DARIAH-DE

Big, complex, heterogeneous..
Laufende Projekte aus dem Arbeitsbereich Big Data in den Geisteswissenschaften

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, die historische Entwicklung narrativer Techniken und in weiterer Folge auch die Entwicklung darauf aufbauender literarischer Kategorien anhand quantitativer Verfahren nachzuvollziehen. Ausgehend von einer automatischen Erkennung bestimmter Merkmale - wie zum Beispiel von Eigennamen oder Passagen direkter Rede - werden induktiv höhergelagerte Strukturen wie Handlungstypen und Figurennetzwerke, die sich für eine direkte Bearbeitung literaturwissenschaftlicher Fragestellungen eignen, gebildet. Daten: Ein Korpus 2000 deutschsprachiger literarischer Werke aus dem Zeitraum von 1500 bis 1930 und eine Sammlung 200 französischer Kriminalromane aus dem 19. und 20. Jahrhundert stellen die Datengrundlage dar.

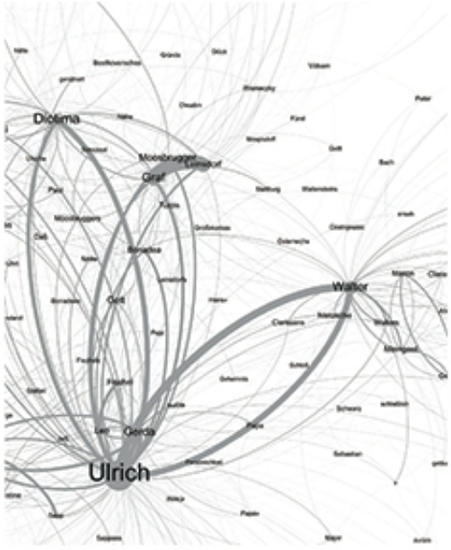

Abb. 1: Soziales Netzwerk aus Robert Musil *Der Mann ohne Eigenschaften*

Methoden: Aufbauend auf der modularen Werkzeugsammlung DKPro (Darmstadt Knowledge Processing Software Repository) wird eine generische Pipeline aus Vorverarbeitungsschritten (POS-Tagging, Named Entity Recognition, Parsing,..) erstellt. Ein eigenes Plaintext-Ausgabeformat macht eine vielfältige Weiterverarbeitung der dabei gewonnenen Informationen möglich. Um eine Induktion höhergelagerter Strukturen im Text zu erreichen, kommen Verfahren wie Netzwerkanalyse, Klassifikation, Clustering und Topic Modeling zum Einsatz. Deliverables: Entwickelte Methoden und Arbeitsabläufe werden als frei verfügbare Lernmaterialien aufbereitet, insbesondere für Scriptsprachen wie Python und R. Die Anleitungen sollen auch über Anpassung und Training der Werkzeuge für andere Forschungsfragen informieren.

Kontakt:
Dr. Steffen Pielström
Stefan Pernes, MA
Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Julius-Maximilians-Universität Würzburg

Identifikation grenzübergreifender Lebensläufe in europäischen Nationalbiografien

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 -

Abb. 2: Korrelationsgraph zu *Johann Wolfgang von Goethe*

miteinander korreliert und durch eine gezielte Erhebung sämtlicher Zusammenhänge mitunter Beobachtungen gemacht, die in den Geschichtswissenschaften noch nicht theoretisch erfasst sind. Daten: Wikidata, Wikipedia, ADB, Oxford Dictionary of National Biography, andere Nationalbiografien. Strukturierte Daten sowie unstrukturierte Texte in mehreren Sprachen werden miteinander verschränkt. Methoden: Es werden Korrelationsgraphen aus Personen, Zeitpunkten und Aktivitäten erzeugt, wobei eine Triangulation der Daten (strukturierte, unstrukturierte Daten, manuelle Nachrecherche) und ein Inferieren von Zusammenhängen anhand verwandter Personen stattfindet. Aus den angereicherten Einzelbiografien werden anschließend übergreifende Internationalitätskriterien gebildet. Deliverables: Implementierung als Teil der DARIAH-DE Generic Search Infrastruktur. Beinhaltet die Visualisierung von biographischen Ereignissen und stellt eine quantitative Grundlage für kontrollierte Vokabulare in der Biografieforschung dar.

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. Daten: Datengrundlage ist das OpenMigne Korpus, dessen Texte in griechischer und lateinischer Sprache einen Zeitraum vom Ursprung des Christentums bis in das 15. Jahrhundert abdecken

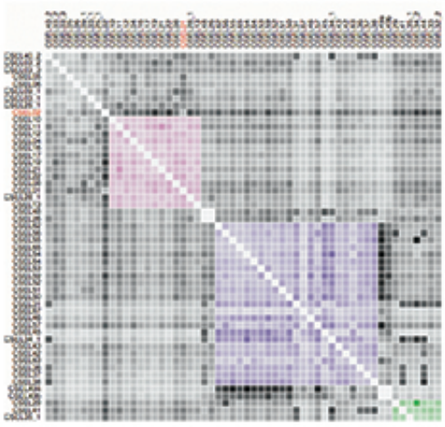

Abb. 3: Text-Reuse im *Corpus Scriptorum Ecclesiasticorum Latinorum*

Methoden: Grundlage bildet eine automatische Extraktion von Metadaten, anhand genre- und autorenspezifischer Schemata und OCR Layout Detection, sowie ein System zur Referenzierung von Textpassagen, das von einzelnen Worten bis hin zu ganzen Teilkorpora (Canonical Text Services) skaliert. Im Anschluss an diese Aufbereitung des Korpus wird eine Alignierung von Textpassagen und eine Erschließung diachroner Zitationsspuren anhand von Machine Learning Verfahren durchgeführt. Deliverables: Eine klassische Big Data Architektur (verteilter Speicher und Datenverarbeitung nach dem Map-Reduce Prinzip), deren Quellcode zur freien Verfügung gestellt wird. Auf das OpenMigne Korpus kann über eine API-Schnittstelle (Application Program Interface) und eine Weboberfläche zur Exploration des Korpus zugegriffen werden.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

UNIVERSITÄT LEIPZIG

TECHNISCHE
UNIVERSITÄT
DARMSTADT

UNIVERSITY OF BAMBERG
OTTO-FRIEDRICH-UNIVERSITÄT BAMBERG

IEG
Leibniz-Institut für Europäische Geschichte