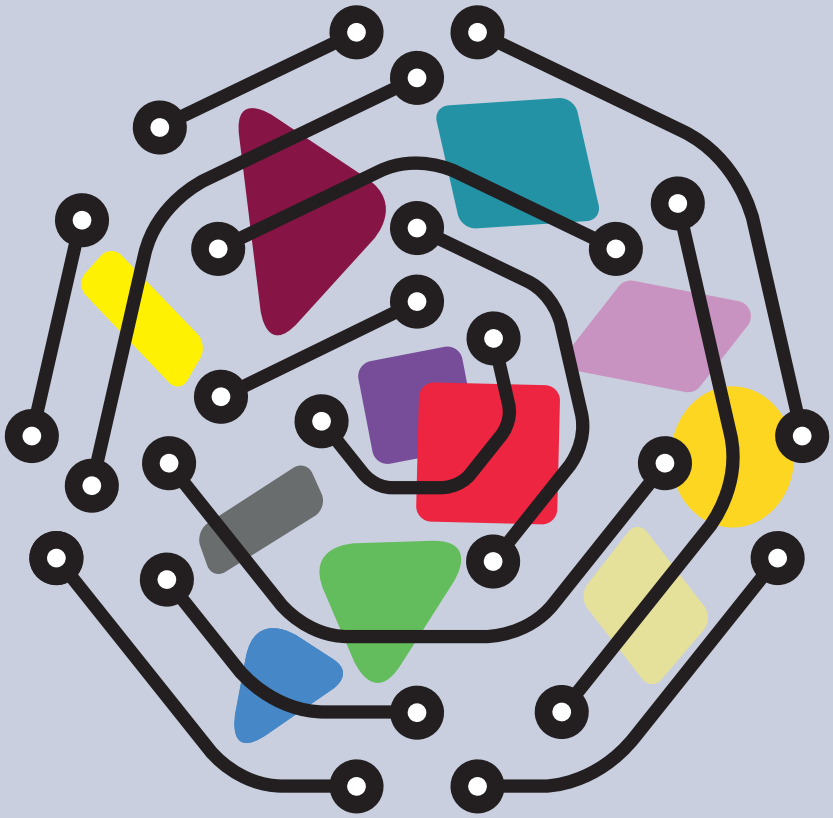
Konferenzabstracts



DHd 2016

MODELLIERUNG
VERNETZUNG
VISUALISIERUNG

DIE DIGITAL HUMANITIES
ALS FÄCHERÜBERGREIFENDES
FORSCHUNGSPARADIGMA

UNIVERSITÄT LEIPZIG
7. – 12. MÄRZ 2016

Digital Humanities im deutschsprachigen Raum (DHd)

DHd 2016

Modellierung - Vernetzung - Visualisierung
Die Digital Humanities als fächerübergreifendes Forschungsparadigma

Konferenzabstracts

Universität Leipzig
7. bis 12. März 2016

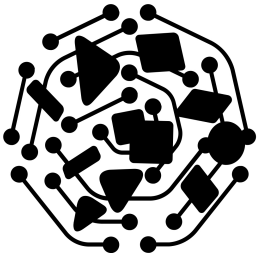
Die Abstracts wurden von den Autorinnen und Autoren in einem Template erstellt und mittels des von Marco Petris, Universität Hamburg, entwickelten DHConvalidators in eine TEI konforme XML-Datei konvertiert.

Koordination der Publikation: Elisabeth Burr
Korrektur der Auszeichnung der Bibliographie:
Julia Burkhardt
Fleur Pfeifer
Rebecca Sierig
Konvertierung TEI nach PDF: Aramís Concepción Durán
TEI to PDF scripts: Karin Dalziel
https://github.com/karindalziel/TEI-to-PDF
Konferenz-Logo und Umschlaggestaltung: Alexander Morgenstern

online verfügbar: http://dhd2016.de/

ISBN 978-3-941379-05-3

3. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V.



DHd 2016

MODELLIERUNG
VERNETZUNG
VISUALISIERUNG

DIE DIGITAL HUMANITIES ALS
FÄCHERÜBERGREIFENDES
FORSCHUNGSPARADIGMA

Plenarvorträge

Workshops

Panels

Sektionen

Vorträge

Poster

Plenarvorträge

Die Rolle von Mensch und Computer in den Digital Humanities

Keim, Daniel A.
Daniel.Keim@uni-konstanz.de
Universität Konstanz, Deutschland

Computerbasierte Analysen sind zentraler Bestandteil der Digital Humanities. Zahlreiche geisteswissenschaftliche Fragen können mit Hilfe algorithmischer Methoden auf dem Computer schneller und auf breiterer Datenbasis beantwortet werden - und in einigen Fällen können auch neuartige Fragen behandelt werden. In zahlreichen Fällen aber reichen die automatischen Analysemethoden nicht aus, um die Daten zu verstehen und valide Schlussfolgerungen zu ziehen. Der Mensch mit seinen Fähigkeiten - seinem Hintergrundwissen, seiner Kreativität und seinem Urteilsvermögen – muss integraler Bestandteil des Analyseprozesses sein und effektiv durch automatische Verfahren unterstützt werden. Die Darstellung der Daten und Analyseergebnisse mit Hilfe von Visualisierungen spielt dabei eine wichtige Rolle. Für eine effektive Datenexploration müssen interaktive Visualisierungen eng mit automatischen Analysemethoden verknüpft werden. Beispiele aus den Digital Humanities zeigen das Potential dieses Ansatzes, aber auch seine Grenzen.

Von den 'digital humanities' zu einer humanen Digitalisierung

Zweig, Katharina Anna
zweig@cs.uni-kl.de
TU Kaiserslautern, Deutschland

Über Jahrzehnte hatte die Informatik mit den Geisteswissenschaften nur wenig Berührungspunkte - dies hat sich in den letzten Jahren gründlich geändert und die Verwendung informatischer Werkzeuge in den Geisteswissenschaften erlebt einen wahren Boom. Und dabei geht es nicht mehr nur um die reine Verwaltung von Daten in großen Datenbanken, sondern mehr und mehr um die Verwendung von künstlicher Intelligenz auf sinnhaft modellierte Informationen, um implizite Beziehungen sichtbar zu machen. Anhand einer spezifischen Modellierung von Daten aus dem geisteswissenschaftlichen Bereich als komplexe Netzwerke werde ich zeigen, dass die Analyse solcher Netzwerke wissenschaftstheoretisch oftmals gleich zwei Modelle enthält, die aber häufig höchstens implizit angesprochen werden. Dadurch kommt es immer wieder zu Fehlinterpretationen, wenn Algorithmen aus fertigen Softwarepackages unbedacht auf derart modellierte Daten angewendet werden.

Wenn es auch bisher so aussieht, als sei die nutzbringende Komponente der Beziehung zwischen Informatik und den Geisteswissenschaften rein auf der Algorithmenseite zu finden, zeigt die obige Analyse, dass wir Informatiker mehr denn je der Geisteswissenschaften bedürfen, um die Digitalisierung human gestalten zu können. Algorithmen werden heute dazu verwendet, um "kriminelle Persönlichkeiten" zu identifizieren, Kredite zu verleihen, oder um Versicherungstarife zu bestimmen. Wie können wir in einer solchen Situation vermeiden, dass Algorithmen verzerren, diskriminieren oder gar manipulieren? Wie können sensible Machtbalancen zwischen Privatheit und Sicherheit, ökonomischem Erfolg und Transparenz hergestellt werden? Hier brauchen wir die Analogiebildung der Historiker, die besten Einsichten in das Werden von Gesellschaft von Soziologinnen und Wirtschaftswissenschaftlern, fundierte Einblicke in die menschliche Psyche durch Psychologinnen, und nicht zuletzt das rechte Maß an Regulierung durch Rechtswissenschaftlerinnen und Rechtswissenschaftler. Es könnte das Jahrhundert der Geisteswissenschaften werden.

Workshops

Komponisten-Datenbanken / -Portale: Entwicklungsmöglichkeiten, Austauschformate und Vernetzungspotential

Blanken, Christine
blanken@bach-leipzig.de
Bach-Archiv Leipzig, Deutschland

Rettinghaus, Klaus
rettinghaus@bach-leipzig.de
Bach-Archiv Leipzig, Deutschland

Siegert, Christine
Christine.Siegert@beethoven-haus-bonn.de
Beethoven-Haus Bonn, Deutschland

Dubowy, Norbert
dubowy@mozarteum.at
Mozarteum, Mozart-Institut, Digitale Mozart-Edition, Salzburg, Oesterreich

Schwinger, Tobias
Tobias.Schwinger@sbb.spk-berlin.de
Staatsbibliothek zu Berlin, KoFIM-Projekt, Berlin, Deutschland

Muehlberger, Guenter
guenter.muehlberger@uibk.ac.at
Universitaet Innsbruck, Institut fuer Germanistik, Innsbruck, Oesterreich

Christlein, Vincent
vincent.christlein@fau.de
Universitaet Erlangen-Nuernberg, Institut fuer Informatik, Erlangen, Deutschland

Stadler, Peter
stadler@weber-gesamtausgabe.de
Carl-Maria-von-Weber-Gesamtausgabe, Detmold, Deutschland

Schildt, Maria
maria.schildt@musik.uu.se
Uppsala Universitet, Department of Musicology, Dueben Collection Database Uppsala, Schweden

Wiermann, Barbara
Barbara.Wiermann@slub-dresden.de
Saechsische Landesbibliothek - Staats- und Universitaetsbibliothek Dresden, Musiksammlung, Dresden, Deutschland

Schmidt, Frieder
F.Schmidt@dnb.de
Deutsche Nationalbibliothek, Papierhistorische Sammlungen, Leipzig, Deutschland

Schneider, Dietmar
Dietmar.Schneider@Startmail.com
Privatier, Nuernberg (common science)

Hausmann, Christiane
hausmann@bach-leipzig.de
Bach-Archiv Leipzig, Deutschland

Morgenstern, Anja
morgenstern@mozarteum.at
Mozarteum, Mozart-Institut, Digitale Mozart-Edition, Salzburg, Oesterreich

Wollny, Peter
wollny@bach-leipzig.de
Bach-Archiv Leipzig, Deutschland

Kupferschmidt, Jens
kupferschmidt@rz.uni-leipzig.de
Universitaet Leipzig, Rechenzentrum, Leipzig, Deutschland

Bärwald, Manuel
barwald@bach-leipzig.de
Bach-Archiv Leipzig, Deutschland

Fragen/Probleme für die Zukunft von Datenbanken musikalischer Quellen

Die Entwicklung digitaler Medien und die daraus resultierenden Chancen für eine Weiterentwicklung computergestützter Verfahren zeitigt weitreichende Folgen auch für die musikwissenschaftliche Grundlagenforschung. Die philologisch arbeitenden Disziplinen profitieren ungemein von den Digitalisierungsvorhaben in Bibliotheken oder haben selbst an solchen Vorhaben ihren Anteil. Und sie haben starke gemeinsame Interessen: Neben Tools zur digitalen Edition sind dies vor allem Schreiber-

Erkennung, Papier- und Wasserzeichenforschung sowie Provenienzrecherchen. Das Interesse, über Standards zu diskutieren und die Grundlagen für gemeinsame digitale Standards weiterzuentwickeln ist bei Editionsvorhaben genauso vorhanden wie bei Bibliotheken. Derzeit bereits gegebene Vernetzungsmöglichkeiten werden genutzt und sollten weiter ausgebaut werden; ein Beispiel dafür ist eine übergeordnete Forschungsinfrastruktur, wie sie etwa in Bezug auf Papierforschung das Wasserzeichen-Informationssystem und die Papierhistorischen Sammlungen der DNB zur Verfügung stellen.

Darüber hinaus werden in vielen musikwissenschaftlichen Forschungseinrichtungen seit Jahrzehnten Daten zu Komponisten und ihren Werken zusammengetragen, seit etwa 2000 erfolgt dies im deutschsprachigen Raum auch per Datenbanken. Vernetzungen dieser Daten sind dabei aber bislang die Ausnahme. Ein Austausch könnte also auch auf dieser Ebene intensiviert werden.

Ein weiterer Aspekt betrifft die in den letzten Jahren entwickelten Methoden der Auswertung strukturierter Daten. Auch wenn sie im Bereich der Musikwissenschaft quantitativ wohl noch nicht unter den Begriff „Big Data“ fallen, so stellen diese Daten einen Fundus dar, welcher mit Hilfe vieler, in verschiedenen Projekten unter dem Label Digital Humanities laufender Methoden einer Auswertung harrt. Voraussetzung dafür wäre allerdings eine stärkere Vernetzung.

In der Bach-Forschung sind engmaschige Untersuchungen zur Überlieferung jedes einzelnen Musikwerks seit langem ein essentieller Bestandteil, denn viele Werke J. S. Bachs sind weder autograph überliefert noch genau zu datieren. Dies hat zur Folge, dass ein großer Teil der Untersuchungen von Bach-Quellen Handschriften des gesamten 18. und frühen 19. Jahrhunderts betreffen müssen. Sie stammen von oft unbekannten Schreibern mit unklarer Provenienz. Ihren Bezug zu verschollenen originalen Quellen zu ermitteln, ist damit seit den 1950er Jahren – angestoßen und betrieben durch die Arbeit an der Neuen Bach-Ausgabe – ein essentieller Bestandteil der Bach-Forschung. Diese hat sich so auf einigen Feldern zu einem Vorreiter in der paläographisch und philologisch orientierten Quellenforschung entwickelt. Die entsprechenden Erkenntnisse wurden in den Kritischen Berichten dieser Gesamtausgabe ausgewertet; mit Blick auf die gesamte Bach-Familie darüber hinaus in gedruckten Katalogen über einzelne Quellenbestände, vor allem in den Leipziger Beiträgen zur Bach-Forschung: Brüsseler Bibliotheken (1997), Singakademie zu Berlin (2005), Wien und ‚Alt-Österreich‘ (2011). Sowohl das Wissen als auch die Methoden wurden im Laufe der vergangenen Jahrzehnte ebenfalls für Forschungen zu anderen Komponisten nachgenutzt. Durch diese Impulse konnten wiederum auch für die Bach-Überlieferung Erkenntnisse gewonnen werden. So haben beispielsweise durch die Recherchen zu Berliner Überlieferungskreisen, namentlich der Singakademie, gerade auch die Gesamtausgaben zu den Söhnen Bachs sehr profitiert.

