Dokument Text Edition
Graduiertenkolleg 2196
gefördert durch: DFG Deutsche Forschungsgemeinschaft

Workshop

Modellieren, vernetzen, analysieren – Methoden und Techniken zur digitalen Erschließung und Strukturierung geisteswissenschaftlicher Textkorpora

Termin: 13. Februar 2017, 9.00–18.30 Uhr
Ort: Bergische Universität Wuppertal, Campus Grifflenberg (Hauptcampus)
Raum: K.11.07 (Senatssaal)

Programm

09:00 Begrüßung (*Jochen Johrendt*)
Einführung/Moderation: *Julia Nantke, Frederik Schlupkothen, Daniela Schulz*

09:15 *Øyvind Eide* (Köln): Modelling in scholarly editing: between investigation and production

10:30 *Stefan Gradmann* (Leuven): Beyond Infrastructure!

11:45 Pause

12:00 *Immanuel Normann* (Tübingen): Semantic-Web-Technologien für die Textwissenschaft

13:15 Mittagspause

14:35 *Steffen Lohmann* (Sankt Augustin): Möglichkeiten der IT-gestützten Textanalyse mittels NLP und Linked Data

15:50 *Ulrike Henny* (Köln/Würzburg): Von Themen zu Topics: über die Einsatzmöglichkeiten von Topic Modeling für quantitativ gestützte Inhaltsanalysen in den Geisteswissenschaften

17:05 Pause

17:20 *Gioele Barabucci* (Köln): Computer-assisted collation and digital documents: possibilities, limits and open research questions

anschließend: Abschlussdiskussion

Zum Thema

Im Zuge der Digitalisierung der Geisteswissenschaften halten vermehrt Technologien Einzug, welche auf eine Erfassung und Untersuchung von Texten und Textkorpora mit Hilfe digitaler Methoden und Werkzeuge abzielen. Dabei liegen unterschiedliche Erkenntnis- und Anwendungsinteressen zu Grunde, die ihren gemeinsamen Nenner in der systematischen Strukturierung größerer Textmengen haben. Daher bieten diese Methoden und Werkzeuge insbesondere im editionswissenschaftlichen Kontext Einsatzmöglichkeiten, die von der Authentifizierung strittiger Überlieferungssituationen bis hin zur Erfassung und Darstellung komplexer Textbestände im Rahmen einer Edition reichen und die sich in Zukunft wohl noch erweitern werden.

Aus geisteswissenschaftlicher Sicht erscheint es unabdingbar, sich mit den Funktionsweisen der digitalen Werkzeuge auseinanderzusetzen, denn diese verändern die Modalitäten der Archivierung, Strukturierung, Präsentation und damit der Wahrnehmung geisteswissenschaftlicher Gegenstände in entscheidender Form. Zudem zielen verschiedene digitale Werkzeuge nicht (nur) auf die Erfassung, sondern vielmehr (auch) auf eine (halb)automatische Analyse von Texten ab, wobei die Grenzen zwischen methodologischen Ansätzen und technischen Umsetzungen keinesfalls klar umrissen sind.

Zu einer Auseinandersetzung gehören daher sowohl die Frage der Nutzbarkeit bestimmter Werkzeuge und Techniken für die eigene Arbeit als auch die bewusste Reflexion des formierenden/perspektivierenden Einflusses, welchen die jeweiligen Methoden und Technologien auf die mit ihrer Hilfe erfassten Gegenstände haben.

Im Rahmen des geplanten Workshops sollen verschiedene Einsatzmöglichkeiten derartiger Technologien zur digitalen Strukturierung und Analyse geisteswissenschaftlicher Texte und Textkorpora hinsichtlich ihrer Funktionsweise, der ihnen zu Grunde gelegten methodologischen Prämissen sowie ihrer Einsatzmöglichkeiten in Bezug auf editions- und fachwissenschaftliche Erkenntnisinteressen vorgestellt und diskutiert werden. Der Fokus soll hierbei auf zwei Bereiche gelegt werden:

1. Technische Ansätze aus dem Umfeld der Dokumentenanalyse und des Semantic Web, die dazu dienen, etwa mit Hilfe von Data Mining und ‚Ontologien' systematische Beziehungsnetze in und zwischen Texten zu etablieren und auf diese Weise Zusammenhänge sichtbar zu machen und/oder eben überhaupt erst zu generieren. Innerhalb dieses Bereichs werden bspw. mit RDF und Topic Modelling unterschiedliche Strategien verfolgt.
2. Werkzeuge, die für konkrete analytische Zwecke wie stilometrische Verfahren zur Identifizierung von Autorschaft oder computergestütztes Kollationieren eingesetzt werden. Wie für die Technologien des ersten Bereichs gilt aller-

dings auch hier, dass die Werkzeuge nicht unabhängig von den hinter ihnen stehenden Vorstellungen bestehen, welche die ihnen zu Grunde liegenden Algorithmen entscheidend prägen.

