„Kontrastive Analyse literarischer Texte mit Zeta“. Einführung in die Implementierung und Evaluation von Distinktivitätsmaßen

Keli Du (Universität Trier), Julia Dudar (Universität Trier), Cora Rok (Universität Heidelberg), Christof Schöch (Universität Trier)

In den ‚Computational Literary Studies‘ finden eine Reihe quantitativer Methoden Anwendung, die für die vergleichende Textanalyse herangezogen werden. Dazu gehören insbesondere statistische Distinktivitätsmaße, die „charakteristische“ Elemente (Lemmata, Wortarten usw.) einer Gruppe von Texten, bspw. von Romanen eines bestimmten Autors, aus einer bestimmten Epoche oder einer bestimmten Gattung, im Vergleich mit einer anderen Textgruppe herausstellen. Ziel unseres Projekts „Zeta und Konsorten“ ist es, durch die Implementierung und Evaluation verschiedener Distinktivitätsmaße zu einem tieferen Verständnis ihrer Funktionsweise zu gelangen und ihren Nutzen für die (digitale) Literaturwissenschaft zu beschreiben.

Im Rahmen der vDHd 2021 möchten wir einen zweiteiligen Workshop während der ersten „Eventtage“ im März und der zweiten Event-Phase im September anbieten, in dem wir verschiedene Distinktivitätsmaße vorstellen und erklären, wie sich zwei Textgruppen miteinander vergleichen lassen. Die geplante Dauer des ersten Teils des Workshops inklusive Diskussionsphasen beträgt 90 Minuten, des zweiten Teils circa 1 Std.

Im Zuge des Workshops soll zunächst unsere Untersuchung eines literarischen Korpus präsentiert werden; mithilfe von pyzeta wurden insgesamt 160 französische Romane der 1980er Jahre aus vier verschiedenen Gattungen (Hochliteratur, Krimi, Science-Fiction und Romance) verglichen. Die ausgegebenen Listen enthalten Wörter, die für die jeweils 40 Romane einer bestimmten Gattung im Vergleich zu den Romanen der anderen drei Gattungen charakteristisch sind. Die Wortlisten werden zur besseren Verständlichkeit aus dem Französischen ins Englische übersetzt. Anschließend wird die Python-Implementierung des „Zeta“-Maßes („pyzeta“) vorgestellt, das die gleichmäßige Verteilung charakteristischer Merkmale innerhalb einer Textgruppe im Vergleich zu einer anderen Textgruppe berechnet. Den Teilnehmenden werden statistische Grundbegriffe vermittelt und die notwendigen Instrumente an die Hand gegeben, um unter Anleitung eine vergleichende Analyse zweier Korpora mittels Pyzeta durchzuführen. Ein Textkorpus sowie der Programmcode werden zur Verfügung gestellt, eine Installation wird durch den Einsatz von Google Colab nicht notwendig sein. Grundkenntnisse in Python sind zwar erwünscht, werden aber nicht vorausgesetzt, da das Skript für Google-Colab so angepasst wird, dass die Teilnehmer nur Parameter wie „Gattung“ oder „Einheit“ (Lemma, Wortart) einstellen müssen.

Um die Ergebnisse dieser Untersuchung validieren zu können sowie die Stärken von pyzeta (und später von anderen Maßen) zu erfassen, benötigen wir zuverlässige Evaluationsstrategien. Eine Option dafür ist die menschliche Evaluation. Um diese Evaluationsmöglichkeit experimentell zu testen, möchten wir die Workshop-Teilnehmenden darum bitten, im Anschluss an den Workshop einen Online-Fragebogen auszufüllen. In diesem Fragebogen werden diverse, zufällig ausgewählte Sätze aus unserem Korpus aufgelistet. Die Teilnehmenden werden gebeten, jeden Satz einer der vier literarischen Gattungen zuzuordnen und außerdem anzugeben, mithilfe welcher Wörter sie die Gattung bestimmt haben. Die Anweisungen zum Ausfüllen des Fragebogens werden am Ende des Workshops gegeben. Für die Teilnehmenden, die kein Französisch sprechen, werden wir einen Fragebo-

gen mit englischen Übersetzungen der Sätze anbieten. Die Teilnehmenden haben dann zwei Wochen Zeit, den Fragebogen auszufüllen. Mit dem Fragebogen-Experiment möchten wir die Beteiligten aktiv in die Entwicklung und Ausarbeitung unserer Evaluationsstrategie involvieren. Die Wörter, die von den Teilnehmenden als entscheidend für die Gattungsbestimmung angegeben wurden, werden von uns mit den Wörtern verglichen, die wir bei der Korpusanalyse mit pyzeta gewonnen haben. Auf diese Weise möchten wir feststellen, in welchem Ausmaß pyzeta in der Lage ist, die distinktiven Wörter einer bestimmten Gattung zu identifizieren.

Bei einem weiteren einstündigen Treffen während der zweiten „Eventtage“ im September möchten wir den Workshop-Teilnehmern die Ergebnisse der menschlichen Evaluation sowie den Vergleich mit den Ergebnissen von pyzeta präsentieren und diese mit ihnen diskutieren.

Anm.: Die vDHd 2021 ist eine durch die Community des Verbandes Digital Humanities im deutschsprachigen Raum organisierte virtuelle Konferenz. Die Abstracts haben kein Peer-Review-Verfahren durchlaufen.