Um die Fülle der auf viele Kritische Berichte und andere Publikationen verteilten Forschungsergebnisse strukturiert recherchierbar zu machen, wurde 1999 in Göttingen am dortigen Johann Sebastian Bach-Institut die Bach-Quellen-Datenbank erstellt (seit 2001 als bach.gwdg.de online, Blanken 2002), die 2010 in das Portal Bach digital integriert wurde, das nunmehr nicht allein Informationen zu den Werken und ihren Quellen bietet, sondern auch hochauflösende Digitalisate der Handschriften selbst. Seit einigen Jahren werden sukzessive auch Daten / Digitalisate zu den Werken weiterer Komponisten der Bach-Familie berücksichtigt (Alt-Bachisches Archiv, Carl Philipp Emanuel, Wilhelm Friedemann und Johann Christoph Friedrich Bach), so dass bach-digital.de mittlerweile eine Datenbank zur gesamten Bach-Familie ist, mit derzeit knapp 7800 Quellen-, 3500 Werk-Datensätzen sowie 1750 Digitalisaten. Durch die Zusammenschau von Quellen und die Möglichkeit des strukturierten Zugriffs auf die hierzu gehörenden Informationen wird eine immer neue Beschäftigung mit den Werken der Bach-Familie herausgefordert.

Bach digital versteht sich dabei als ein Work in Progress, das es täglich weiterzuentwickeln und mit neuen Inhalten zu befüllen gilt. Dafür werden Anregungen von Nutzern und auch die aktive Mitarbeit einzelner externer registrierter Nutzer in Anspruch genommen. Die Zugriffsstatistik zeigt, dass Bach digital auch international sehr gut angenommen wird. Derzeit wird daher mittels mehrsprachiger Datenvorhaltung (Teilübersetzungen in Englisch, Japanisch und Französisch) gerade die internationale Ausrichtung gestärkt. Über die Bestände der drei derzeitigen Kooperationspartner Bach-Archiv Leipzig, Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden und das Rechenzentrum der Universität Leipzig hinaus ist es nun das Ziel, die Zahl der Bibliothekspartner zu erhöhen, um den Zugang zu den weltweit verstreuten Quellen zu erleichtern. Neben etlichen kleineren Sammlungen in Deutschland sind dies u. a. auch die British Library und die Library of Congress, die zugesagt haben, digitalisierte Bach-Quellen über Bach digital zur allgemeinen Verfügung zu stellen.

Die seit 16 Jahren ununterbrochene und tägliche Arbeit an einer open source-basierten Datenbank (Blanken et al. 2015) und ihre technische wie inhaltliche Weiterentwicklung sind nun an einem Punkt, da richtungweisende Entscheidungen zum Ausbau, aber auch zur Vernetzung mit anderen Projekten anstehen. Der Grundbestand der Daten von bach-digital.de ist jederzeit für andere Projekte nachnutzbar. Diese ganz oder teilweise erfolgende Überführung von Daten in andere Datenbanken hat Konsequenzen, über die grundsätzlich zu sprechen ist.

Hier nun sollen Erfahrungen, Perspektiven und Wünsche anderer und eventuell vergleichbarer

Datenbanken oder Digitalisierungsprojekte einbezogen werden.

Fragen / Probleme für die Zukunft von Datenbanken musikalischer Quellen

Inhaltliche Fragestellungen

Potential von Provenienz-Recherchen

Ausschöpfung des Potenzials der Gemeinsamen Normdatei (GND) für die

Provenienz-Forschung (Dokumentation historischer Musiksammlungen, Digitalisierung von Besitz- oder Auktionskatalogen, Geo-Referenzierung etc.

Vernetzung mit anderen Projekten, externe Nutzung dieser Daten)

Schreiberforschung

- Entwicklung von Standards für Schreiber-Nomenklaturen (Leitfragen: Wie sollten Beispiel-Sammlungen von Schriftproben strukturiert sein? Wie lassen sich gemeinsame Schreiber-Portale aufbauen?)
- Vernetzung von bereits vorhandenen Schriftproben-Datenbanken
- Automatische Schreiberhanderkennung (Leitfragen: Wo stehen wir in der Musikpaläographie? Was lässt sich von außermusikalischen Projekten lernen? Gibt es überhaupt Bedarf, wissenschaftliche Anstrengungen zu einer automatischen Schreibererkennung mithilfe der Informatik zu unternehmen?)

Wasserzeichen/Papier-Forschung

- Gemeinsamer sukzessiver Ausbau von Wasserzeichen-Recherche-Portalen: Wasserzeichen-Informationssystem (WZIS), Bernstein / Memory of Paper (WZ-Pause versus Aufnahmen mit moderner Kameratechnik, z. B. mittels Thermographie)

Technische Fragestellungen

Vernetzung

- Nutzerfreundliche Anbindung externer Angebote mit zusätzlichen Informationen (Crosslinking) durch den breiten Einsatz von Normdaten und dem BEACON-Format

Austauschformate

- Bereitstellung von Forschungsdaten in standardisierten und etablierten Formaten zur einfachen Weiterverwendung und automatisierten Auswertung
- Verwendung freier Lizenzen für wissenschaftskonforme Nachnutzung (Prinzipien guter wissenschaftlicher Praxis)

Notenincipits und Libretti

- Nutzen und Potenzial von TEI und MEI in Musiker-Datenbanken

Tabellen oder Ontologien?

- Datenmodelierung zwischen Standardisierung und individuellen sowie praktischen Bedürfnissen

Dokumentation und Lizenzen

- Offenlegung von Struktur und Inhalt zur langfristigen Verfügbarmachung

Wissenschaftskommunikation

- Stringenter Ausbau einer Common Science-Plattform; Installierung eines Redaktionsteams
- Nutzerfreundliche Weiterentwicklung von bisher primär wissenschaftlich orientierten Plattformen, Öffentlichkeitswirksamkeit (unter Einbeziehung weiterer digitaler Medien: Audio / Video / Editionen)

Datenqualität

- RISM-OPAC (Chancen und Probleme bei Datenübernahmen aus Komponisten-basierten Datenbanken)
- Konsistenz von Daten (innerhalb eines Projekts und projektübergreifend)
- Identifizierung und Auffindbarkeit durch Verwendung von Normdaten

Bibliographie

Betz, Florian (2016): "Papiermacher und Papiermühlen in der Gemeinsamen Normdatei (GND). Das Normdaten-Projekt 'Papiermacherkatalog' des Deutschen Buch- und Schriftmuseums der Deutschen Nationalbibliothek", in: Eckhardt, Wolfgang /

Neumann, Julia / Schwinger, Tobias / Staub, Alexander (eds.): *Wasserzeichen – Schreiber – Provenienzen*. Neue Methoden der Erforschung und Erschließung von Kulturgut im digitalen Zeitalter: Zwischen wissenschaftlicher Spezialdisziplin und 'Catalog Enrichment' (= Zeitschrift für Bibliothekswesen und Bibliographie Sonderband 118). Frankfurt am Main: Vittorio Klostermann 243-254

Blanken, Christine (2002): *Göttinger Bach-Katalog* http://www.bach.gwdg.de/

Blanken, Christine (2016): "Die Komponisten-Datenbank 'Bach digital'. Erfahrungen und Perspektiven abseits einer Präsentation von Digitalisaten", in: Eckhardt, Wolfgang / Neumann, Julia / Schwinger, Tobias / Staub, Alexander (eds.): *Wasserzeichen – Schreiber – Provenienzen*. Neue Methoden der Erforschung und Erschließung von Kulturgut im digitalen Zeitalter: Zwischen wissenschaftlicher Spezialdisziplin und 'Catalog Enrichment' (= Zeitschrift für Bibliothekswesen und Bibliographie Sonderband 118). Frankfurt am Main: Vittorio Klostermann 135-148.

Blanken, Christine / Rettinghaus, Klaus / Hausmann, Christiane / Kupferschmidt, Jens / Freitag, Stefan (2015): *Bach digital.* Dokumentation: Umsetzung des Projektes auf Basis der Content Management Anwendung des MyCoRe-Arbeitskreises. http://www.bach-digital.de/docs/BachDigital_Doku.pdf?XSL.lastPage.SESSION=/docs/BachDigital_Doku.pdf .

Eckhardt, Wolfgang (2016): "Digitale Dokumentation von Wasserzeichen in Musikhandschriften im Rahmen des Projekts KoFIM", in: Eckhardt, Wolfgang / Neumann, Julia / Schwinger, Tobias / Staub, Alexander (eds.): *Wasserzeichen – Schreiber – Provenienzen*. Neue Methoden der Erforschung und Erschließung von Kulturgut im digitalen Zeitalter: Zwischen wissenschaftlicher Spezialdisziplin und 'Catalog Enrichment' (= Zeitschrift für Bibliothekswesen und Bibliographie Sonderband 118). Frankfurt am Main: Vittorio Klostermann 167-196

Mühlberger, Günter (o. J.): Die automatisierte Volltexterkennung historischer Handschriften als gemeinsame Aufgabe von Archiven, Geistes- und Computerwissenschaftlern. Das Modell einer zentralen Transkriptionsplattform als virtuelle Forschungsumgebung https://www.academia.edu/7451967/Die_automatisierte_Volltexterkennung_historischer_Handschriften_als_gemeinsame_Aufgabe_von_Archiven_Geistes-_und_Computerwissenschaftlern._Das_Modell_einer_zentralen_Transkriptionsplattform_als_virtuelle_Forschungsumgebung

Rettinghaus, Klaus (2014): "Bringing together Bach and MEI – Future prospects for Bach digital", Vortrag bei der “Music Encoding Conference” 2014, Charlottesville, Virginia / USA: University og Virginia.

Stadler, Peter (2016): "Zum Einsatz von Normdaten bei der Carl-Maria-von-Weber-Gesamtausgabe", in: Eckhardt, Wolfgang / Neumann, Julia / Schwinger, Tobias / Staub, Alexander (eds.): *Wasserzeichen – Schreiber – Provenienzen*. Neue Methoden der Erforschung und Erschließung von Kulturgut im digitalen Zeitalter: Zwischen wissenschaftlicher Spezialdisziplin und 'Catalog Enrichment' (= Zeitschrift für Bibliothekswesen und Bibliographie Sonderband 118). Frankfurt am Main: Vittorio Klostermann 19-26.

Transkribus. Universität Innsbruck https://transkribus.eu/ .

Wenger, Emanuel (2016): "Metasuche in Wasserzeichendatenbanken (Bernstein-Projekt): Herausforderungen für die Zusammenführung heterogener Wasserzeichen-Metadaten", in: Eckhardt, Wolfgang / Neumann, Julia / Schwinger, Tobias / Staub, Alexander (eds.): *Wasserzeichen – Schreiber – Provenienzen*. Neue Methoden der Erforschung und Erschließung von Kulturgut im digitalen Zeitalter: Zwischen wissenschaftlicher Spezialdisziplin und 'Catalog Enrichment' (= Zeitschrift für Bibliothekswesen und Bibliographie Sonderband 118). Frankfurt am Main: Vittorio Klostermann 289-295

CATMA - Eine Plattform zum kollaborativen und automatisierten Annotieren und Analysieren von Texten

Bögel, Thomas
thomas.boegel@informatik.uni-heidelberg.de
Universität Heidelberg, Deutschland

Gius, Evelyn
evelyn.gius@uni-hamburg.de
Universität Hamburg, Deutschland

Petris, Marco
marco.petris@uni-hamburg.de
Universität Hamburg, Deutschland

Strötgen, Jannik
jannik.stroetgen@mpi-inf.mpg.de
Max-Planck-Institut für Informatik Saarbrücken, Deutschland

Beschreibung

Dieser Workshop widmet sich der Textannotation und der Textanalyse mit der web-basierten Annotationsplattform CATMA (*Computer Aided Text Markup and Analysis*) (Meister et al. 2015), welche

seit 2008 an der Universität Hamburg entwickelt wird. Die Bedarfe der Modellierung und Operationalisierung geisteswissenschaftlicher Konzepte und die Anwendung dieser Modelle auf Textdaten stand bei der Entwicklung von CATMA im Fokus. CATMA ist Open Source und außerdem XML / TEI-kompatibel, dadurch ist die Nachnutzbarkeit der mit CATMA erstellten Annotationen und Analyseergebnisse sichergestellt. Der Workshop wird neben einer Einführung in die Nutzung der Plattform für manuelles Annotieren auf zwei weitere – gerade bei der Umsetzung größerer Annotationsprojekte wesentliche – Aspekte genauer eingehen: Kollaboration und Automatisierung.

Im Workshop wird gezeigt werden, welche Funktionalitäten in CATMA für kollaboratives Arbeiten zur Verfügung stehen und wie das kollaborative Arbeiten unter den erschwerten Bedingungen der literaturwissenschaftlichen Praxis, wie z. B. der Polyvalenz literarischer Texte, möglich ist. Beim automatischen Erstellen von Annotationen hingegen spielen die kürzlich im Rahmen des heureCLÉA -Projektes [1] in CATMA integrierten Möglichkeiten eine zentrale Rolle.