Abstracts und Viten

Gioele Barabucci (Köln)

Computer-assisted collation and digital documents: possibilities, limits and open research questions

Abstract:

The task of finding differences among hundreds of documents (what textual critics call „collation“) look like a task that can easily be delegated to computers. However, expert philologists often frown upon these computer-generated collations: they are deemed too detailed („too noisy“) or not enough detailed („too superficial“), sometimes even too detailed and not enough detailed at the same time. How is this possible? This talk will discuss the state of the art in computer-assisted collation of medieval and modern texts, the limitations of the current tools and the profound connection between the task of comparing texts and the epistemological question „what is a (digital) document?“.

Vita:

Gioele Barabucci is a Marie Curie Experienced Researcher at the Cologne Center for eHumanities of the University of Cologne. He received his PhD in Computer Science from the University of Bologna.

His main research topics are the design of knowledge (how to represent and store information) and the evolution of information (understanding and forecasting how data and its structure will change over time). He is currently working on the automatic collation of the writings (in Latin, Arabic and Hebrew) of the Arab philosopher Ibn Rušd (Averroes). He also worked on the Edition of the Frankish Capitularies, the Cologne Sanskrit Dictionary, the Akoma Ntoso standard for legal documents and many other academic and open source projects. http://gioele.io/uni.

E-Mail: gioele.barabucci@uni-koeln.de

Øyvind Eide (Köln)
Modelling in scholarly editing: between investigation and production

Abstract:
The presentation will start with a general introduction to modelling, contextualising modelling in digital humanities in the broader context of computer science, philosophy of science, semiotics, as well as the humanities at large. An analytical distinction between modelling as a process of coming to know and modelling as a method for for production will be presented and connected to scholarly editing. Based on this, types of text based modelling will be presented, including text encoding, semantic tagging, and conceptual modelling. This will be linked to visualisations, which will always be based on more or less conscious interpretations of texts and can be used both to present results and as learning strategies.

Vita:
Øyvind Eide holds a PhD in Digital Humanities from King's College London (2013). He was an employee in various positions at The University of Oslo from 1995 to 2013, working on digital humanities and cultural heritage informatics. From 2013 to 2015 he was a Lecturer and research associate at The University of Passau. From October 2015 he is acting professor at the University of Cologne. He is an active member of the CIDOC-CRM SIG and a co-founder of the TEI Ontologies SIG. His research interests are focused on transformative digital intermedia studies, using critical stepwise formalisation as a method for conceptual modelling of cultural heritage information. This is used as a tool for critical engagement with media differences, especially the relationships between texts and maps as media of communication. He is also engaged in theoretical studies of modelling in the humanities as well as beyond.
E-Mail: oeide@uni-koeln.de

Stefan Gradmann (Leuven)
Beyond Infrastructure!

Vita:
Stefan Gradmann is a full Professor in the Arts department of KU Leuven (Belgium). Besides his continuing focus on knowledge management and semantically based operations his research and teaching covers digital libraries and web based information architectures, with a special emphasis on the digital humanities and a focus on semiological aspects of meaning and interpretation in the information ecosystem of the WWW.

Humanities Computing: He was an international advisor for the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences, and as such has contributed to the report “Our Cultural Commonwealth”. Furthermore, he has been leading the EC funded project Digitised Manuscripts to Europeana (DM2E).
Europeana: He has been heavily involved in building Europeana, the European Digital Library, from its very beginnings. More specifically he was leading work on technical and semantic interoperability and has been a co-author of the graph based Europeana Data Model (EDM) and triggered Europeana‘s involvement in the LoD community.
His working languages are German, English, French and Dutch.
E-Mail: stefan.gradmann@kuleuven.be

Ulrike Henny (Köln/Würzburg)

Von Themen zu Topics: über die Einsatzmöglichkeiten von Topic Modeling für quantitativ gestützte Inhaltsanalysen in den Geisteswissenschaften

Abstract:
Topic Modeling ist eine Methode der quantitativen Textanalyse, die ursprünglich entwickelt wurde, um die Abfrage und Suche von Informationen in großen Dokument-Sammlungen zu erleichtern. Inzwischen wird sie auch in den Geisteswissenschaften eingesetzt, um thematische Strukturen und die Verteilung von Themen in Textsammlungen aufzudecken. Mit dem Vortrag werden Beispiele von Topic-Model-Analysen aus verschiedenen Fachbereichen und basierend auf Sammlungen unterschiedlicher Textsorten vorgestellt, um die Möglichkeiten des Einsatzes der Methode zu demonstrieren. Zugleich soll das Verhältnis der Methode zu klassischen Verfahren inhaltlicher Textanalyse diskutiert werden. Welche Konzepte und Begriffe liegen „Topic Modeling“ zugrunde, wie unterscheiden sich diese von Begriffen z.B. aus der Literatur- und Sprachwissenschaft? Wo werden Grenzen der Anwendbarkeit des Verfahrens für geisteswissenschaftliche Fragestellungen und der Interpretierbarkeit der Ergebnisse sichtbar?

