Poster

Big, complex, heterogeneous.. Laufende Projekte aus dem Arbeitsbereich *Big Data in den Geisteswissenschaften* in DARIAH-DE

Stefan Pernes, Uni Würzburg

Der Begriff *Big Data* wird in den unterschiedlichsten Kontexten gebraucht, er umfasst unterschiedliche Größenordnungen und Strategien der Datenverarbeitung, und kann aufgrund dieser heuristischen Schwäche im besten Fall als *Buzzword*, mit Sicherheit aber nicht als trennscharfes Konzept bezeichnet werden. Zu Recht wird die Aufmerksamkeit zunehmend auf Fragen der Reliabilität und Validität der Verfahren gelenkt (Jordan 2014) und übergroße Heilsversprechen sowie Ankündigungen eines *Endes der Theorie* (Anderson 2008) kritisch hinterfragt. In geisteswissenschaftlichen Kontexten rückt diese allgemein geführte Diskussion jedoch in den Hintergrund. Hier gilt es, Textbestände in einer zuvor nicht da gewesenen Größe unter Berücksichtigung ihrer Vielschichtigkeit und Heterogenität zu verwalten und festzustellen, welchen Beitrag quantitative Methoden zu hermeneutischen Interpretationsverfahren leisten können. Diese Spezifika führen auch dazu, dass einige Voraussetzungen erst geschaffen werden müssen; so zum Beispiel das Training bestehender Verfahren der Sprachverarbeitung für literarische Textsorten, das Erstellen spezifischer Korpora und Vokabulare, oder die Verbesserung der Texterkennung von mittelalterlichen Handschriften. Das sind einige der Aufgaben, zu denen die *Use Cases* des DARIAH-DE Clusters *Big Data in den Geisteswissenschaften* einen Beitrag leisten und die im Folgenden vorgestellt werden sollen. Die *Use Cases* bearbeiten Fragestellungen aus Literaturwissenschaft, Philologie und Geschichte und werden jeweils in Kooperation eines fachwissenschaftlichen Partners und eines Partners aus dem Bereich der angewandten Informatik durchgeführt.

Narrative Techniken und Untergattungen im Deutschen Roman

Lehrstuhl für Computerphilologie, Uni Würzburg / Ubiquitous Knowledge Processing Lab, TU Darmstadt

Fachwissenschaftlicher Gegenstand des *Use Case* ist es, anhand quantitativer Verfahren die historische Entwicklung narrativer Techniken und in weiterer Folge auch die Entwicklung darauf aufbauender literarischer Kategorien nachzuvollziehen. Die Textgrundlage bildet ein Korpus 2000 deutschsprachiger Romane aus dem Zeitraum von 1500 bis 1930 und eine Sammlung von 200 französischen Kriminalromanen aus dem 19. und 20. Jahrhundert. Zum Einsatz kommen Verfahren zur automatischen Erkennung bestimmter Merkmale, wie zum Beispiel der Erkennung von Eigennamen oder von Passagen direkter Rede. Sämtliche Merkmale werden im Anschluss als *Features* zueinander in Bezug gesetzt, um die Texte

zu gruppieren und Gattungsbegriffe nachprüfen zu können. Parallel dazu werden Lernmaterialien erstellt, welche die dabei entwickelten Arbeitsabläufe in Form allgemeinverständlicher *Rezepte* zugänglich machen und interessierte ForscherInnen dazu ermächtigen sollen, *state of the art* Werkzeuge der Sprachverarbeitung für ihre jeweiligen Forschungsvorhaben einzusetzen und auf ihre jeweiligen Daten anzupassen.

Identifikation grenzübergreifender Lebensläufe in nationalen Biografien

Leibniz-Institut für Europäische Geschichte Mainz / Lehrstuhl für Medieninformatik, Uni Bamberg

Der *Use Case* erforscht die Verbindungen von individuellen historischen Lebensläufen und Internationalitätskriterien auf Grundlage von *Wikipedia* und mehreren europäischen Nationalbiografien. Dabei werden verschiedene Merkmale von Mobilität - wie zum Beispiel Geburts-, Wirkungs- und Sterbeorte, Tätigkeiten und verwandtschaftliche Beziehungen - miteinander korreliert und durch eine gezielte Erhebung sämtlicher Zusammenhänge mitunter Beobachtungen gemacht, die in den Geschichtswissenschaften noch nicht theoretisch erfasst sind. Die Datengrundlage umfasst strukturierte Daten sowie unstrukturierte Texte in mehreren Sprachen die miteinander verschränkt werden. Zusätzlich zum fachwissenschaftlichen Erkenntnisgewinn, stellt das Vorhaben eine quantitative Grundlage für kontrollierte Vokabulare in der Biografieforschung dar und zeigt auf, welche inhaltlichen und formalen Kategorien für *Semantic Web*-Ansätze in der Biografieforschung erforderlich und nutzbringend sein können.

Spuren der Zitation und Wiederverwendung im OpenMigne Korpus

Lehrstuhl für Digital Humanities, Uni Leipzig / Lehrstuhl für Medieninformatik, Uni Bamberg

Ausgehend von Editionen der Texte frühchristlicher Kirchenväter durch Jacques Paul Migne im 19. Jahrhundert entwickelt der *Use Case* Verfahren zur Erschließung vollständiger diachroner *Zitationsspuren.* Es handelt sich dabei um ein Feststellen chronologisch nachvollziehbarer Verläufe in einem Netzwerk von Zitationen, welches sich über ein gesamtes Korpus spannt. Textgrundlage bildet das *OpenMigne* Korpus, dessen Texte in griechischer und lateinischer Sprache einen Zeitraum vom Ursprung des Christentums bis in das 15. Jahrhundert abdecken. Die technische Umsetzung verläuft schrittweise: Beginnend mit der Erkennung von Zitationen in direkter Rede und in gleichsprachigen Ursprungstexten werden die Verfahren dahingehend erweitert, dass auch eine Erkennung von Paraphrasierungen und Zitationen in sprachlich heterogenen Korpora möglich wird. Die entwickelten Verfahren werden weiters für eine Anwendung über den *Use Case* hinaus aufbereitet und bereitgestellt.

Fazit

Anhand der vorgestellten Projekte wird ein Mal mehr deutlich, wie unterschiedlich die Voraussetzungen und Fragestellungen sein können, die unter dem Begriff *Big Data* verhandelt werden. Dabei tritt jedoch auch in den Vordergrund, was in diesem Feld die gemeinsamen, spezifisch geisteswissenschaftlichen Interessen sein können - ein methodologischer Austausch, von dem alle beteiligten Disziplinen profitieren.

Literatur

Jordan, Michael (2014): *Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts.* Interview by Lee Gomes, IEEE Spectrum. http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts (10.11.2014)

Anderson, Chris (2008): *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, Wired Magazine 16.07. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory (10.11.2014)