Ziel von heureCLÉA ist die Bereitstellung einer digitalen Heuristik zur Annotation von einfachen bis komplexen Konzepten der Narratologie. Unter einer digitalen Heuristik verstehen wir ein Werkzeug zur automatischen und semiautomatischen Annotation. Das heureCLÉA-Projekt konzentriert sich hierfür auf die Analyse temporaler Phänomene. Auf einer einfachen, an der Textoberfläche orientierten Ebene handelt es sich hierbei unter anderem um die Erkennung von Zeitausdrücken in literarischen Texten, auf einer komplexeren, metatextuellen Ebene beispielsweise um die Erkennung von Phänomenen der zeitlichen Ordnung wie Prolepse und Analepse (Gius / Jacke 2014). Dafür entwickelten wir einen auf manuellen und automatischen Annotationen basierten Ansatz, in dem die regelbasierte Extraktion und Normalisierung von Zeitausdrücken als Ausgangspunkt für Machine Learning Verfahren verwendet wurde (Bögel et al. 2015). Eine ganz wesentliche Komponente dieses Ansatzes ist das an der Universität Heidelberg entwickelte System HeidelTime (Strötgen / Gertz 2013). Sowohl die automatische Annotation von Zeitausdrücken, als auch linguistischer Oberflächenphänomene (Wortarten und Satzgrenzen), sowie Tempusannotationen sind bereits in CATMA integriert und können mit manuellen Annotationen kombiniert werden.

Im Workshop werden zunächst wesentliche Aspekte wie Modellierung, Annotation, Analyse, Kollaboration und Automatisierung mit CATMA anhand der Erfahrungen des heureCLÉA-Projektes vorgestellt. Anschließend haben die Teilnehmer_innen die Gelegenheit in einer praktischen hands-on-Session CATMA sowie HeidelTime und andere Komponenten der in CATMA integrierten NLP-Pipeline auszuprobieren. Automatische Annotationen können evaluiert werden und mit manuellen Annotationen kombiniert in die Analyse einfließen. Es kann entweder mit eigenen oder kollaborativ mit von uns zur Verfügung gestellten Texten gearbeitet werden.

Wir erhoffen uns außerdem durch den Workshop kritisches Feedback zur weiteren Verbesserung von CATMA und eine Diskussion über die Anforderungen für Textanalyse-Plattformen in verschiedenen Bereichen der Digital Humanities.

Beitragende

Alle Veranstalter_innen sind Mitglieder des heureCLÉA-Projektes. Wir haben auf zahlreichen nationalen und internationalen Tagungen und Konferenzen unsere Arbeiten zu CATMA, HeidelTime und heureCLÉA vorgestellt. Dieser Workshop baut auf Erfahrungen aus anderen Workshops zum selben Thema sowie der Einbettung von CATMA in die Lehre auf. Auch das sehr positive Feedback vergangener Workshops, z. B. zu unserem Tutorial im Rahmen der DH 2014, hat uns dazu motiviert, erneut einen Workshop anzubieten bzw. einen Antrag dafür einzureichen.

Thomas Bögel , Institut für Informatik, Universität Heidelberg,

Nach seinem Computerlinguistikstudium begann Thomas Bögel sein Promotionsstudium am Institut für Informatik an der Universität Heidelberg, wo er auch wissenschaftlicher Mitarbeiter ist. Seine Forschung beschäftigt sich vor allem mit "event extraction" und "timeline generation" sowie mit der Entwicklung von Machine Learning Systemen für die Extraktion von temporalen Relationen in narratologischen Texten.

Evelyn Gius , Institut für Germanistik, Universität Hamburg,

Evelyn Gius forscht und lehrt im Bereich der Digital Humanities mit einem Fokus auf computergestützter Textanalyse und der Hermeneutik digitaler Zugänge zu Texten. In ihrer Promotion hat sie mithilfe von CATMA an einem Korpus von Erzählungen über Arbeitssituationen untersucht, inwiefern narratologische Kategorien aus der Literaturwissenschaft für die Analyse der Konflikthaftigkeit von Alltagserzählungen genutzt werden können.

Marco Petris , Institut für Germanistik, Universität Hamburg,

Marco Petris ist Informatiker mit starker Affinität für die Geisteswissenschaften und hat von Beginn an CATMA federführend aufgebaut. Als Software Entwickler ist er in zahlreiche Projekte für die Digital Humanities involviert, wobei er sich dabei vor allem um die Konzeption und Implementierung kümmert.

Jannik Strötgen , Max-Planck-Institut für Informatik, Saarbrücken,

Bevor Jannik Strötgen als Postdoc zum MPI wechselte, studierte er in Heidelberg Computerlinguistik

und promovierte und arbeitete am Institut für Informatik der Universität Heidelberg. Im Rahmen seiner Dissertation beschäftigte er sich vor allem mit Informationsextraktion sowie Information Retrieval und begann die Entwicklung von HeidelTime, einem frei verfügbaren Temporal Tagger, der für verschiedene Domänen und Sprachen geeignet ist.

Kapazität und Ausstattung

Die Teilnehmerzahl ist auf 20 Personen begrenzt. Jede_r Teilnehmer_in braucht einen Laptop (ein Tablet PC reicht nicht aus!) und ein Google Mail Konto für den CATMA Login.

Notes

1. heureCLEA ist ein BMBF-gefördertes ehumanities Projekt zwischen der Universität Hamburg und der Universität Heidelberg (Gertz / Meister 2016).

Bibliographie

Bögel,Thomas / Strötgen, Jannik / Gertz, Michael (2015): „A Hybrid Approach to Extract Temporal Signals from Narratives“, accepted at: *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL'15)*, Duisburg-Essen, Germany.

Gertz, Michael / Meister, Jan Christioph (2016): *heureCLÉA*. Collaborative Literature exploration & annotation. Hamburg: Universität Hamburg http://heureclea.de/ [letzter Zugriff 08. Oktober 2015].

Gius, Evelyn / Jacke, Janina (2014): „Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets“. Hamburg 2014 [letzter Zugriff 08. Oktober 2015].

Meister, Jan Christoph / Gius, Evelyn / Petris, Marco / Meister, Malte / Jacke, Janina (2015): *CATMA*. Computer Aided Textual Markup Computer Aided Textual Markup & AnalysisAnalysis. Hamburg: Universität Hamburg http://www.catma.de/ [letzter Zugriff 08. Oktober 2015].

Strötgen, Jannik / Gertz, Michael (2013): „Multilingual and cross-domain temporal tagging“ in: *Language Resources and Evaluation* 47, 2: 269-298.

nodegoat Workshop: Einführung in die Nutzung einer multifunktionalen webbasierten Datenbankapplikation für Geisteswissenschaftler

Kessels, Geert
geert@lab1100.com
LAB1100, Niederlande

van Bree, Pim
pim@lab1100.com
LAB1100, Niederlande

Einleitung

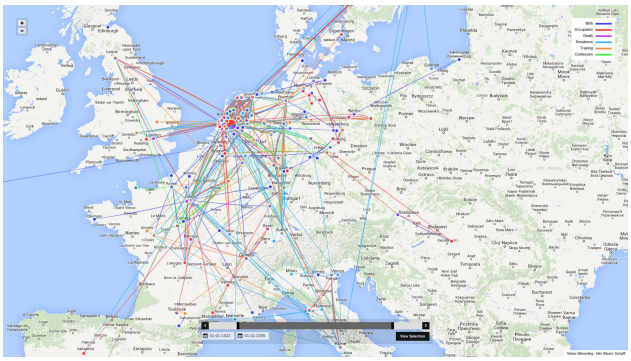
nodegoat ist eine multifunktionale webbasierte Datenbankapplikation, in der man mit den eigenen Forschungsdaten und einer selbst angelegten Struktur, Visualisierungen und geographische Verknüpfungen herstellen kann. Die intuitive Benutzung von nodegoat macht es auch für Datenbank-Laien möglich, eine Datenbank mit grafischer Oberfläche zu erstellen.

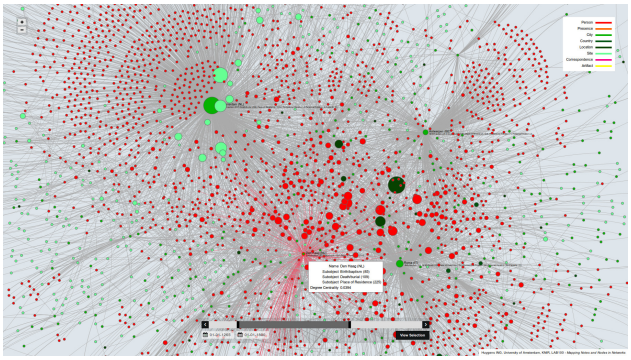
Zunächst werden einige laufende nodegoat Projekte vorgestellt. Anschließend erhalten die Workshop-Teilnehmer und Teilnehmerinnen die Möglichkeit, über ihre Forschungsprojekte zu berichten, für die Sie eine Datenbankapplikation verwenden möchten. Danach lernen die Teilnehmenden, wie Sie Daten in nodegoat eingeben können. Im letzten Teil des Workshops werden die Teilnehmerinnen und Teilnehmer die Möglichkeit haben, in zeitlich / thematisch / methodisch organisierten Gruppen eigene Projekte in nodegoat zu entwickeln. Die Workshop-Teilnehmenden sind herzlich willkommen, ihre eigenen Projektdatensätze mitzubringen.

Über nodegoat

nodegoat is a web-based data management, network analysis and visualisation environment. nodegoat allows scholars to build datasets based on their own data model and offers relational modes of analysis with spatial and diachronic contextualisation. By combining these elements within one environment, scholars are able to instantly process, analyse and visualise complex datasets relationally, diachronically and spatially; trailblazing. nodegoat follows an object-oriented approach throughout

its core functionalities. Borrowing from actor-network theory this means that people, events, artefacts, and sources are treated as equal: objects, and hierarchy depends solely on the composition of the network: relations. This object-oriented approach advocates the self-identification of individual objects and maps the correlation of objects within the collective.





Beispielprojekte

- GRIMMWELT (Kassel) Interactive Museum Installation : 20.000 letters of the Grimm Brothers visualised through time and space.
- Mapping Nodes and Nodes in Networks (Huygens Institute for the History of the Netherlands, University of Amsterdam, Royal Dutch Institute in Rome).
- SpInTime - Dynamically visualizing how cultural patterns, networks and exchanges evolve in space and time (University of Amsterdam), and ERNiE , the *Encyclopedia of Romantic Nationalism in Europe*.
- Memory Landscapes and the Regime Change of 1965-66 in Semarang (NIOD Institute for War, Holocaust and Genocide Studies, Semarang University, Radboud University Nijmegen).
- The transnational dynamics of social reform, 1840-1940 (Ghent University, Maastricht University), .
- Comparative transnational study of national movements (NISE).
- Diplomatic Letters 1683-1744 (Indonesian National Archive , Corts Foundation).

Vorbereitung

- Schauen Sie sich *nodegoat Video Tutorials* auf YouTube an https://www.youtube.com/watch?v=eLDRNiJrRUc&list=PLXc6y7l7xxxIwd64QppyAA0G2ECsNGJ
- Bitte lesen Sie den Blog-Beitrag *Enter, Curate & Explore Data* und die *nodegoat FAQ* :

Ablauf des Workshops

- Einführung in nodegoat und Vorstellung von nodegoat Projekten.
- Diskussion der Forschungsfragen, die von den Teilnehmer_innen vorgestellt werden.
- Dateneingabe in nodegoat (einschließlich, aber nicht beschränkt auf: relationales Datenmodell, geographische Visualisierung, soziale Netzwerk-Visualisierung)
- Feedbackrunde
- Break
- Diskussion und Inventarisierung von Forschungsprojekten.
- In Gruppen: Konzeption eines neuen Datenmodels in nodegoat.
- Abschlussdiskussion

Voraussetzungen

Die Teilnehmer und Teilnehmerinnen müssen ihren eigenen Laptop (mit einer aktuellen Version von Firefox, Chrome oder Safari) mitbringen.

Teilnehmerzahl

ca. 15 - 20

Lebensläufe

LAB1100 is a research and development firm established in 2011 by Pim van Bree and Geert Kessels. Their joint skill set in new media, history, and software development allows them to conceptualise and develop complex software applications. Working together with universities and research institutes, LAB1100 has built digital research platforms and interactive data visualisations.

Pim van Bree received his MA in New Media at the University of Amsterdam and his BA in Digital Imagineering at the NHTV University of Applied Sciences in Breda. He graduated with a thesis on the actor

network of transnational online dating, investigating the crossroads between the local, national, global, and the online assemblage. His work experience in the field of new media: digital strategist at Tribal DDB Amsterdam and software developer at KIWA.

Geert Kessels received his BA in History from Radboud University Nijmegen and completed the research master program in History at the University of Amsterdam. He graduated with a thesis on the influences of German Idealism on the Slovak romantic intellectual Ľudovit Štúr. During his studies he completed an internship at the Study Platform on Interlocking Nationalisms and worked as a project manager for EUROCLIO - The European Association of History Educators.

Wissenschaftliches Bloggen bei de.hypotheses.org

König, Mareike
mkoenig@dhi-paris.fr
Deutsches Historisches Institut Paris, Frankreich

Baillot, Anne
anne.baillot@gmail.com
Humboldt Universität Berlin

Einführung in das wissenschaftliche Bloggen

Wissenschaftliche Blogs haben ein hohes Potential für die schnelle Verbreitung und Diskussion aktueller Forschungsinhalte. Als neue Form der fachwissenschaftlichen Kommunikation nutzen Blogs die Möglichkeiten des Internets und des Web 2.0 für eine direkte und interaktive Publikation. Als öffentlich geführte wissenschaftliche Notizbücher eignen sich Blogs zur selbstkritischen Reflexion des eigenen Forschungsprozesses wie auch zur Dokumentation desselben. Nicht nur Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftlern bietet Bloggen die Möglichkeit, bereits in einem frühen Stadium auf ihr Projekt aufmerksam zu machen, mit erfahrenen Wissenschaftlerinnen und Wissenschaftlern in Austausch zu treten und sich zu vernetzen.