Vita:
Ulrike Henny hat in Köln und Lissabon Regionalwissenschaften Lateinamerika studiert. Nachdem sie an der Universität zu Köln das IT-Zertifikat für Geisteswissenschaften erwarb, arbeitete sie als Webentwicklerin bei der Universität der Vereinten Nationen in Bonn. Zwischen 2011 und 2015 war sie am Cologne Center for eHumanities (CCeH) für Datenmodellierung und Programmierung in verschiedenen digitalen Vorhaben der Nordrhein-Westfälischen Akademie der Wissenschaften zuständig. Aktuell ist sie Teil der Nachwuchsgruppe CLiGS (Computergestützte

Literarische Gattungsstilistik) an der Universität Würzburg und arbeitet an ihrer Dissertation zu Topics und Stil in Untergattungen des hispanoamerikanischen Romans im 19. Jahrhundert. Sie ist darüber hinaus Mitglied des Instituts für Dokumentologie und Editorik (IDE).
E-Mail: ulrike.henny@uni-koeln.de oder ulrike.henny@uni-wuerzburg.de

Steffen Lohmann (Sankt Augustin)
Möglichkeiten der IT-gestützten Textanalyse mittels NLP und Linked Data

Abstract:
Die Digitalisierung von Literatur und die zunehmende Verfügbarkeit von elektronischen Büchern eröffnen neue Möglichkeiten der IT-gestützten, teilautomatisierten Textanalyse. Methoden der Computerlinguistik ermöglichen es, Entitäten und deren Zusammenhänge aus digitalen Texten zu extrahieren. Diese können dann als strukturierte Informationen in Form von Linked Data repräsentiert werden. Dies wiederum erlaubt automatisierte Abfragen und die Visualisierung der extrahierten Informationen sowie darauf aufbauend die statistische, semantische und visuelle Analyse der Texte. Der Vortrag führt in das Thema der IT-gestützten Analyse von Literatur ein und veranschaulicht die Möglichkeiten und Potentiale der Anwendung von Linked Data in Kombinationen mit interaktiven Abfrage- und Visualisierungstechniken anhand ausgewählter Beispiele.

Vita:
Dr. Steffen Lohmann ist Wissenschaftler und Kompetenzfeldleiter am Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS). Seine Forschungsschwerpunkte liegen in den Bereichen semantische Technologien, Wissensverarbeitung, interaktive Systeme und Informationsvisualisierung. Er studierte Informatik, Psychologie und Sozialwissenschaften an den Universitäten Duisburg-Essen und Durham. Er promovierte in Informatik an den Universitäten Duisburg-Essen und Carlos III de Madrid. Nach seiner Promotion arbeitete er am Institut für Visualisierung und Interaktive Systeme der Universität Stuttgart, bevor er im November 2015 zur Fraunhofer-Gesellschaft wechselte. Dr. Steffen Lohmann ist Autor von mehr als 100 wissenschaftlichen Publikationen und in verschiedenen Forschungsgremien aktiv. Er ist regelmäßiger Gutachter von Beiträgen für Fachzeitschriften und Tagungen, hält Vorträge und organisiert Workshops zu den oben genannten Themenbereichen.
E-Mail: steffen.lohmann@iais.fraunhofer.de

Immanuel Normann (Tübingen)

Semantic-Web-Technologien für die Textwissenschaft

Abstract:

In der Text Encoding Initiave (TEI) ist über ein Zeitraum von bald drei Jahrzehnten ein gleichnamiges, XML-basiertes Dokumentenformat gereift, welches sich als De-facto-Standard zur Kodierung geisteswissenschaftlicher Textkorpora durchgesetzt hat. Mit XML ist aber eine bestimmte Entscheidung getroffen, nämlich den zu kodierenden Gegenstand möglichst als eindeutig geschachtelte Struktur aufzufassen. Naturgemäß kommt es da zu Spannungen, wo Strukturen überlappen (z.B. Vers vs. Zeile oder Absatz vs. Seite). Um die Einschränkung auf baumartige Strukturierung (XML) zu umgehen bietet sich eine netzwerkartige Kodierung an. Das Resource Description Framework (RDF) ist diesbezüglich der Standard des Semantic Web. Insbesondere wenn es darum geht, auch Relationen aus dem Text in einen prinzipiell unbegrenzten Kontext zu erfassen, zeigt RDF seine Stärken. Nach einer allgemeinen Einführung in semantische Netze soll in dem Vortrag skizziert werden, wie RDF für die Textwissenschaft eingesetzt werden könnte.

Vita:

Immanuel Normann wurde 2009 an der Jacobs University Bremen in Computer Science zum Thema „Automated Theory Interpretation“ promoviert. Als wissenschaftlicher Mitarbeiter an der Universität Bremen hat er auf dem Gebiet formaler Sprachen geforscht. Am Birkbeck College in London war er als knowledge engineer in einem Museumsprojekt über Textilkultur der Andenvölker tätig. Seit 2012 arbeitet er als Entwickler in der Abteilung digital humanities bei der pagina GmbH in Tübingen - derzeit an Online-Editionen der Werke Walter Benjamins und Arno Schmidts.

E-Mail: immanuel.normann@pagina-tuebingen.de