large datasets is often too difficult to interprete.

Global interpretation of clustering

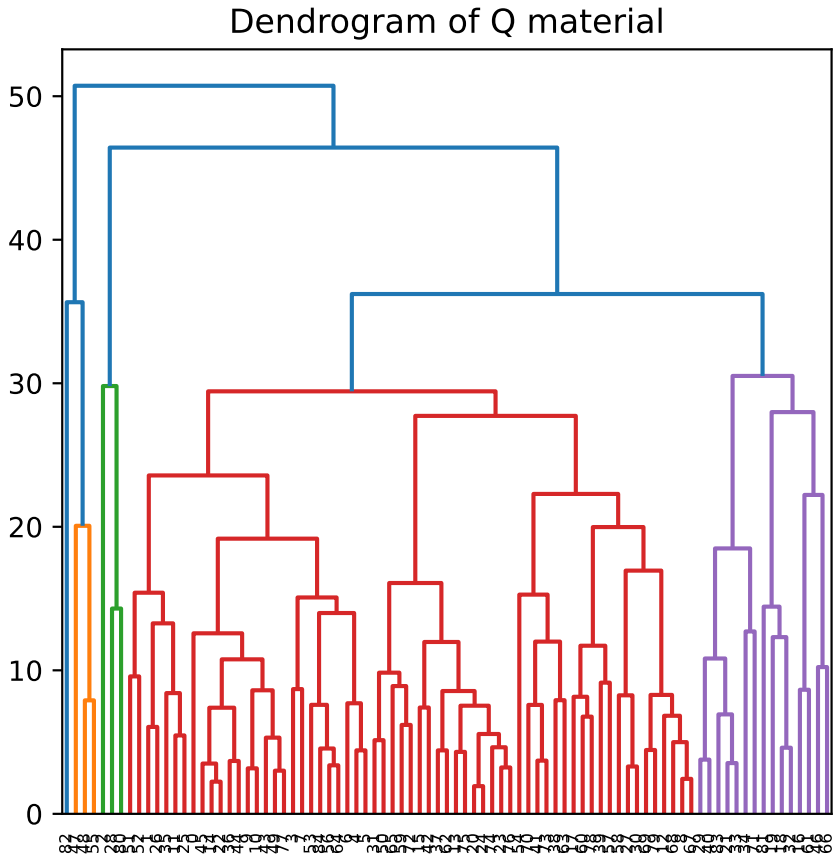
Interpretation:

- Longer Lukan material comes from a different source;
- No significant style changes within double and triple material;
- We are not able to confirm or infirm the 2ST using stylometry analysis, **different results than Mealand 2011 on Matthew !**
- ... but clustering on so much data can lead to difficult to interprete results.

Zooming in on Q material within Luke

Q clustering results

Rather homogeneous Q sayings edited by **Luke** ...3 clusters ?



Comparison to RHK's analysis

A large homogeneous Q cluster with **some notable disputed Q sayings marked as outliers**:

Sermon on the Mount sayings:

- All in the same cluster;
- except ~~*Woes against the Rich*~~ (Q6:24-26) and *Renouncing Ones Own Rights* (Q6:29);

Q 11

- All in the same cluster, including *[Looting a Strong Person]* (Q11:21);
- except *The Lord's Prayer* (Q11:2); *The friend at midnight* (Q11:5-8) (genre effect ?);

Comparison to RHK's analysis

Q 12

- All in the same cluster including ~~*Fleeing amond the Towns of Israel*~~;
- except *Children Against Parents* (Q12:49); *Not Fearing the Body's Death* (Q12:4-5); ~~*The Rich Fool*~~ (Q12:16-21).

Q19:12-27

The Entrusted Money as an outlier: probably linked to genre or to the disputed content.

Interpretation:

- Lukan sayings from Q are relatively homogeneous in style;
- Except for some notable outliers, confirming doubts/non inclusion of RHK.

Conclusion and further works

Conclusion

- Computational analysis takes into account data correlation that are impossible for a human mind;
- **Too large clustering renders result interpretation complicated** and clustering tasks should be small to be useful;
- Adds some **further arguments concerning the inclusion/exclusion of some saying material in Q**;

Automatic analysis of the Synoptic Gospels **can bring valuable information regarding Gospels compositions** and further studies should be performed to fully leverage results.

Further works

Expected further works:

- Lexicographic analysis: grouping together sayings according to their vocabulary;
- Perform direct analysis on RHK's Q text to analyze several styles/vocabulary within the Q source;
- Take variants into account instead of eclectic text;
- Add insights by adding the analysis of John's gospel;

Thank you for your attention !

Any questions ?

My e-mail : sophie.robert@univ-lorraine.fr

Bibliography I

- Bergemann, Thomas (1993). *Q auf dem Prüfstand: Die Zuordnung des Mt/Lk-Stoffes zu Q am Beispiel der Bergpredigt*. FRLANT, 158; Göttingen: Vandenhoeck Ruprecht.
- Carlston, C. and D. Norlin (1971). "Once More — Statistics and Q". In: *Harvard Theological Review* 64(01), pp. 59–78.
- Honoré, A.M. (1968). "A Statistical Study of the Synoptic Problem". In: *Novum Testamentum* 10 (2/3), pp. 5–147.
- Libby, James (2015). "Disentangling Authorship and Genre in the Greek New Testament: History, Method, and Praxis". PhD thesis. McMaster Divinity College.
- Linmans, A. (1998). "Correspondence Analysis of the Synoptic Gospels". In: *Literary and Linguistic Computing* 13(1), pp. 1–13.

Bibliography II

- Mattila, S. (2004). "Negotiating the Clouds around Statistics and ■Q■: A Rejoinder and Independent Analysis". In: *Novum Testamentum* 16(2), pp. 105–131.
- Mealand, D. (1995). "Correspondence Analysis of Luke". In: *Literary and Linguistic Computing* 10(3), pp. 171–182.
- — (1997). "Measuring Genre Differences in Mark with Correspondence Analysis". In: *Literary and Linguistic Computing* 12(4), pp. 227–245.
- — (2011). "Is there Stylometric Evidence for Q?" In: *New Testament Studies* 57(4), pp. 483–507.
- Morgenthaler, Robert (1971). *Statistische Synopse*. Ed. by Stuttgart: Gotthelf-Verlag.

Bibliography III

- O'Rourke, J. (1974). "Some Observations on the Synoptic Problem and the Use of Statistical Procedures". In: *Novum Testamentum* 16(4), pp. 272–277.
- Poirier, J. (2008). "Statistical Studies of the Verbal Agreements and their Impact on the Synoptic Problem". In: *Currents in Biblical Research* 7(1), pp. 68–123.
- Robinson, J. et al. (2000). *The Critical Edition of Q*. 1517 Media.
- Ronning, Halvor (1989). "Word Statistics and the Minor Agreements of the Synoptic Gospel". In: *Actes du Second Colloque International Bible et Informatiques: Méthodes, Outils, Résultats*.
- Rosché, Theodore R. (Sept. 1960). "The Words of Jesus and the Future of the 'Q' Hypothesis". In: *Journal of Biblical Literature* 79.3, pp. 210–220. ISSN: 0021-9231. DOI: 10.2307/3263927.