Mit Wissenschaftsblogs entwickelt sich rasant ein neues Genre, das bislang nicht im Methoden-Kanon und den bisherigen Reputationsmechanismen geistes- und sozialwissenschaftlicher Disziplinen vorgesehen war. Was genau bedeutet Bloggen für das akademische Schreiben und Publizieren? Wie verändert diese Kommunikationsform den wissenschaftlichen Alltag? Mit *de.hypotheses.org* wurde Anfang 2012 eine deutschsprachige Plattform für geistes- und sozialwissenschaftliche Blogs geschaffen, in deren Umfeld seither eine stetig wachsende deutschsprachige Community als Teil eines europäischen Netzwerks entstanden ist.

In einer Einführung werden zunächst konzeptionelle und wissenschaftliche Aspekte des Bloggens generell thematisiert und einige *best practice* Beispiele aus dem Bereich der Geistes- und Sozialwissenschaften vorgestellt. Anschließend beginnt der praktische Workshop mit Schulungsblogs für alle Teilnehmerinnen und Teilnehmer auf der Blogplattform de.hypotheses.org (Wordpress). Während des Workshops werden außerdem Tipps für die Anfangsphase eines wissenschaftlichen Blogs gegeben sowie rechtliche Belange erörtert.

Besondere technische Kenntnisse sind nicht notwendig. Die Teilnehmenden sollten ihre eigenen Laptops mitbringen.

Inhalte des Workshops

- Bloggen in den Geisteswissenschaftenfigurftsblog: Impressum, Über das Blog, Foto
- Veröffentlichung erster Beiträge: Artikel und Seiten
- Inhalte ordnen: Kategorien und Schlagworte, Kommentare verwalten
- Erweiterte Konfigurationsmöglichkeiten
- Veröffentlichung komplexer Inhalte (Einfügen verschiedener Objekte und Medien wie Videos)

Entwicklung und Nutzung interdisziplinärer Repositorien für historische textbasierte Korpora

Odebrecht, Carolin
carolin.odebrecht@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Lüdeling, Anke
anke.lueding@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Dreyer, Malte
malte.dreyer@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Zielke, Dennis
dennis.zielke@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Ziele des Workshops

Der Workshop setzt sich das Ziel, die Möglichkeiten, Aufgaben und Herausforderungen bei der Wiederverwendung von historischen Korpora zu identifizieren und zu diskutieren. Insbesondere sollen dabei deren Architektur, Dokumentation, Veröffentlichung und Speicherung betrachtet werden. So wollen wir versuchen, Methoden und Strategien für das interdisziplinäre Forschungsparadigma der Digital Humanities zu entwickeln und diese in den Fragestellungen der Konferenz der DHd 2016 zu verorten. Der Fokus wird auf die spezifischen und fächerübergreifenden Anforderungen historischer Texte in Bezug auf deren Aufbereitung und Speicherung in Repositorien zum Zweck der Wiederverwendung gelegt. Damit richtet sich der Workshop gleichermaßen an Korpuserstellende, Entwickelnde und an Betreiber_innen von Repositorien und deren Nutzer_innen.

Forschungsfragen und Kontextualisierung des Workshops

Historische Texte bilden den Forschungsgegenstand verschiedener geisteswissenschaftlicher Fächer wie der Linguistik, der Geschichtswissenschaft, der Literaturwissenschaft und vieler weiterer. Jede Disziplin hat dabei ihre eigenen Forschungsfragen und Arbeitsweisen, die sich in beispielsweise den genutzten Formaten und Annotationsweisen zeigen, wie beispielsweise die TEI Guidelines und deren TEI-XML-Format (TEI Consortium 2015) für digitale Editionen oder das Stand-Off-Format PAULA (Dipper 2005) für linguistische Korpora. Dennoch gibt es Ähnlichkeiten bei der Textauswahl und -aufbereitung, die eine gemeinsame Nutzung der vorhandenen Daten sinnvoll erscheinen lassen. In vielen Fällen wird zwischen den digitalisierten historischen Texten – den Primärdaten – und den hinzugefügten Interpretationen in Form von Metadaten und Annotationen unterschieden. Diese Begriffe – Primärdatum, Annotation, Metadatum – werden sowohl fachübergreifend aber auch innerhalb einer Disziplin oft sehr unterschiedlich genutzt. [1] Dadurch entsteht der Eindruck, dass die Daten der Disziplinen grundverschieden sind. Da sich die oft unterschiedlichen Korpusarchitekturen jedoch auf den gleichen Forschungsgegenstand, nämlich den gleichen Text, beziehen können, gibt es durchaus große Schnittmengen zwischen den verschiedenen Disziplinen.

Ein Beispiel für eine ähnliche Textauswahl sind historische Zeitungen, auf deren Grundlage ganz unterschiedliche Fragestellungen adressiert werden (für einen kleinen Überblick siehe bspw. Burr et al. 2015). Die korpusbasierten Aufbereitungsarten reichen bei diesem Register beispielsweise von digitalen Editionen (z. B. „ Korpus romanischer Zeitungssprachen “, Burr 1994-2007), über registerspezifische Referenzkorpora (z. B. „ German Manchester Corpus “, Bennett et al. 2008) bis hin zu syntaktisch tief annotierten Korpora (z. B. Mercurius Baumbank , Demske 2003-2005). Ein weiteres Beispiel sind historische Briefe, die als digitale Edition literaturwissenschaftlich untersucht werden (z. B. „Briefe und Texte aus dem intellektuellen Berlin um 1800“, Baillot / Seiffert 2013b), oder die als linguistisch annotierte Korpora zur Untersuchung von historischen Sprachständen dienen (z. B. „ Kasseler Junktionskorpus “, Ágel / Hennig 2007-2009).

Die Beantwortung der jeweiligen Forschungsfrage stützt sich dann häufig auf Interpretationen in Form von Annotationen in einem Korpus, deren Formen sich disziplinübergreifend ähneln können. Dennoch existieren vielfältige manuell zu erstellende oder automatisch generierbare Annotationsarten wie zum Beispiel Named-Entity-Recognition, Referenzierung auf Personendatenbanken (wie z. B. die Gemeinsame Normdatei [2]), GEO Tagging (vgl. z. B. Elliot / Gillies 2009), Lemmatisierung (z. B. Schmid 1994) oder syntaktisches Parsing (z. B. Malt Parser [3]), die die interpretatorischen Analysegrundlagen stellen. Da die Digitalisierung der historischen Texte (auf Grundlage von Handschriften, Drucken oder Editionen) und deren Annotation aufwändig und vielschichtig sind (vgl. u. a. Rissanen 2008, Kytö / Pahta 2012) kann die Wiederverwendung von historischen Korpora von Vorteil sein. Die Vorstellung der verschieden disziplinspezifischen Sicht- und Zugriffweisen auf solche textbasierten Daten (vgl. Pitti 2004) und deren Wiederverwendung ist ein zentraler Themenbereich des Workshops.

Damit diese heterogenen historischen Korpora von unterschiedlichen Disziplinen genutzt und wiederverwendet werden können, müssen sie über eine gemeinsame Plattform zugreifbar sein. Diese Plattform muss das Durchsuchen der Daten sowie der Metadaten, ggf. das Evaluieren sowie das Anreichern mit weiteren Annotationen und erneute Hochladen der Daten ermöglichen. Idealerweise können Repositorien diese Funktion übernehmen. Sie funktionieren dann wie eine Art Marktplatz, auf dem historische Korpora fachübergreifend ausgetauscht und mit Informationen angereichert werden können.

Der Workshop nimmt diesen Startpunkt, dessen Voraussetzungen und Konsequenzen zum Thema und begreift ihn als einen Teilbeitrag zu einem Fragenkomplex der DHd-Konferenz:

> „Was sind die Daten der Geisteswissenschaften? Wie müssen die Daten der Geisteswissenschaften (digitalisierte bzw. digitale Texte, Bilder, Musik, Audio, Filme / Videos etc.) aufgearbeitet und vorgehalten werden, um sie über die Fächer hinweg nicht nur für unterschiedliche, sondern auch derzeit

noch unbekannte Fragestellungen nutzen zu können?“ [4]

Der Workshop versucht für historische Korpora zu ergründen, wie und welche Wiederverwendungsszenarien unter welchen Voraussetzungen möglich sind und was der aktuelle Stand der Forschung ist. Dabei ist es enorm wichtig, dieses Thema vielschichtig und aus mehreren Perspektiven zu beleuchten. Fallstudien für die Wiederverwendung historischer Daten können exemplarisch Erfahrungen, Herausforderungen und Aufgaben thematisieren. Anhand von Korpusarchitekturen, die die Wiederverwendung unterstützen, können wichtige Konzepte und Modelle diskutiert und verglichen werden. Die Beschreibung von konkreten Technologien für die Umsetzung eines Repositoriums erlaubt es, die theoretischen Datenmodelle auf ihre Praxistauglichkeit zu untersuchen. Die Nutzer dieser Technologien tragen durch ihre Erfahrungen über die potentiellen Vorteile der Wiedernutzung und die Bereiche, in denen sie Sinn machen, maßgeblich zur Diskussion bei.

Damit stellen sich folgende Fragen in Bezug auf die historischen Korpora, deren Aufbereitung, die Repositorien bzw. Technologien und deren Nutzung:

- Können dieselben Primärdaten unter verschiedenen Forschungsfragen unterschiedlich genutzt werden?
- Welche Gemeinsamkeiten, welche Unterschiede weisen die Korpora hinsichtlich ihrer umfangreichen Aufbereitung historischer Texte auf.
- In wie weit fördern / erschweren die Annotationen als theoretische Konzepte und Interpretationen eine Wiederverwendung?
- Welche Arten von Annotationen und Analysen können wie wiederwendet werden?
- Welche Arten der Wiederverwendung können sich ergeben?
- Wie unterschiedlich bewerten Disziplinen die Qualität eines Korpus?
- Welche interdisziplinären Nutzer- und Nutzungsszenarien ergeben sich?
- Welche Anforderungen ergeben sich hinsichtlich der Korpusarchitektur inklusive Annotationsarten und Format?
- Welche Speicherformate eignen sich für die Wiederverwendung von Forschungsdaten?
- Wie können Lizenzen den Austausch und die Wiederverwendung fördern?
- Was sind die relevanten Metadaten über ein Korpus?
- Welche Art von Zugriff auf die Korpora ist notwendig, um eine Wiederverwendung zu erleichtern? Wie müssen Repositorien beschaffen sein?
- Welche Vor- und Nachteile besitzen disziplinspezifische / interdisziplinäre oder / und formatabhängige oder -unabhängige Repositorien?
- Eine Diskussion und mögliche Beantwortung dieser Fragen wollen wir durch einen fächerübergreifenden Austausch von Entwicklern, Korpuserstellern und Nutzern im Rahmen des Workshops ermöglichen.

Form des Workshops

Der Workshop soll bestehend aus zwei impulsgebenden Keynotes und sechs Vorträgen an einem Tag vor der DHd-Konferenz stattfinden. Eine Keynote wird das Thema des interdisziplinären Zugangs und der Wiederverwendung zu historischen Daten allgemein thematisieren und problematisieren (Lüdeling und Dreyer, Projektleiter des LAUDATIO -Repositoriums für historische Texte.

Eine zweite Keynote wird die Frage nach einem Qualitätsmanagement im Rahmen von Wiederverwendungsszenarien, dessen Umfang und Zweck aufwerfen und diskutieren (Laurent Romary, DARIAH Director .

Die Vorträge, die aus dem offenen Call des Workshops hervorgehen, sollen Korpuserstellende, Repositorienentwickler_innen und -nutzer_innen aus ganz verschiedenen Fachbereichen die Gelegenheit geben, die oben aufgeworfenen Fragen aufzunehmen und aus einer notwendigerweise interdisziplinären Sicht die Möglichkeiten, Herausforderungen und Lösungen für die Wiederverwendung von historischen Korpora zu diskutieren. Die Keynotes erhalten je 30 Minuten Redezeit sowie 15 Minuten Diskussion und die Vorträge je 20 Minuten und 10 Minuten Diskussion. Eine Diskussion wird den Workshop abschließen. Die primäre Sprache des Workshops ist Deutsch.

Notes

1. Zur Diskussion über Primärdatum, Transkriptionen, Normalisierungen siehe bspw. Claridge 2008, Himmelmann 2012, Kramer 2014; über Metadaten siehe bspw. Odebrecht 2015, Haynes 2004; über Annotationen siehe bspw. Lüdeling 2011, Kübler / Zinsmeister 2015).
2. „Die Gemeinsame Normdatei (GND) ist eine Normdatei für Personen, Körperschaften, Konferenzen, Geografika, Sachschlagwörter und Werktitel, die vor allem zur Katalogisierung von Literatur in Bibliotheken dient, zunehmend aber auch von Archiven, Museen, Projekten und in Webanwendungen genutzt wird.“ (DNB 2015).
3. Tool zum automatischen Annotieren von syntaktischen Dependenzen (Hall et al. 2014).
4. Siehe den Call for Papers DHd 2016 http://www.dhd2016.de/node/9 [letzter Zugriff: 12.September 2015].

