university of groningen / faculty of science and engineering / mathematics and applied mathematics

# Computational Stemmatology

MSc Applied Mathematics Project Proposal

**Keywords**: Computational Stemmatology, Natural Language Processing, Data Science

**Starting Date**: 16th October 2023

**Midpoint Date**: 16th March 2024

**Final Date**: 12th July 2024

**Student**: Darren Zammit s5284236

**First Supervisor**: Dr Julian Koellermeier

**Second Supervisor**:

# 1 A Brief Introduction to Computational Stemmatology

Computational stemmatology is a discipline wherein one aims to reconstruct the genealogy (stemma) of a text on the basis of differences between surviving manuscripts (witnesses) [1]. Over time a text is copied which usually results in changes. One can analyse the differences between different witnesses to construct the stemma. Philologists construct stemmata to study the errors made and how scribes influenced a text which may be useful to study the evolution of cultures, values, and ideas.

Graphical modelling takes a central role in constructing stemmata [1]. Contemporary methods – which are primarily analogous to those in phylogenetics, the study of the evolutionary history of species – can be split into distance-based [2], parsimony-based [3], and statistical methods [4]. In these methods every word in the text is equally important. In reality there are many different kinds of errors and changes, some more significant than others, leading to the algorithm failing to faithfully construct a stemma. For example, spelling mistakes or changing symbols should have far less impact on a script than adding or redacting text. Another issue which these methods struggle to address is that of contamination, that is when a text is copied from multiple manuscripts [1]. If different scripts were used for different chapters then this could lead to outputs which faithfully reflect the stemma of most chapters but not all if different chapters were copied from different manuscripts.

Since Lee's seminal paper [5] in applying computational phylogenetics models in stemmatology multiple advances in computational stemmatology have been made. However, most papers in computational stemmatology are authored by computer scientists and philologists with mathematicians and statisticians representing a small minority of authors [1].

# 2 Research Aims

The first phase of the project will consist of a literature review of modern methods in computational stemmatology and phylogenetics. The aim is to extend existing methods or adapt novel methods from phylogenetics and test them for accuracy and robustness against contamination, missing strings and missing manuscripts on artificial data sets and texts for which the stemma is already known such as in [6]. Due to the similarity of the tasks in stemmatology and phylogenetics, advances in one field can usually be adapted and applied in the other. The resulting stemmata will be compared to the outputs of commonly used phylogenetic programs such as PAUP [7], Phylip [8] and Mr Bayes [9] which have been used as standard benchmarks in computational stemmatology [1]. This can be achieved by converting the textual data into genetic data which can be fed directly into the aforementioned programs [1].

In the final phase of the project we plan on working with actual philologists to generate stemmata using tailored models for previously unstudied texts. Some of the texts considered were impenetrable even to manual stemmatological methods and thus this would be the first time a stemma is produced for these manuscripts.

# 3 Timeline

We plan on starting the project on the 16th of October 2023. The student shall have 25 credits of extra courses in 2023/24. 5 credits in term 1, 10 in term 2, 5 in term 3 and the Research Seminar in Applied Mathematics. Thus, the midterm date will be taken as the 16th of March and the end date will be the 12th of July.

# References


[1] Roelli, Philipp. (2020). Handbook of stemmatology : history, methodology, digital approaches. Berlin, Germany: De Gruyter.

[2] M. Spencer and C. J. Howe, "Estimating distances between manuscripts based on copying errors," Literary and Linguistic Computing, vol. 16, p. 467–484, 2001.

[3] W. M. Fitch, "Toward defining the course of evolution: minimum change for a specific tree topology," Systematic Biology, vol. 20, p. 406–416, 1971.

[4] Q. Nguyen and T. Roos, "Likelihood-based inference of phylogenetic networks from sequence data by phylodag," in International Conference on Algorithms for Computational Biology, 2015.

[5] A. Lee, 'Numerical Taxonomy Revisited: John Griffith, Cladistic Analysis and St. Augustine's Quaestiones in Heptateuchum.' Studia Patristica XX: 24-32, 1989.

[6] T. Roos, T. Heikkilä, Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets, Literary and Linguistic Computing, Volume 24, Issue 4, December 2009, Pages 417–433, https://doi.org/10.1093/llc/fqp002.

[7] D. L. Swofford, PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts. 2003.

[8] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005.

[9] F. Ronquist, and J. P. Huelsenbeck. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574. 2003.