A Meta-Study on Contemporary Computational Analysis of the Biblical Texts: Textual Criticism

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Ethan Stokey
Spring, 2025

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Rosanne Vrugtman, Department of Computer Science
Briana Morrison, Department of Computer Science

A Meta-Study on Contemporary Computational Analysis of the Biblical Texts: Textual Criticism

CS4991 Capstone Report, 2025

Ethan Stokey
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
gac8sb@virginia.edu

ABSTRACT
Textual criticism is advancing by integrating computational methodologies into the scholarly process. In this meta study, I explore the use of clustering, Bayesian phylogenetic models, machine learning, and recursive translation analysis as methods in textual criticism. These computational approaches demonstrated a minimization of bias and were able to trace manuscript relationships more objectively. They showed great promise in accurately reconstructing original Biblical texts to a high degree. From this examination, I have concluded that further work can be conducted by expanding manuscript data samples and unifying the computational methods into a single framework. This integration will reduce bias, enhance textual reconstruction accuracy, and further advance the role of technology in Biblical manuscript analysis. These advancements could significantly enhance the accuracy, speed, and accessibility of biblical textual criticism for scholars worldwide.

1. INTRODUCTION

In Biblical textual criticism and manuscript analysis, scholars aim to identify variations between Biblical manuscripts across the centuries. This is done to reconstruct the most accurate version of the original text, understand the transmission history of the documents and uncover intentional or accidental changes made by scribes over time.

With the use of computers in the field, more computer science systems are being applied to the discipline of textual criticism. Using computational tools such as natural language processing (NLP), machine learning algorithms, image recognition software, etc., simplifies the process of comparing manuscripts, identifying anomalies and classifying data, making it more effective than traditional manual methods.

2. RELATED WORKS

In the reconstruction of the New Testament through, the prolific New Testament scholar Ehrman (1995) noted that certain readings of the New Testament are elevated in priority due to certain criteria. These include how many manuscripts support a textual tradition, how the oldest manuscripts read, the geographic spread of manuscript readings and the textual families into which manuscripts are grouped by their similarity [1]. That final criterion is one of the most important in determining where and when a manuscript likely originated.

In 2013, an experiment was conducted to determine similarities and differences between ancient Greek manuscripts of an apocryphal work known as the *Protoevangelion of James*. Using

computational methods such as k-means, hierarchical clustering and correspondence analysis, these manuscripts were grouped by each reading's characteristics, then compared with the traditional textual families to which modern scholars ascribe each manuscript. Although the majority of manuscripts were correctly correlated with each other, some manuscripts were distanced from their known families through these methods [2].

3. ADVANCEMENTS IN TECHNOLOGICAL TEXTUAL CRITICISM

In a recent 2023 study, a new analytical approach was utilized, combining Bayesian probability theory and phylogenetic trees. The latter are commonly used in biology to construct relations between animals to show common ancestors. Various manuscripts across different textual families were given prior probabilities of textual trustworthiness based on general New Testament scholarship confidence of its contents being authorial. In addition, the Lewis Mk substitution model was applied to handle the transcriptional behavior that variation units have in their arbitrary number of states (versions of reading).

Researchers used TEI XML to craft the tree of manuscript data collations, which was then converted through the teiphy python package into BEAST XML. The BEAST 2 program incorporated the Baesyian analysis parameters set upon the data, including a clock model that displayed a tree branch's length as a unit of time in accordance with a manuscript's likely date of composition [3]. Using this methodology, human bias was supplanted with pure quantitative probabilities, prior and post of the execution of the models. Using a sample collation of the Epistle to the Ephesians, the researchers showed the feasibility of use of these methods in contemporary textual criticism.

One of the more recent studies pertaining to technology's influence in the field of textual criticism came alongside the advent of AI. As a test of its capabilities in the field, a team of researchers used machine learning to distinguish between the most likely readings of verses and any variant readings. They used six Greek manuscripts of the first three chapters of the Gospel of John as a starting point. In comparison to modern critical New Testament Greek texts, the accuracy and precision rate at correctly identifying the likely renderings of the readings was very high [4]. However, it is clear that the study was limited in its data sample size. With over 5,800 Greek manuscripts, not including over 20,000 across multiple languages before the printing press, a large data sample size like that would build an incredible foundation for easier textual criticism and manuscript analysis.