Bibliographie

Ágel, Vilmos / Hennig, Mathilde (2007-2009): *KAJUK* (Version 1.1). Justus-Liebig-Universität Gießen http://www.uni-giessen.de/kajuk/index.htm , http://hdl.handle.net/11022/0000-0000-2102-8 [letzter Zugriff 10. September 2015].

Baillot, Anne / Seifert, Sabine (2013a): „The Project "Berlin Intellectuals 1800–1830" between Research and Teaching“ in: *Journal of the Text Encoding Initiative* 4. DOI: 10.4000/jtei.707.

Baillot, Anne / Seifert, Sabine (2013b): *Briefe und Texte aus dem intellektuellen Berlin um 1800* http://tei.ibi.hu-berlin.de/berliner-intellektuelle/ [letzter Zugriff: 22.Dezember 2015].

Bennett, Paul / Durrell, Martin / Ensslin, Astrid / Scheible, Silke / Whitt, Richard (2008): *GerManC Project* (Version 1.0), University of Manchester http://www.llc.manchester.ac.uk/research/projects/germanc/ , http://hdl.handle.net/11022/0000-0000-2D1B-1 [letzter Zugriff 10.September 2015].

Burr, Elisabeth (1994-2007): *Korpus Romanischer Zeitungssprachen*. Duisburg - Bremen - Leipzig http://www.uni-leipzig.de/~burr/CorpusLing/ [letzter Zugriff 10. Sepzember 2015].

Burr, Elisabeth / Burkhardt, Julia / Potapenko, Elena / Sierig, Rebecca / Concepción Durán, Arámis (2015): „Das Duisburg-Leipzig Korpus romanischer Zeitungssprachen und sein Textmodell“, in: *Von Daten zu Erkenntnissen. 2. Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum*, DHd 2015, Graz. https://www.conftool.pro/dhd2015/index.php/Burr-Das_Duisburg-Leipzig_Korpus_romanischer_Zeitungssprachen-1771016.pdf?page=downloadPaper&filename=Burr-Das_Duisburg-Leipzig_Korpus_romanischer_Zeitungssprachen-1771016.pdf&form_id=177 [letzter Zugriff: 11.September 2015].

Claridge, Claudia (2008): “Historical Corpora” in: Lüdeling, Anke / Kytö, Merja (eds.): *Corpus Linguistics. An International Handbook.* Volume 1. Berlin: De Gruyter 242–259.

Demske, Ulrike (2003-2005): *Mercurius* (Version 1.1). Universität Potsdam http://www.uni-potsdam.de/guvdds/projekte/abgproj.html , http://hdl.handle.net/11022/0000-0000-467D-6 [letzter Zugriff: 10.September 2015].

Dipper, Stefanie (2005): “XML-Based Stand-Off Representation and Exploitation of Multi-Level Linguistic Annotation”, in: *Proceedings of Berliner XML Tage Berlin* 39-50.

DNB = Deutsche Nationalbibliothek (2015): *Gemeinsame Normdatei (GND)* http://www.dnb.de/gnd [letzter Zugriff 10.September 2015].

Elliott, Tom / Gillies, Sean (2009): “Digital Geography and Classics. Changing the Center of Gravity”, in: *Digital Humanities Quarterly* 3, 1 http://www.digitalhumanities.org/dhq/vol/3/1/000031/000031.html [letzter Zugriff 10. September 2015].

Hall, Johan / Nilsson, Jens / Nivre, Joakim (2014): *MaltParser* http://www.maltparser.org/ [letzter Zugriff 10. September 2015].

Haynes, David (2004): *Metadata for information management and retrieval*. London: Facet publishing.

Himmelmann, Nikolaus. P. (2012): “Linguistic Data Types and the Interface between Language Documentation and Description”, in: *Language Documentation and Conservation* 6: 187-207.

Kramer, Michael J. (2014): “Defining Data for Humanists: Text, Artifact, Information or Evidence?”, in: *Journal for Digital Humanities* 3, 2 http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/ [letzter Zugriff: 10. September 2015].

Kübler, Sandra / Zinsmeister, Heike (2015): *Corpus Linguistics and Linguistically Annotated Corpora*. London / New York: Bloomsbury.

Kytö, Merja / Pahta, Päivi (2012): “Evidence from historical corpora up to the twentieth century” in: Nevalainen, Terttu / Traugott, Elizabeth C. (eds.): *The Oxford Handbook of the History of English.* Oxford o.a.: Oxford University Press 123-133.

Lüdeling, Anke (2011): “Corpora in Linguistics: Sampling and Annotation” in: Grandin, Karl (ed.): *Going Digital*. Evolutionary and Revolutionary Aspects of Digitization. Nobel Symposium 147. New York: Science History Publications 220-243.

Odebrecht, Carolin (2015): „Interdisziplinäre Nutzung von Forschungsdaten mithilfe einer technisch-abstrakten Modellierung“, in: *Von Daten zu Erkenntnissen. 2. Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum* DHd 2015, Graz. https://www.conftool.pro/dhd2015/index.php/Odebrecht-Interdisziplin%C3%A4re_Nutzung_von_Forschungsdaten_mithilfe_einer_technisch-abstrakten_Modellierung-63110.pdf?page=downloadPaper&filename=Odebrecht-Interdisziplin%C3%A4re_Nutzung_von_Forschungsdaten_mithilfe_einer_technisch-abstrakten_Modellierung-63110.pdf&form_id=63 [letzter Zugriff 12.September 2015].

Pitti, Daniel V. (2004): “Designing Sustainable Projects and Publications” in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A Companion to Digital Humanities.* Oxford: Blackwell 471–487.

Rissanen, Matti (2008): “Corpus Linguistics and Historical Linguistics” in: Lüdeling, Anke / Kytö, Merja (eds.): *Corpus Linguistics. An International Handbook.* Volume 1. Berlin: Mouton de Gruyter 53-68.

Schmid, Helmut (1994): “Probabilistic Part-of-Speech Tagging Using Decision Trees“, in: *Proceedings of International Conference on New Methods in Language Processing*, 1994. Manchester.

TEI Consortium (eds.) (2015): *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 2.8.0. 2015-04-06.* TEI Consortium. http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 11. August 2015].

Crossmediales Publizieren mit TUSTEP – ein Workshop der International TUSTEP User Group

Recker-Hamm, Ute
recker@uni-trier.de
Akademie der Wissenschaften und der Literatur Mainz

Schneider, Matthias
schneiderm@uni-trier.de
Universität Trier, Trier Center for Digital Humanities

Einleitung

Geisteswissenschaftliche Projekte – von der studentischen Studienarbeit bis hin zu groß angelegten Forschungsverbünden – sind heute mit der Anforderung konfrontiert, ihre Arbeitsergebnisse nicht nur in einer einzigen Publikationsform (z. B. in gedruckter Form) zu veröffentlichen, sondern vielmehr auch in diversen elektronischen Formaten (z. B. für Ebook-Reader und / oder als dynamische Webseiten). Dabei kann das Verhältnis der verschiedenen Publikationsformen zueinander ganz verschieden sein: Während die Ebook- oder PDF-Varianten oftmals elektronische Abbilder der Druckfassung mit überschaubarem eigenen Mehrwert (z. B. Volltextsuche, Bookmarks usw.) sind, können dynamische Publikationsformen sehr weit über die Druckfassungen hinausgehen, indem sie mehr Inhalte (z. B. die Grunddaten oder alternative Lesefassungen) und zusätzliche Funktionen anbieten. Bei wissenschaftlichen Publikationen darf erwartet werden, dass die verschiedenen Fassungen zitierfähig und aufeinander bezogen sind.

Der Siegeszug von XML / TEI als Auszeichnungsformat auch in den geisteswissenschaftlichen Fächern hat die Grundlage für das "Single-Source"-Prinzip bereitet, indem es durch inhaltsbezogenes, nicht layout-spezifisches Markup die Voraussetzung dafür schafft, aus einer einzigen Datenquelle mithilfe geeigneter Werkzeuge mehrere Ziel-Publikationsformen zu bedienen. Neben den bekannten XML-Standard-Techniken wie XSLT, XQuery und XSL:FO stehen Werkzeuge für die Verarbeitung von XML-Daten nach dem Single-Source-Prinzip zur Verfügung, die speziell geisteswissenschaftliche Arbeitsweisen und Anforderungen unterstützen, darunter das Tübinger System von Textverarbeitungsprogrammen (TUSTEP).

Der Workshop *Crossmediales Publizieren mit TUSTEP* wird anhand eines überschaubaren, konkreten Beispiels einer Briefedition zeigen, wie XML-Daten in TUSTEP für die Print-, die Ebook- und die dynamische HTML-Publikation verarbeitet werden können. Er richtet sich an Teilnehmerinnen und Teilnehmer, die sich über die allgemeinen Methoden des Single-Souce-Prinzips informieren und erfahren wollen, wie sie in TUSTEP umgesetzt werden. Vorausgesetzt werden grundlegende Kenntnisse im Umgang mit XML-Daten. Im Workshop werden sich präsentierende Abschnitte, in denen die Möglichkeiten von TUSTEP in diesem Bereich vorgestellt werden, mit Anteilen abwechseln, in denen die Teilnehmerinnen und Teilnehmer die Bearbeitung von XML-Daten mit TUSTEP selbst erproben können.

Das Tübinger System von Textverarbeitungsprogrammen

Das Tübinger System von Textverarbeitungsprogrammen ist eines der ältesten, seit über 40 Jahren bis heute beständig weiterentwickelten und in zahlreichen geisteswissenschaftlichen Projekten (für eine Auswahl siehe itug). Es bietet ein umfangreiches Paket von speziell auf philologisch-sprachwissenschaftliche Anforderungen zugeschnittenen, aufeinander abgestimmten Werkzeugen, die in hohem Maß an projektspezifische Erfordernisse angepasst werden können. Die Bandbreite der Programmmodule reicht dabei vom Vergleichen und Kollationieren, dem Segmentieren, Annotieren und Indexieren von Texten in XML oder anderen Formaten über das Skripting bis hin zum professionellen Satz und kann – beispielsweise für dynamische Webpublikationen – auch direkt auf Servern eingesetzt werden. Seit Kurzem verfügt TUSTEP über eine XML-basierte, alternative Kommandosyntax namens TXSTEP , die eine Steuerung des Programms von außerhalb, z. B. aus der XML-Entwicklungsumgebung Oxygen erlaubt.

Die International TUSTEP User Group e.V. (ITUG) unterstützt TUSTEP-Anwender und -Lerner seit über 20 Jahren durch Kurse, jährlich stattfindende Fach-Kolloquien (und das TUSTEP-Wiki . Außerdem fördert sie die Weiterentwicklung von TUSTEP, seit es seit 2011 als Open Source Software frei und kostenlos verfügbar ist.

Workshop-Inhalte

Überblick über TUSTEP

Zu Beginn des Workshops wird ein kurzer orientierender Überblick über die Programmbausteine und die Arbeitsweise von TUSTEP gegeben, wobei

insbesondere die einschlägigen Hilfsmittel vorgestellt werden.

XML-Daten im TUSTEP-Editor

Die Teilnehmerinnen und Teilnehmer unternehmen anhand des Beispieltexts ihre ersten Schritte in TUSTEP, indem sie sich den Editor für die Bearbeitung von XML-Daten einrichten. Sie erfahren, wie sie Kommandos und einfache Pattern-Matching-Anweisungen ausführen.

TEI-Beispieltexte setzen (PS- / PDF-Herstellung)

Die in TEI ausgezeichneten Beispieltexte werden mithilfe des schnell zu erschließenden und einfach zu benutzenden, dabei dennoch auf eine hohe Qualität abzielenden TUSTEP-Moduls *SATZ für die Druckstufe (PS / PDF) vorbereitet. Nach einer Vorstellung der Funktionsweise und der Einstellungsmöglichkeiten können die Workshopteilnehmenden die PDF-Ausgabe nach eigenem Belieben variieren.

Von TEI nach EPUB: TUSTEP im Rahmen von Toolchains

Für die Herstellung eines Ebooks im EPUB-Format wird aus den TEI-Daten zunächst mit der TUSTEP-eigenen Skriptsprache TUSCRIPT eine HTML-Datei erzeugt, die anschließend mit dem Open source-Konvertierer Calibre nach EPUB transformiert wird. Neben der Erläuterung und gemeinsamen Anwendung der verschiedenen Komponenten soll in diesem Bereich deutlich werden, auf welche Weise TUSTEP und externe Tools gemeinsam verwendet und automatisiert genutzt werden können.

Einsatz von TUSTEP als Publikationsserver

Abschließend werden die zuvor vorgestellten Verarbeitungsmöglichkeiten für XML-Daten auf einem Webserver eingesetzt und gemeinsam mit den Teilnehmenden getestet. Diese Zusammenführung soll insbesondere die Stärken von TUSTEP im Rahmen dynamischer Online-Projekte (z. B. Editionen mit fortlaufend aktualisierten Daten) aufzeigen.

Organisatorisches

Technische Ausstattung

Alle Teilnehmenden sollten ein eigenes Notebook (Linux, MAC OS X, Windows) mitbringen. Die nötige Software (TUSTEP und weitere benötigte Komponenten) werden von den Veranstaltern auf USB-Sticks mitgebracht. Hierzu wird vorab eine Informationsmail an die angemeldeten Workshopbesucher versendet. – Für den Workshop werden ein Beamer und WLAN-Zugang für alle Teilnehmenden benötigt.

Zahl der möglichen Teilnehmenden

Die Zahl der Teilnehmenden sollte auf max. 20 Personen begrenzt sein, um eine angemessene, interaktive Durchführung gewährleisten zu können.

Zeitplanung

Der Workshop ist auf zweimal 90 Minuten Dauer zuzüglich einer etwa 30-minütige Pause konzipiert.

