Computational Literary Genre Stylistics

Julius-Maximilians-
UNIVERSITÄT WÜRZBURG
Lehrstuhl für Computerphilologie (Department for Literary Computing)

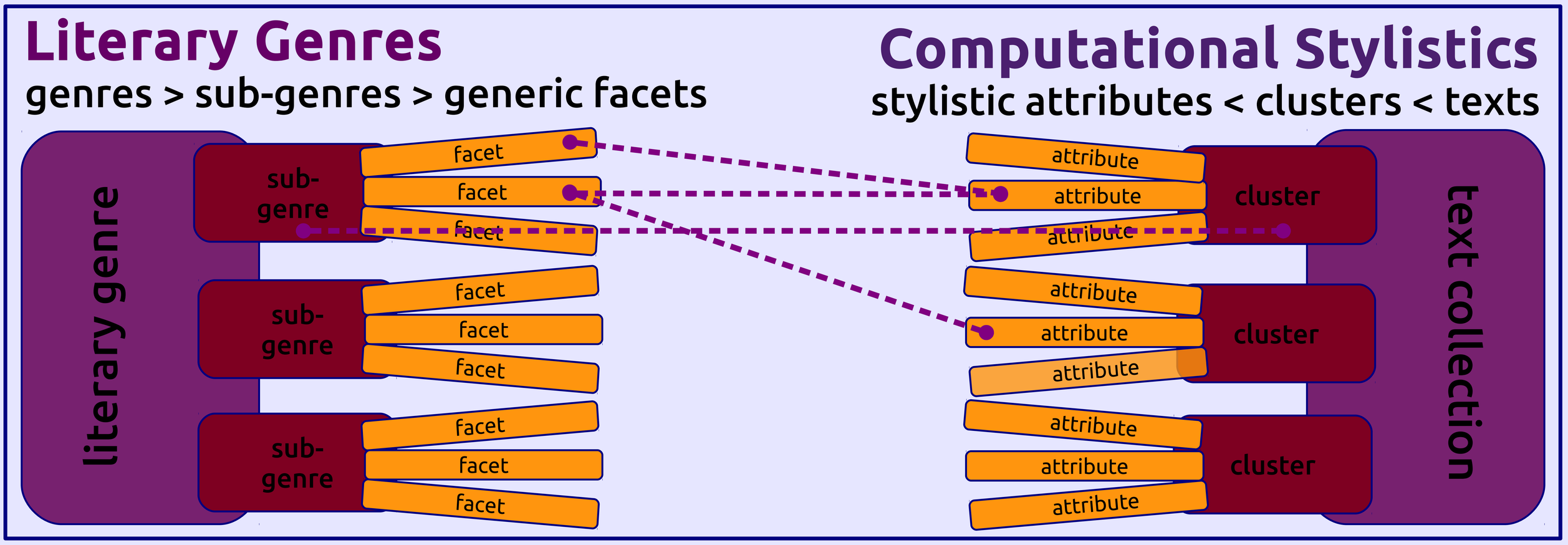
Project lead: Dr. Christof Schöch – http://go.uni-wuerzburg.de/schoech

The project in a nutshell

This project will investigate the relation between literary genres (generic facets) and style (stylistic attributes). Analyses are based on large collections of literary texts and use or develop state-of-the-art methods in computational stylistics and text analysis. This will enable the project to combine attention to minute stylistic details with the investigation of large trends and patterns in literary history.

What does the project aim to accomplish?

The project brings together scholars from Romance Philology and Computer Scientists, building a common ground in Computational Philology. It aims to change the way we think about the concepts of genre and style as well as about the relation between literary interpretation and computation. In particular, the project aims to strengthen the engagement with digital methods in Romance Philology.

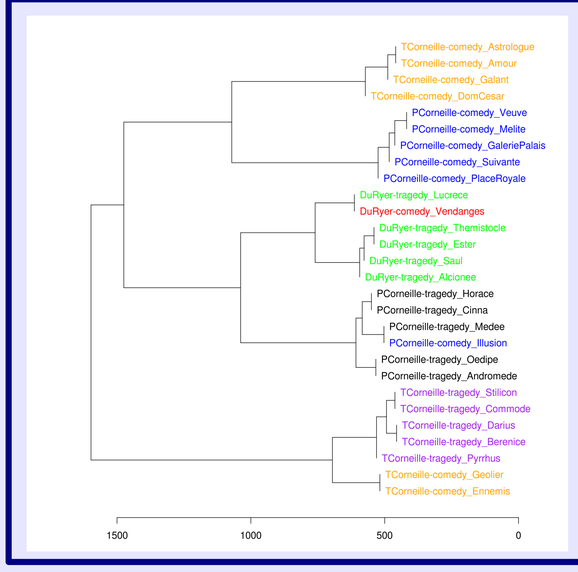


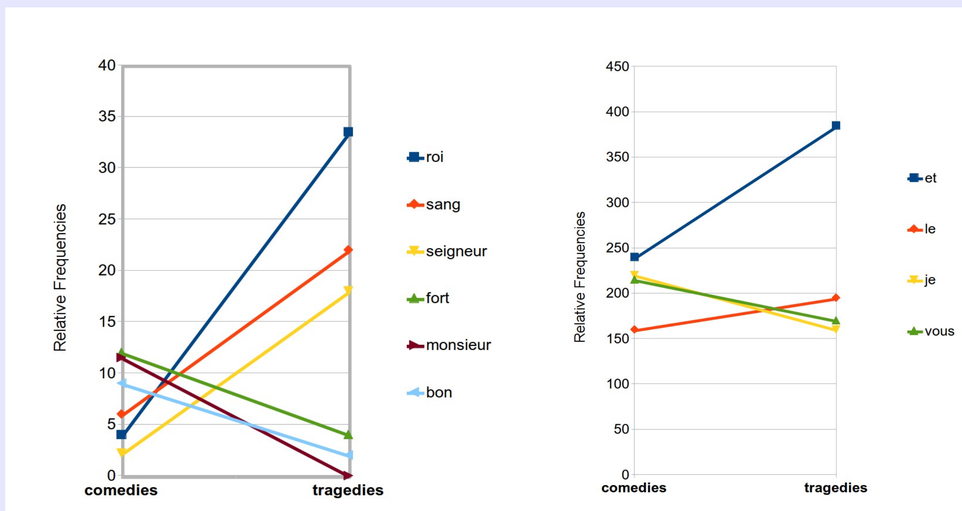
Which research questions are being pursued?

- How can the relation between style and genre be conceptualized in a productive way?
- On what stylistic basis can differentiations of genre be made?
- Which automatically identifiable, linguistic features, on which linguistic levels, are indicators of genre?
- Which relations exist between genres and sub-genres?
- How do generic styles and author-dependent styles relate to each other?

Which digital research methods are being used?

- Linguistic preprocessing using machine learning to add stylistically relevant annotations to the texts.
- The most relevant analytic methods are cluster analysis, principal component analysis, topic modeling, etc.
- Visualization techniques are used heuristically to make patterns and trends in data discoverable.
- Validation is performed regarding statistical validity and robustness and regarding correlations with existing knowledge from literary history.





Example #1: Comedies and tragedies by three authors

- Left: Cluster Analysis based on 1000 most frequent words (Eder's Delta).
- Right: Distinctive content words and distinctive function words (Voyant)

Example #2: Comedies in prose and verse by five authors

- Left: Principal Component Analysis based on 100 most frequent words (stylo)
- Right: Proportions of the form (blue), author (green) and date (red) signals in PC1 and PC2 for 5-200 MFW.