Another method for increasing efficacy within textual criticism is through proper translation analysis. This is done by recursively observing differences between major textual families, identifying whether they are personal scribal preferences for writing or actual critical variants, and then rechecking those conclusions and resolving uncertainties. In a 2020 study, the Old Testament book of Jonah had its Masoretic Text reading compared with that of the Greek Septuagint. Through this process, a translator's personal writing style could be distinguished with the actual variants of the text so scholars could focus on that latter area [5].

4. EXPECTED OUTCOMES

The Bayesian phylogenetic model visualizes the distance in time and relation between manuscripts. By showcasing relationships as a tree, one can see the "fathers" of each manuscript and their "brothers." Subjectivity is always present in the manual work of

textual criticism, so applying prior probabilities to remove unnecessary bias is a welcome advantage that scholars can use to better showcase manuscript relationships. Alongside this, through machine learning, scholars can focus on certain prioritized criteria in deciding which text is likely authorial in textual criticism. Finally, recursive translation analysis enables scholars to refine their understanding of textual shifts, distinguishing genuine textual variants from a translator's personal influence. As these technologies continue to develop, they have the potential to revolutionize textual criticism by minimizing human bias.

5. CONCLUSION

It is demonstrably possible to incorporate computer science applicative techniques to the aim of reconstructing a text from its varying manuscripts. By applying techniques such as Bayesian probability modeling, machine learning and recursive translation analysis, scholars can identify textual variants, manuscript relationships and scribal behaviors with far greater precision. These techniques not only streamline the immense task of comparing thousands of ancient manuscripts but also help reduce subjective bias, bringing the discipline closer to reconstructing the most likely original text of the Bible. Overall, this push to further bring technology into the fold of textual criticism is looking promising, and these works detailing the fascinating complexities of incorporating technology within textual criticism disciplines inspired much of this paper.

6. FUTURE WORK

To expand upon the research and its results, future work should seek to increase the size of the dataset of manuscripts used for computational analysis. While some studies have yielded impressive results, such as the machine learning model applied to six Greek manuscripts of John's Gospel, the limited sample size restricts broader applicability. Scaling up these studies to incorporate a much larger and more diverse corpus would not only increase the robustness of the findings but also allow for more refined insights into the development of textual traditions over time and geography.

Future work should also focus on integrating the various computational approaches explored in this meta-study into a unified model. A composite framework that combines Bayesian phylogenetic analysis, machine learning classification and recursive translation analysis could yield a more comprehensive system for evaluating manuscript relationships, variant readings and translator influence. By bringing these tools together, scholars could gain a more holistic view of textual transmission, while also benefiting from the strengths of each method.

REFERENCES

[1] Bart Ehrman (1995) 'Textual Criticism of the New Testament', in *Hearing the New Testament: Strategies for Interpretation*. W.B. Eerdmans Publishing Company, pp. 127–145.

[2] Eddie Dunn, Curry Guinn, and George Zervos (2013) 'Proceedings on the International Conference on Artificial Intelligence (ICAI)', in. Athens, Greece: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), pp. 1–7.

[3] Joey McCollum and Robert Turnbull (2023) 'Using Bayesian Phylogenetics to Infer Manuscript Transmission History', *Digital Scholarship in the Humanities*, 39(1), pp. 258–279. doi:10.1093/llc/fqad089.

[4] Mason Jones, Francesco Romano, and Abidalrahman Mohd (2022) 'Machine learning in textual criticism', *2022 7th*

International Conference on Machine Learning Technologies (ICMLT), pp. 1–5. doi:10.1145/3529399.3529400.
[5] Eric Tully (2020) 'A model for distinguishing between textual variants and translation shifts in Old Testament textual criticism', *Scandinavian Journal of the Old Testament*, 34(2), pp. 245–266. doi:10.1080/09018328.2020.1805207.