Kontaktdaten

Ute Recker-Hamm, M.A.
Akademie der Wissenschaften und der Literatur Mainz
Arbeitsstelle »Mittelhochdeutsches Wörterbuch« an der Universität Trier
54286 Trier
www.uni-trier.de/index.php?id=19336
Forschungsinteressen: Datenverarbeitung für philologische und lexikographische Zwecke, Digitale Editionen, Native XML-Datenbanken
Zweite Vorsitzende der International TUSTEP User Group

Matthias Schneider, M.A.
Universität Trier
Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften
54286 Trier
http://kompetenzzentrum.uni-trier.de/ / www.m-schneider.eu
Forschungsinteressen: Digitale Geisteswissenschaften, strukturelle Textdatenverarbeitung sowie politische, militärische und Historiographiegeschichte der Neueren und Neuesten Zeit

Visualisierungs-methoden und -instrumente für historische Quellenkorpora

Schrade, Torsten
Torsten.Schrade@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz

Andreas, Kuczera
Andreas.Kuczera@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz

Thomas, Kollatz
kol@steinheim-institut.org
Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte

Kurzbeschreibung

Im Verlauf der letzten 10 Jahre hat die Menge an digital verfügbaren, fachwissenschaftlich annotierten Volltexten für die historische Forschung stark zugenommen. Damit einher geht auch eine Veränderung sowohl der Nutzungsformen digitaler Quellen als auch der Möglichkeiten der historischen Arbeitsweise. Bestand um die Jahrtausendwende noch enger Kontakt zwischen Historiker_innen und Quellen, nimmt dies mit zunehmender Digitalisierung perspektivisch ab. Hat der / die Forscher_in früher die für seine Forschungsfragen relevanten Quellen in der Regel alle mindestens einmal gelesen, scheint dies bei den heute recherchierbaren Mengen an digitalen Quellen kaum noch möglich. Ein Hauptproblem ergibt sich hier aus der Schnittstelle zwischen Forscher_innen und den im Netz erreichbaren Quellendatenbanken. Die Suchinterfaces der Datenbanken sind oft für die Nutzung durch Expert_innen des jeweiligen Materials optimiert. Dies ist auf der einen Seite zu begrüßen, da sie den Fachwissenschaftler_innen damit besten Zugriff auf das Material gewähren. Daneben sollten aber weitere Zugriffsmöglichkeiten für übergreifende Text-Mining- oder Big-Data-Recherchen bereitgestellt werden, mit denen verschiedene Quellenkorpora parallel im Hinblick auf übergreifende Fragestellungen untersucht werden können.

Neben Such- bzw. sammlungsorientierten Zugriffen auf die Daten wird die Fähigkeit, mittels bestimmter Visualisierungsmethoden und -instrumente neue Zusammenhänge in den Daten zu erkennen und diese dann für die historische Analyse zu nutzen immer wichtiger. Insbesondere im Bereich der graphorientierten Visualisierungsmethoden ist im Moment ein regelrechter Boom an Softwarebibliotheken und Online-Tools zu beobachten. Auch in den einschlägigen Forschungsumgebungen bzw. Forschungsinfrastrukturen für die Geisteswissenschaften wie TextGrid und DARIAH werden zunehmend Visualisierungsinstrumente für annotierte Fachdaten eingebunden.

Im Workshop zweier Partnerinstitutionen aus dem DARIAH-DE Cluster "Fachwissenschaftliche Annotationen" (Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte / Akademie der Wissenschaften und der Literatur Mainz, Digitale Akademie) soll es mit konkretem Praxisbezug um die Potentiale, Methoden und Instrumente zur Visualisierung von historischen Fachdaten gehen. Als Anwendungsbeispiel dienen die historischen Fachdatenrepositorien, die beide Partner in den Workshop mit einbringen (bspw. Epidat – epigraphische Datenbank; die Deutschen Inschriften Online ; die Regesta Imperii Online).

Visualisiert werden zunächst klassische Fragestellungen, wie beispielsweise die nach bestimmten Familienbeziehungen, nach Orts- oder Zeitbezügen in den Daten. Genutzt werden Instrumente wie beispielsweise die Graphdatenbank neo4j, das Netzwerkvisualisierungsinstrument Gephi, aber auch JavaScript-Tools auf Basis von sigma.js , dracula oder auch d3 . Der Workshop wird schrittweise vorgehen. Ein grundlegendes Verständnis für TEI/XML, JSON, RDF sowie JavaScript-Webtechnologien ist für die Teilnahme am Workshop hilfreich, aber nicht zwingend. Nach einer allgemeinen Einführung in den Bereich des fachwissenschaftlichen Annotierens wird es um praktische Beispiele gehen, die in den Daten vorhandenen Annotationen für verschiedene Visualisierungsinstrumente aufzubereiten und dann die jeweiligen Visualisierungen zu erzeugen. In kurzen Impulsreferaten werden sich die Workshop-Teilnehmer_innen die gemeinsam erarbeiteten Visualisierungen und die Instrumente, mit denen diese Visualisierungen erzeugt worden sind, gegenseitig vorstellen.

Im Fazit soll der Workshop ein Bewusstsein und auch schon erste Fähigkeiten entwickeln, sich mit Hilfe von Visualisierungsinstrumenten neue Sichten auf das Quellenmaterial und somit neue Forschungsperspektiven zu eröffnen. Deutlich werden wird aber auch, dass dieser Ansatz nicht automatisch zu einer „Antwort-Maschine“ führt, die dem / der Wissenschaftler_in die interpretative Arbeit abnimmt. Vielmehr können sich durch Visualisierungen neue Interpretationsmöglichkeiten des Quellenmaterials bieten, die vorher einfach auf Grund der Datenmasse nicht sichtbar gemacht werden konnten.

Weiterführende Literatur:

Kuczera, Andreas (2015): *Graphdatenbanken für Historiker*. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi http://mittelalter.hypotheses.org/5995 .

Schrade, Torsten (2013): "Datenstrukturierung", in: Frietsch, Ute / Rogge, Jörg (eds.): *Über die Praxis des kulturwissenschaftlichen Arbeitens*. Ein Handwörterbuch. Bielefeld: transcript 91–97.

Teilnehmerzahl

10 - 15 Personen

Benötigte Ausstattung

- WLAN-Zugang
- Beamer
- Workshop-Teilnehmer sollten ihre eigenen Laptops mitbringen

Es geht auch ohne Formeln – der Einsatz von TeX in den Digital Humanities am Beispiel kritischer Editionen

Sievers, Martin
sievers@uni-trier.de
Universität Trier, Deutschland

Einleitung

Die Diskussion rund um digitale Editionen als Ergänzung oder sogar Ersatz für die klassische Buchausgabe ist in vollem Gange. Grund dafür ist auch der Siegeszug der plattform- und implementationsunabhängigen Metasprache XML in den „Digital Humanities“. Insbesondere der TEI-Standard (Burnard / Bauman 2007) hat dafür gesorgt, dass Informationen aller Art in einem Sammelformat vorliegen, das von vielen neu entwickelten Werkzeugen als Ausgabe- und Austauschformat verwendet wird.

Der weitere Bearbeitungsprozess bis hin zur Fertigstellung einer Edition fußt daher heutzutage stark auf XML. Gleichwohl bleibt das Buch als notwendiges Ergebnis eines Editionsprojekts weiterhin die Regel. Somit stehen viele Wissenschaftler zu Beginn eines solchen Projekts vor dem Problem, einen Workflow auf XML-Basis zu definieren, der am Ende möglichst komfortabel – mit oder ohne Zutun eines Verlags – auch einen hochwertigen Buchdruck erlaubt.

Projekte wie *XML-Print* oder *Apache FOP* setzen hier an und wollen das Textsatzproblem innerhalb der „X-Technologien“ [1] lösen. Es ist in den vergangenen Jahren jedoch deutlich geworden, dass der wissenschaftliche Textsatz von diesen Werkzeugen (noch) nicht in all seinen Facetten erfasst werden kann. Daher greifen aktuelle Editionsprojekte in der Regel auf etablierte Werkzeuge wie *TUSTEP* oder andere nicht notwendigerweise XML-basierte Ansätze zurück, indem der XML-Eingabetext mittels XSLT in die entsprechenden Zwischenformate überführt wird.

Genau hier möchte der Workshop ansetzen und die Möglichkeiten des Textsatzsystems TeX [2] im Bezug auf die Erstellung einer historisch-kritischen Ausgabe (kritische Edition) vorstellen. Leider wird dieser Weg aus Unwissenheit bzw. wegen falscher oder schlicht veralteter Informationen bzgl. des Funktionsumfangs viel zu selten beschritten. Umso erfreulicher sind Forschungs- und Arbeitsumgebungen wie *FuD* oder *ediarum* , die am Ende des editorischen XML-basierten Workflows eine mit TeX erzeugte PDF-Datei zur Kontrolle bzw. als Vorstufe des Druckergebnisses ausgeben.

Funktionsweise von LaTeX

Allgemein

Das quelloffene Textsatzsystem TeX und die gleichnamige Programmiersprache wurden Ende der Siebziger Jahre vom amerikanischen Mathematikprofessor Donald E. Knuth für den Druck seiner eigener Bücher entwickelt. Das Problem des Textsatzes – „Wie bringe ich unter Beachtung verschiedener Regeln möglichst schön Zeichen aller Art aufs Papier?“ – wurde von ihm als mathematisches Optimierungsproblem definiert und mit neuartigen Algorithmen gelöst. Die *subjektive* Schönheit wurde dadurch von Knuth auf Basis typographischer Traditionen und Methoden *objektiviert*.

Die so entstandenen Algorithmen, z. B. derjenige für den Zeilenumbruch (Knuth / Plass 1981) waren bahnbrechend und sind bis heute „State of the Art“. Entsprechend werden sie auch in jüngerer Software wie *Adobe InDesign* oder *Apache FOP* nahezu unverändert verwendet. Für den Autor eines Texts bedeutet dies, dass er sich vollständig auf die inhaltliche bzw. strukturelle Gestaltung konzentrieren kann. Dies entspricht der Arbeit mit XML-Quelldaten, die in der Regel keinerlei typographische Anweisungen enthalten.

Somit lebt die klassische Trennung zwischen Autor und Setzer wieder auf, die durch Text*verarbeitungs*programme in den vergangenen Jahrzehnten stückweise aufgeweicht worden ist – mit negativen Folgen für die Druckqualität. In heutigen digitalen Arbeitsumgebungen entspricht der Setzer einem Satzprogramm, das eine Druckvorlage – heutzutage oft „hinter den Kulissen“ einer virtuellen Arbeitsumgebung – auf die Quelldokumente eines Autors anwendet.

Anwendungsfall Kritische Edition

Eine kritische Edition stellt besondere Anforderungen an ein Textsatzwerkzeug. Daher eignet sich dieser Dokumenttyp besonders gut, um die Qualität von LaTeX zu demonstrieren. Plachta definiert für eine kritische Edition zehn elementare Bestandteile (Plachta

2006: 14-15). Diese lassen sich zu den folgenden drei Themenkomplexen zusammenfassen:

- Edierter Text,
- Apparate,
- Verzeichnisse.

Zu all diesen Bereichen liefert LaTeX entweder direkt oder über Erweiterungen (Pakete) Möglichkeiten, hochwertige Ergebnisse zu erzielen. Sie stehen über das zentrale Paketarchiv CTAN [3] allen Nutzern kostenfrei zur Verfügung.

Inhalte des Workshops

Entlang der im vorherigen Abschnitt genannten Bestandteile einer kritischen Edition wird der Workshop den Teilnehmern die Gelegenheit geben, LaTeX als Satzprogramm kennenzulernen bzw. vorhandene Kenntnisse auszubauen. Im Detail werden die folgenden Inhalte vermittelt:

- Edierter Text: Neben den grundlegenden Satzalgorithmen werden mikrotypographische Fragen thematisiert. Dazu gehören neben der Nutzung typographisch korrekter Zeichen (z. B. bei Anführungszeichen, Gedankenstrich oder Ellipse) auch der automatische optische Randausgleich oder die Laufweitenänderung, die beide durch das Paket *microtype* bereitgestellt werden. Mehr zum Thema Mikrotypographie findet man bei Beinert (2015) .
- Apparate: Die Anforderungen an den Satz von Apparaten gehen weit über diejenigen „normaler“ Fußnoten hinaus. Es wird das Paket *reledmac* vorgestellt und neben verschiedenen Anpassungen auch Lösungen für Probleme wie z. B. überlappende Lemmata erarbeitet.
- Verzeichnisse: Eine kritische Edition enthält neben einem Quellenverzeichnis verschiedene Register. Diese sollen im Idealfall direkt aus den Druckdaten dynamisch erzeugt werden. Dazu werden die Erweiterungen *biblatex* sowie *imakeidx* vorgestellt, die genau dies bewerkstelligen.

Teilnehmerkreis / Technische Ausstattung

Der Workshop richtet sich an alle interessierten Wissenschaftler_innen, die Wert auf einen hochwertigen Druck ihrer Edition legen. Auch Entscheidungsträger, die ein Textsatzsystem für ihr Editionsprojekt suchen, sind ausdrücklich angesprochen. Für die praktischen Beispiele werden grundlegende Kenntnisse der Textsatzsprache LaTeX vorausgesetzt.

Da es sich um eine „Hands-on“-Sitzung handelt, sollte die Teilnehmerzahl 15 nicht übersteigen, wobei ich eine Warteliste begrüßen würde. Die Teilnehmerinnen und Teilnehmer benötigen einen Laptop mit einer (möglichst aktuellen) TeX-Distribution (MiKTeX oder TeX Live). Für die Präsentation selbst wird ein Beamer benötigt.

