Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Dr. Katrin Dennerlein
Institut für Deutsche Philologie
Philosophische Fakultät I

Ausschreibung für eine studentische Hilfskraft für mind. fünf (gerne bis zu 20) Stunden pro Woche) ab sofort für zunächst 4 Monate (mit der Option auf eine langfristige Zusammenarbeit)

Extraktion von Strukturen aus Dramen für die quantitative Dramenanalyse

Die automatische Analyse von Texten ist derzeit ein wichtiges Thema in den Digital Humanities. Wir benötigen Ihre Kenntnisse für die Vorbereitung von Dramen für die Analyse, für die automatische Extraktion von Konfigurationen und für die mathematische Auswertung der Ergebnisse. Ihre Aufgabe besteht darin, in Dramen Strukturen wie Akt- und Szenenzählungen, Figurenauftritte und -abgänge, Beiseitsprechen etc. automatisch zu erkennen. Die Kennzeichnung der entsprechenden Ereignisse in verschiedenen Dramen variiert einerseits, andererseits ist unser Korpus zu groß um alle Dramen zu lesen und die Ereignisse manuell auszuzeichnen. Sie sollen deshalb ein Programm weiterentwickeln, das die entsprechenden Ereignisse möglichst gut erkennt. Da nicht davon auszugehen ist, dass alles eindeutig erkennbar ist, sollte das Programm Konsistenzprüfungen durchführen (z.B. eine Person muss aufgetreten sein, bevor sie spricht) und Warnungen für nicht klar erkannte Strukturen ausgeben. Einen Eindruck von Strukturen der TEI-Ausgangsdaten können Sie sich im TextGrid-Repository machen, etwa am Beispiel der TEI-Fassung von *Minna von Barnhelm*. (http://textgridrep.de/services/tgcrud-public/rest/textgrid:rjh9.0/data)

Bisher wurde mit einem kleinen Trainingskorpus http://scikit-learn.org der Algorithmus Random Forest auf die Erkennung von Informationen in Bühnenanweisungen trainiert. Dies ist im NLTK-Trainer geschehen (http://nltk-trainer.readthedocs.org/en/latest/). Alle bisherigen Ergebnisse können Sie hier einsehen: https://github.com/pouyana/teireader/tree/ner . Sie werden von Ihrem Vorgänger eingearbeitet.

Literaturwissenschaftliche Vorkenntnisse sind nicht nötig, Python-Kenntnisse und die Bereitschaft mit maschinellem Lernen zu arbeiten, dagegen schon. Software und Dokumentation sollen weiterhin unter einer Open-Source-Lizenz stehen.

Kontakt: Dr. Katrin Dennerlein, katrin.dennerlein@uni-wuerzburg.de
Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte
Institut für Deutsche Philologie
Am Hubland
97074 Würzburg