Direkte Rede im französischen Roman: Automatische Erkennung und gattungsabhängige Verteilungen

Direkte Rede im französischen Roman: Automatische Erkennung und gattungsabhängige Verteilungen Die zunehmende Digitalisierung des kulturellen Erbes unter anderem durch Bibliotheken verändert die Rahmenbedingungen für literaturwissenschaftliche Forschung. Wenn immer mehr literarische Texte in digitaler Form verfügbar sind, rücken derzeit noch periphere, computergestützte Methoden näher ans Zentrum literaturwissenschaftlicher Arbeit (vgl. Ramsay 2011). Zugleich kann dadurch nicht nur ein Kernbestand kanonischer und repräsentativ gesetzer Werke berücksichtigt werden, sondern eine viel größere Bandbreite auch weniger kanonischer Werke. Dadurch verändert sich unser Bild einer literarischen Gattung oder Epoche. Der vorliegende Beitrag möchte die Chancen und Herausforderungen dieses digitalen Paradigmenwechsels am Beispiel der automatischen Erkennung direkter Rede in französischen Romanen des 19. Jahrhunderts aufzeigen (vgl. zu deutschen Erzähltexten, Brunner 2015). Die hier verwendete Sammlung besteht aus 127 Romanen, die zwischen 1840 und 1889 erschienen und verschiedenen Untergattungen zugeordnet sind.

Direkte Rede ist in französischen Romanen nicht einheitlich durch öffnende und schließende Anführungszeichen gekennzeichnet. Der Beginn ist zwar häufig mittels eines Gedankenstrichs am Zeilenanfang markiert, das Ende hingegen ist nicht besonders hervorgehoben und inquit-Formeln sind meist nur durch Kommata von der direkten Rede abgegrenzt (vgl. Berthelot 2001). Für einen Menschen ist es dennoch leicht, in einem Roman direkte Rede anhand typographischer, semantischer und kontextueller Informationen zu erkennen. Diese Information händisch zu annotieren ist aber sehr zeitaufwändig. Für einen Computer ist diese Aufgabe dagegen zunächst eine Herausforderung, wenn es aber erst einmal funktioniert, können fast beliebig große Textmengen bearbeitet werden.

Durch die Methode des maschinellen Lernens können beide Ansätze verbunden werden (vgl. Han et al. 2011). Hier wird in einem ersten Schritt eine händische Annotation eines kleinen Teils der Sätze in der Textsammlung (direkte Rede oder nicht) vorgenommen. Außerdem werden die Texte mit Werkzeugen aus dem Natural Language Processing automatisch linguistisch annotiert (u.a. Wortarten, Tempus und Modus, Verbtypen, Satzlänge, Interpunktion). Auf dieser Grundlage kann ein Algorithmus diejenigen Merkmale erkennen, die typisch für Sätze mit direkter Rede oder ohne direkte Rede sind und auch weitere, nicht händisch annotierte Sätze entsprechend markieren.

Auf der Grundlage von 81 verschiedenen Merkmalen und mit einem “Random Forest” genannten Algorithmus konnten wir eine sehr zufriedenstellende Erkennungsqualität erreichen: knapp 94% der Sätze wurden korrekt klassifiziert (F-Score: 0.939). Insgesamt ergibt sich für das gesamte

Romankorpus, dass durchschnittlich 61% aller Sätze direkte Rede enthalten. Die Daten lassen auch Aussagen darüber zu, wie sich die direkte Rede in den verschiedenen Untergattungen oder über die Zeit hinweg verteilt. Es zeigt sich, dass die Automatisierung einer scheinbar trivialen Aufgabe doch einigen algorithmischen Aufwand erfordert, um sie mit zufriedenstellender Präzision umzusetzen. Im Gegenzug eröffnet dies aber eine neue Sichtweise auf die Literaturgeschichte, denn es ist nun möglich, grundlegende formale Eigenschaften literarischer Texte nicht nur exemplarisch zu analysieren, sondern ihre Entwicklung und Verteilung auf der Grundlage einer umfassenderen Berücksichtigung der literarischen Tradition zu beschreiben.

Zugleich ist das hier Vorgestellte nur ein erster Schritt zu einer breiteren Erfassung narrativer Techniken. Weil es nun möglich ist, automatisch die Erzählerrede von der Figurenrede zu trennen, könnten in einem nächsten Schritt die unterschiedlichen Formen der Erzählerrede (u.a.: narrativ, deskriptiv, argumentativ) oder auch unterschiedliche Erzählperspektiven (auto-, homo-, heterodiegetisch) in den Fokus der automatischen Erkennung rücken und deren Entwicklung ebenfalls datengestützt beschrieben werden.

Literatur

Durrer, Sylvie. (1994) Le dialogue romanesque. Style et structure. Genf: Droz.

Brunner, Annelen (2015). Automatische Erkennung von Redewiedergabe. Ein Beitrag zur quantitativen Narratologie. Berlin: De Gruyter.

Han, Jiawei et al. (2012). “Classification: Basic Concepts”, in: Data Mining: Concepts and Techniques. Burlington, MA: Elsevier, 327-392.

Stefanie Popp, Daniel Schlör, Christof Schöch, José Calvo Tello, Ulrike Henny

Würzburg, Nachwuchsgruppe CLiGS