Notes

1. Unter den X-Technologien versteht man die W3C-Standards XML, XSL und XPath sowie je nach Kontext weitere im XML-Umfeld entstandene Sprachen und Formate wie XQuery oder XLink.
2. TeX wird seit langem schon nur noch selten direkt angewendet, sondern in der Regel über eine Makrosprache wie LaTeX (siehe z. B. http://www.latex-project.org) oder ConTeXt (siehe z. B. http://wiki.contextgarden.net/What_is_ConTeXt) angesprochen. Der Workshop konzentriert sich auf LaTeX.
3. Das *Comprehensive TeX Archive Network* stellt über zwei zentrale Server und mehr als hundert Spiegelserver (Mirrors) weltweit unter http://www.ctan.org insgesamt über 4500 Erweiterungen bereit.

Bibliographie

Beinert, Wolfgang (2015): *Typolexikon.de*. Das Lexikon der europäische Typographie http://www.typolexikon.de/m/mikrotypographie.html [letzter Zugriff 05. Februar 2016].

Berlin-Brandenburgische Akademie der Wissenschaften (2010): *ediarum*. Eine digitale Arbeitsumgebung für Editionsvorhaben http://www.bbaw.de/telota/software/ediarum [letzter Zugriff 05. Februar 2016].

Burnard, Lou / Bauman, Syd (2007): *TEI P5. Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ [letzter Zugriff 05. Februar 2016].

Knuth, Donald E. / Plass, Michael F. (1981): “Breaking paragraphs into lines”, in: *Journal Software: Practice and Experience* 1 0.1002/spe.4380111102.

Plachta, Bodo (^{2}2006): *Editionswissenschaft: eine Einführung in Methode und Praxis der Edition neuerer Texte* (= Reclams Universal-Bibliothek 17603). Stuttgart: Philipp Reclam jun.

Trier Center for Digital Humanities (TCDH) / Forschungszentrum Europa (FZE) (2014): *FuD*. Eine virtuelle Forschungsumgebung für die Geisteswissenschaften. Universität Trier http://fud.uni-trier.de/de/ [letzter Zugriff 05. Februar 2016].

TextGrid und DARIAH-DE: Forschungsumgebung und Infrastruktur für die Geisteswissenschaften

Vanscheidt, Philipp
vanscheidt@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Rapp, Andrea
rapp@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Schmid, Oliver
oschmid@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Schmunk, Stefan
schmunk@sub.uni-goettingen.de
Niedersächsische Staat- und Universitätsbibliothek Göttingen, Deutschland

Kollatz, Thomas
kol@steinheim-institut.org
Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte Essen, Deutschland

TextGrid [1] ist eine vom Bundesministerium für Bildung und Forschung (BMBF) geförderte virtuelle Forschungsumgebung für Geisteswissenschaftler_innen (TextGrid Konsortium 2006-2014). Sie ermöglicht die Zusammenarbeit von Wissenschaftler_innen an verschiedenen Standorten und nicht zuletzt die Archivierung von Forschungsdaten in einem eigenen Repositorium. TextGrid ist in die digitale Forschungsinfrastruktur DARIAH-DE integriert. DARIAH-DE wird ebenfalls vom BMBF gefördert, mit dem Ziel, für die Geistes- und Kulturwissenschaften in Deutschland eine nachhaltige digitale Forschungsinfrastruktur bereitzustellen, die aus den Säulen, Lehre, Forschung, Forschungsdaten und technische Infrastruktur besteht. Seit 2014 wird eine Vielzahl von Entwicklungen gemeinsam vorangetrieben und zugleich nutzt TextGrid einige der Komponenten von DARIAH-DE. TextGrid setzt sich aus dem TextGrid Repository, einem digitalen Forschungsdaten-XML-Langzeitarchiv, und dem TextGrid Laboratory zusammen.

Das TextGrid Laboratory dient als Einstiegspunkt in die virtuelle Forschungsumgebung mit Werkzeugen und Diensten in einer anpassbaren Software, in die weitere digitale Arbeitsmittel und eigene Tools integriert werden können. Eine Vielzahl von Werkzeugen im Laboratory erlaubt das Arbeiten mit Texten und Bildern, aber auch beispielsweise mit Noten und Digitalisaten. Wichtige Komponenten sind:

Ein frei verfügbarer XML-Editor erlaubt das Bearbeiten von XML-Dateien, Anwender_innen können dabei beliebig zwischen vier Ansichten wechseln: (i) einer Ansicht des XML-Baumes, in den Knoten ergänzt und in dem Elemente verändert werden können, (ii) einer Ansicht des Quelltextes, (iii) einer Ansicht, die eher an der Darstellung eines Textverarbeitungsprogramms orientiert ist und die über CSS an die eigenen Bedürfnisse angepasst werden kann sowie (iv) einer Vorschau, die über XSLT eine HTML-Ansicht erstellt. Eine Unicode-Zeichentabelle ermöglicht das einfache Suchen, Kopieren und Einfügen beliebiger Symbole aus dem Unicode Zeichensatz. Alternativ kann auch der oXygen XML-Editor in die Umgebung eingebunden werden.

Ein Text-Bild-Link-Editor unterstützt den XML-Editor bei der Alignierung von Text und Bildelementen. Ziel ist die Erstellung einer Ausgabedatei, die die Textelemente und die topographische Position von rechtwinkligen und polygonen Bildbereichen in SVG miteinander verknüpft, wie dies zum Beispiel bei der Verbindung von Faksimiles und Transkriptionen in kritischen Editionen der Fall ist. Auch können Bilder auf diese Weise zum Beispiel im Rahmen kunsthistorischer Untersuchungen annotiert werden.

Das Bildbetrachtungs- und Referenzierungstool DigiLib wurde in die Umgebung integriert und umfasst die Galerieansicht mehrerer Bilder, Zoom, Skalierungs-, Markierungs- und Referenzfunktionen.

Mit dem Noten-Editor MEISE können Notentexte in MEI graphisch kodiert, bearbeitet und auf einem einfachen Niveau auch dargestellt werden. So wird unter anderem die Visualisierung von Varianten ermöglicht.

Eine Projekt- und Nutzerverwaltung (Datei- und Rechtemanagement) ermöglicht die Erstellung neuer Projekte durch Projektleiter_innen und die Zuordnung von weiteren Benutzer_innen in bestimmten Rollen zu einem Projekt, sowie die Abfrage und das Setzen von Rechten an TextGrid-Objekten.

Ein Metadaten-Editor dient dem Erstellen und Verwalten von Metadaten der TextGrid-Objekte. Diese Metadaten werden in TextGrid für projektübergreifende Recherchen verwendet. Das Eingabeformular für die Metadaten kann einfach an die individuellen Bedürfnisse angepasst werden.

Eine Wörterbuchsuche ermöglicht die Suche in einer Vielzahl von verschiedenen Wörterbüchern innerhalb der virtuellen Forschungsumgebung von TextGrid. Hierzu wurde das Trierer Wörterbuchnetz in das TextGridLab integriert.

Das Publizieren (im Repository) wird unterstützt durch eine automatische Metadatenvalidierung. Ferner wurde die Software SADE der Berlin-Brandenburgischen Akademie der Wissenschaften als „skalierbare Architektur für digitale Editionen“ in TextGrid eingebunden, um eigene Webportale für die Publikation gestalten zu können.

Ergänzt wird dieses Spektrum durch Werkzeuge und Dienste, um die für ein bestimmtes Thema oder Forschungsinteresse relevanten Objekte auswählen, bündeln, verwalten, importieren und exportieren zu können. Revisionen einzelner Objekte lassen sich speichern und auch einzeln publizieren.

Das TextGrid Repository, das gemeinsam mit DARIAH-DE weiterentwickelt und betrieben wird, dient als digitales Langzeitarchiv für die Geisteswissenschaften, um die langfristige Verfügbarkeit und Zugänglichkeit der Forschungsdaten zu garantieren und über umfangreiche Suchmöglichkeiten, verschiedene Download-Formate und Visualisierungstools zugänglich und über ein Projekt hinaus nutzbar zu machen.

In dem geplanten Workshop soll insbesondere interessierten Fachwissenschaftler_innen die Möglichkeit gegeben werden, die Vielfalt der virtuellen Forschungsumgebung TextGrid und die Werkzeuge der digitalen Forschungsumgebung DARIAH-DE zu erkunden und digitale Analysemöglichkeiten kennenzulernen. Zunächst soll in einer kurzen Einführung ein Einblick in die Geschichte des Projektes, die verschiedenen Komponenten und Erweiterungen der Software sowie das digitale Arbeiten mit TextGrid gegeben werden. Auch sollen Beispiele von Projekten gezeigt werden, die mit der Forschungsumgebung arbeiten. Nachdem auf diese Weise eine Übersicht über das Spektrum der Werkzeuge und Services gegeben worden ist, können die Teilnehmer_innen in parallel stattfindenden Vorführungen auf mehreren „Inseln“, abhängig von ihrem Interesse, die verschiedenen Möglichkeiten der Forschungsumgebung mit Mitarbeiter_innen in halbstündigen Exkursionen ausprobieren und sich dabei von Thema zu Thema, von Insel zu Insel fortbewegen. Dabei sollen Fragen nicht zu kurz kommen. Wer die einzelnen Etappen am eigenen Rechner ausprobieren möchte, muss die Software zuvor installiert und sich registriert haben:

https://textgrid.de/registrierung/download (für die Registrierung)

https://textgrid.de/download (für den Download der Software)

Die Software steht für Linux, MacOS und Windows zur Verfügung. Sie benötigt eine aktuelle Java-Version. Für Unterstützung bei der Einrichtung stehen die Mitarbeiter_innen im Vorfeld des Workshops zur Verfügung. Die Teilnehmerzahl ist auf 30 begrenzt.

Ausstattung

Für die verschiedenen „Inseln“ werden Beamer und Leinwände oder andere Projektionsflächen benötigt. Die Anzahl an Inseln richtet sich auch nach den diesbezüglichen Möglichkeiten. Die Bestuhlung sollte sowohl Vorträge für alle Teilnehmer gemeinsam als auch die Aufteilung in kleine Gruppen („Inseln“) oder einen schnellen Umbau erlauben. Für eine Posterpräsentation wären zwei bis drei Stellwände oder Möglichkeiten praktisch, Poster an der Wand anzubringen.

Kontakte

Thomas Kollatz
Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte
Edmund-Körner-Platz 2, 45127 Essen
Tel.: +49 201 82162910
E-Mail: kol@steinheim-institut.org
Forschungsschwerpunkte: Digital Humanities, Deutsche Literatur in hebräischen Lettern um 1800, Jüdisch-deutsche Presse, Orthodoxie vor 1871

Prof. Dr. Andrea Rapp
Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft
Dolivostraße 15, 64293 Darmstadt
Tel.: +49 6151 16 57408
Fax: +49 6151 16 57411
E-Mail: rapp@linglit.tu-darmstadt.de
Forschungsinteressen: Digitale Paläographie und Kodikologie, Digitale Edition, Annotationen

Dr. Oliver Schmid
Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft
Dolivostraße 15, 64293 Darmstadt
Tel.: +49 6151 16 57452
Fax: +49 6151 16 57411
E-Mail: oschmid@linglit.tu-darmstadt.de
Forschungsinteressen: Nutzerfreundlichkeit in virtuellen Forschungsumgebungen, quantitative Kodikologie.

Dr. Stefan Schmunk
Niedersächsische Staat- und Universitätsbibliothek Göttingen
Papendiek 14, 37073 Göttingen
Tel.: +49 551 39-20326, +49 551 39-19777
E-Mail: schmunk@sub.uni-goettingen.de
Forschungsschwerpunkte: Empirisch-deduktive Methoden, Forschungsinfrastrukturen und Forschungsdaten

Philipp Hegel, geb. Vanscheidt
Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft
Dolivostraße 15, 64293 Darmstadt
Tel.: +49 6151 16 57405
Fax: +49 6151 16 57411
E-Mail: hegel@linglit.tu-darmstadt.de
Forschungsschwerpunkte: Deutschsprachige Literatur seit 1800, Editionsphilologie, Forschungsumgebungen und quantitative Kodikologie.

Notes

1. Ausschreibungen der einzelnen Komponenten und der Geschichte des Projektes finden sich in Neuroth et al. (2015). Eine digitale Fassung ist verfügbar unter https://textgrid.de/projektdokumentation .

Bibliographie

DARIAH-DE. Digital Research Infrasructure for the Arts and Humanities. Göttingen https://de.dariah.eu/ .

Neuroth, Heike / Rapp, Andrea / Söring, Sibylle (eds.) (2015): *TextGrid: Von der Community — für die Community*. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften. Glückstadt: Hülsbusch.

TextGrid Konsortium (2006–2014) *TextGrid* Virtuelle Forschungsumgebung für die Geisteswissenschaften. Göttingen: TextGrid Konsortium textgrid.de .

Panels

Nachhaltigkeit technischer Lösungen für digitale Editionen. Eine kritische Evaluation bestehender Frameworks und Workflows von und für Praktiker_innen

Andorfer, Peter
peter.andorfer@oeaw.ac.at
ACDH, Österreich

Durco, Matej
matej.durco@oeaw.ac.at
ACDH, Österreich

Stäcker, Thomas
staecker@hab.de
HAB, Deutschland

Thomas, Christian
thomas@bbaw.de
BBAW, Deutschland

Hildenbrandt, Vera
hildenbr@uni-trier.de
TCDH, Deutschland

Stigler, Hubert
hubert.stigler@uni-graz.at
ZIM Uni Graz, Österreich

Söring, Sibylle
soering@sub.uni-goettingen.de
SUB Göttingen, Deutschland

Rosenthaler, Lukas
lukas.rosenthaler@unibas.ch
DH Lab, Schweiz

Stand der Dinge

Digitale Editionen machen in den Digital Humanities das „Brot- und Buttergeschäft“ aus. Doch während sich der methodisch-theoretische Hintergrund digitaler Editionen zusehends konsolidiert und sich diese neue Form der Publikation von Forschungsergebnissen im (fach)wissenschaftlichen Diskurs bereits etabliert hat, fehlt es nach wie vor an umfassend dokumentierten und selbstkritisch reflektierten Best-Practice-Beispielen von Frameworks und Workflows zur Erstellung und / oder Publikation von digitalen Editionen, welche als Blaupausen für künftige digitale Editionsprojekte herangezogen werden können. Das Resultat ist so bekannt wie unerfreulich und kann – nur geringfügig überspitzt – auf folgende Formel gebracht werden: So gut wie jedes Projekt erfindet das Rad – das technische Grundgerüst der Edition – wieder neu.

