The Consortium Profiling Premodern Authors (PROPREAU) applies and develops new flexible machine learning-based tools for the analysis and classification of texts in order to explore several fundamental and unresolved questions of authorship in classical and medieval Latin texts. Despite the cultural importance of Latin, many essential texts remain anonymous. This is largely so because of the highly conservative nature of Latin based culture, characterized by imitation of earlier authors and quoting excerpts of their texts. Therefore, authorship attribution requires an analysis and comparison of large quantities of text. PROPREAU incorporates machine learning-based tools developed at the Turku NLP Group (IT Department) into the conventional argumentation of the humanities, allowing a much wider look at textual material than is attainable by single scholars using conventional methods.

The expected results of the Consortium are new, well-grounded answers to questions of authorship that were previously considered unsolvable. PROPREAU will provide guidelines and new computational methods for future endeavors to identify anonymous premodern texts. The subproject at the Department of Cultural History addresses several significant aspects of the Latin literary culture: the identification of the authors of Latin grammatical texts and their relations, the recognition of documents written in the Papal chancery, and the identification of late-medieval polemical treatises. The subproject of the IT Department will create novel approaches that address the specific challenges of the domain through task-oriented feature selection and training, and the proposed approaches will further be implemented as software released under open licenses.
MATERIALS AND METHODS

Computer-assisted authorship attribution benefits from a large background corpus. Individual case-studies of the project draw their background corpora from the abundance of texts available online, such as the Corpus Corporum, one of the most extensive open-access corpora of Latin. We supplement the existing corpora by scanning and preprocessing texts not yet available in machine-readable form, above all late-medieval literary works. In machine-learning-based computational based author attribution, PROPREAU uses two classifiers, (1) a simple linear classifier, Support Vector Machine and (2) Convolutional Neural Network, which represents state of the art in text classification. Within individual case-studies, the machine-learning methods are developed further to answer to particular problems, such as attribution through the exclusion of potential authors. On the other hand, when possible the classifications are based on large corpora with up to 15 000 000 words.

RECENT FINDINGS

In authorship attribution tasks, our preliminary results demonstrate that the classifiers give accurate results based on word uni- and bigrams and character 5-grams (classifications run both with all but 1000 most common words masked, and without masking). At present, we have submitted for peer-review an article on the attributions of the Ciceronian corpus. On-going case-studies of the project concern the authorship of works attributed to Augustine of Hippo, Pope Gregory I, 14th-century inquisitor Petrus Zwicker, and King Henry VIII. In addition to individual attributions, the preliminary results demonstrate the benefits of large corpora of millions of words. Both the Ciceronian and Augustinian corpora show that in addition to initial hypothesis formulated by the human researchers, the classification of a large corpus reveals anomalies that point towards possible misattributions in the oeuvre of the classified authors. The project proposes a new text to be attributed to both Marcus Tullius Cicero and Augustine of Hippo, both major classical authors.

06/2018