Die wichtigsten Gründe für diese Entwicklung lassen sich rasch benennen:

- Digitale Editionen sind nach wie vor eine sehr junge Publikationsform und darüber hinaus abhängig von der raschen Weiterentwicklung gegenwärtiger Webtechnologien und -standards.
- Digitale Editionen erfordern technische Kompetenzen, welche jene traditioneller Geisteswissenschaftler_innen meist übersteigen, weshalb Kooperationen mit Entwickler_innen notwendig sind.
- Digitale Edition werden häufig in Form von Einzelprojekten realisiert, weshalb nicht auf bestehende Lösungsansätze anderer Institutionen zurückgegriffen wird.
- In so gut wie allen Fällen fehlen die Ressourcen, manchmal wohl leider auch der Wille, die Eigenentwicklung, das geschaffene technische Grundgerüst hinreichend gut zu dokumentieren und in einer Form zu veröffentlichen, so dass andere den Quellcode, Workflows, Stylesheets etc. nachnutzen zu können.
- Digitale Editionen sind – wenigstens in der Selbstwahrnehmung – häufig hochgradig speziell und unterscheiden sich in Inhalt, Form und Funktion von allen bereits bestehenden Editionsprojekten, womit eine völlige Neuentwicklung des technischen Grundgerüstes gerechtfertigt wird.

Ablauf:

Im Rahmen des Panels sollen einige der aktivsten Institutionen aus dem Bereich der digitalen Editionen an einen Tisch gebracht werden. Diese erhalten im Vorfeld der Tagung einen Fragebogen zur Vorbereitung einer kurzen (pro Teilnehmer ca. fünf Minuten) Vorstellung ihrer Systeme, wobei darin der Fokus auf dem Thema Reusability der in den Projekten verwendeten Technologien und Workflows liegen sollte. Konkret sollen die Teilnehmer_innen des Panels auf folgende Punkte eingehen:

- Kurze Vorstellung der eigenen Frameworks und Workflows, vor allem hinsichtlich einer Einschätzung

über die Stärken und Schwächen der eigenen Lösungsansätze, aus welcher Tradition/Disziplin (z. B. Philologie, Geschichtswissenschaften) sie kommen und welche konkreten Projekte damit realisiert wurden.
- Gibt es ein weitgehend standardisiertes Prozedere im Falle von Kooperations- bzw. Nachnutzungsanfragen (inklusive der dafür notwendigen Ressourcen)?
- Wer soll die Angebote nutzen; gibt es fachlich, institutionell, qualitativ, budgetär, regional, national, zeitlich oder anderweitig konstituierte Zielgruppen?

Ein Ziel dieser Vorstellungsrunde soll es sein, potenziell interessierten Nutzer_innen im Auditorium einen kompakten Überblick über bestehende Angebote zur Erstellung und / oder Veröffentlichung von digitalen Editionen zu vermitteln.

Podiumsdiskussion (ca. 30 Minuten)

Im Anschluss an diese Kurzvorstellung erfolgt eine moderierte Podiumsdiskussion, worin folgende Punkte weiter thematisiert werden:

Dokumentation von Technologie und Workflows: Wie gut sind die technischen Aspekte dokumentiert? Ist es für Dritte möglich, anhand dieser Dokumente ähnliche Projekte zu realisieren? Welche technische Infrastruktur ist dafür notwendig?

Veröffentlichung von Code: Ist der für das System geschriebene oder adaptierte Code für andere nachnutzbar veröffentlicht (z. B. auf GitHub)?

Wird das entwickelte Framework auch als Service angeboten?

Ressourcen und Organisation der Entwicklung: Wie groß war / ist der Aufwand der Entwicklung des technischen Grundgerüstes, eventuell in Personenmonaten. Woher stammt das verwendete Know-How (Eigenentwicklung oder Adaption bestehender Konzepte)?

Wie generisch ist das verwendete technische Framework bzw. wie aufwändig sind die Änderungen, die bei der Adaption an ein neues Projekt notwendig werden? Sprich: Wie leicht und wie weit kann das System an projektspezifische Bedürfnisse angepasst werden (unterstützte Datenformate, Funktionalitäten)? Welche Kompetenzen sind notwendig, um Anpassungen auf unterschiedlichen Ebenen (Anzeige, Projektstruktur usw.) vorzunehmen?

Wie groß ist der laufende Aufwand für Wartung einzelner Projekte bzw. des Frameworks an sich?

Es sollen gemeinsame Problemfelder identifiziert und reflektiert werden. Auf dieser Basis kann dann über mögliche (gemeinsame) Lösungen diskutiert werden.

Publikumsdiskussion (ca. 30 Minuten)

Im letzten Drittel des Panels wird die Diskussion zum Publikum hin geöffnet werden. Dabei sollen vor allem potentielle Nutzer_innen die Möglichkeit bekommen, gezielt konkrete und ggf. eigene Projekte betreffend Fragen zu stellen und direkt mit möglicherweise zukünftigen Projektpartnern ins Gespräch zu kommen.

Teilnehmer

Bei der Auswahl der Teilnehmer wurde einerseits darauf geachtet, vornehmlich etablierte Institutionen anzusprechen, die sich als Dienstleister im Bereich digitaler Editionen profiliert haben, deshalb an möglichst generischen Lösungen zur Erstellung und Publikationen von digitalen Editionen interessiert sind und dafür selbst Frameworks und Workflows entwickelt haben. Außerdem wurde versucht, bei der Auswahl der Teilnehmer möglichst den gesamten deutschsprachigen Raum abzudecken.

ACDH-ÖAW

Matej Durco und Peter Andorfer

Das ACDH verwendet ein eXistdb-basiertes Framework zur Veröffentlichung digitaler Editionen names cr-xq-mets . cr-xq-mets basiert lose auf SADE. Die Idee von cr-xq-mets ist die konsequente Trennung von Code und einzelnem Projekt, mit dem Ziel einen hohen Grad an generischer Projektentwicklung bei gleichzeitig geringem Aufwand an Projektmaintainance zu erreichen.

Das ACDH übernimmt auch die Organisation und Moderation des Panels.

Herzog August Bibliothek

Thomas Stäcker

Die Herzog August Bibliothek (HAB) hat für ihre digitalen Editionsprojekte die Reihe Editiones Electronicae Guelferbytanae gegründet), in der bisher 20 Werke erschienen sind. Hinzu kommen zahlreiche kleinere und umfangreichere Editionen, die außerhalb dieser Reihe erschienen sind. Hervorzuheben sind die beiden im Langzeitförderprogramm der DFG (je 12 Jahre) erscheinenden Editionen der Kritische Gesamtausgabe der Schriften und Briefe Andreas Bodensteins von Karlstadt und die Digitale Edition und Kommentierung der Tagebücher des Fürsten Christian II. von Anhalt-Bernburg (1599-1656) . Die Kodierung erfolgt in TEI (nach den übergreifenden Festlegungen der HAB). Die

Anzeige der Webdarstellung und Suchfunktionaltäten basieren auf PHP sowie eXistdb.

TextGrid

Sibylle Söring

TextGrid ist eine Virtuelle Forschungsumgebung für die text- und quellenbasierten Geisteswissenschaften, die u. a. die Erstellung digitaler Editionen mithilfe Open Source-basierter Tools und Dienste unterstützt. Neben der Software, dem TextGrid Laboratory, bietet TextGrid mit dem TextGrid Repository die Möglichkeit, vielfältige Forschungsdaten - u. a. XML / TEI-kodierte Texte, Bilder und Datenbanken - langfristig zu speichern sowie nach internationalen Standardformaten zitierbar zu publizieren und zur Nachnutzung zur Verfügung zu stellen, wie etwa zur Recherche und Visualisierung. Mit TextGrid wurden und werden verschiedene Editionsvorhaben umgesetzt, so u. a. die Digitale genetisch-kritische Edition von Theodor Fontanes Notizbüchern und die Bibliothek der Neologie .

BBAW – DTA, CLARIN-D

Christian Thomas

Das DFG-geförderte Projekt Deutsches Textarchiv der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) erstellt bzw. publiziert digitale Editionen von derzeit (September 2015) 1#665 Werken im DTA-Kernkorpus unter Verwendung eines selbst entwickelten, auf Open-Source-Software basierenden Frameworks. Zu diesem Framework gehört mit dem Modul (DTA-Erweiterungen ein elaborierter Workflow zur Integration hochwertiger Textressourcen aus externen Editions- und Forschungsprojekten. Über DTAE konnten als Ergänzungen des DTA-Kernkorpus bislang weitere 1#097 Einzelwerke sowie der gesamte Bestand zweier herausragender Zeitschriften des 19. / 20. Jahrhunderts, das von J. G. Dingler begründete Polytechnische Journal (346 Bände, 1820–1931) und die Zeitschrift *Die Grenzboten* (270 Bände, 1841–1922) in das DTA integriert werden. Der Textbestand aus DTA und DTAE umfasst ca. 200 Mio. Tokens und ca. 1,2 Mrd. Zeichen. Die DTA-Korpora sind einheitlich gemäß dem TEI-basierten und ausführlich dokumentierten DTA-Basisformat (DTABf) kodiert und werden mit Hilfe computerlinguistischer Werkzeuge automatisch annotiert, was unter anderem spezifizierte Suchanfragen nach bestimmten Metadatenfeldern, Wortarten, grammatischen Kategorien, X-Pfaden etc. ermöglicht. Zudem wird der historische Textbestand automatisch in Richtung moderner Orthographie ‘normalisiert’, was schreibweisentolerante Suchen über das gesamte Korpus ermöglicht (siehe allgemein zur Suche im DTA www.deutschestextarchiv.de/doku/DDC-suche_hilfe). Die Qualitätssicherung sämtlicher Korpusressourcen geschieht kollaborativ in der webbasierten Umgebung DTAQ , in der derzeit 866 registrierte Nutzer mögliche Fehler auf der Text-, Annotations- und Metadatenebene melden und, je nach Rechtestatus, auch direkt online korrigieren können.

Trier Center for Digital Humanities

Thomas Burch, Vera Hildenbrandt

Das TCDH kann inzwischen auf eine mehr als langjährige Erfahrung in der Planung, Durchführung und Betreuung von Projekten im Bereich der Digital Humanities verweisen. Neben dem Schwerpunkt im Bereich der Erstellung und Publikation digitaler Editionen verfügt das Zentrum über eine ausgewiesene Expertise in der Entwicklung von Software-Umgebungen für geisteswissenschaftliche Großvorhaben. In mehreren von der DFG, dem BMBF sowie im Rahmen des Akademienprogramms geförderten Projekten entstanden und entstehen am TCDH digitale Editionen wie z. B. das Heinrich-Heine-Portal , das Christian-Dietrich-Grabbe-Portal , das Cusanus-Portal , die elektronische Publikation der Korrespondenz August-Wilhelm Schlegels sowie die digitale Rekonstruktion der Textgenese und Entstehungsgeschichte von Wolfgang Koeppens ‚Jugend'. Im Bereich der Softwareentwicklung konzipiert, betreut und entwickelt das Team des TCDH virtuelle Arbeitsumgebungen für Projekte mit hohen Anforderungen an Workflow und grafische Benutzerschnittstellen. Hervorgehoben seien hier Systeme wie das internetbasierte Artikelredaktionssystem für die Produktion und Publikation von Wörterbüchern in dezentralen Arbeitsstellen, das gemeinsam mit dem Forschungszentrum Europa und dem Sonderforschungsbereich 600 entwickelte " Forschungsnetzwerk und Datenbanksystem " (FuD 2015), die Redaktions- und Publikationsplattform zur Europäischen Geschichte Online oder das interaktive Werkzeug “ Transcribo” zur Erstellung von Transkriptionen.

ZIM – ACDH Graz

Hubert Stigler

Das ZIM hat im Rahmen einer Vielzahl von Editionsprojekten forschungsgetrieben ein objektorientiertes Framework auf Basis von FEDORA Commons und weiteren Open-Source-Projekten (Apache Cocoon, Blazegraph u. a.) entwickelt, das Aspekte der Publikation von Digitalen Editionen mit jenen der Langzeitarchivierung von Forschungsdaten zu verbinden sucht. Als österreichischer Beitrag zu DARIAH steht es einer breiten Öffentlichkeit zur Nachnutzung zur Verfügung.

DH Lab Basel

Lukas Rosenthaler

Während die anderen Teilnehmer vor allem an Lösungen für XML / TEI-basierte digitale Editionen arbeiten, legt SALSAH (System for Annotation and Linkage of Sources in Arts and Humanities) den Schwerpunkt (zurzeit noch) auf die Verknüpfung und Verlinkung von vornehmlich digitalen Faksimiles.

Der Modelle Tugend 2.0 – Vom digitalen 3D-Datensatz zum wissenschaftlichen Informationsmodell

Hoppe, Stephan
email@stephan-hoppe.de
Ludwig-Maximilians-Universität München

Pfarr-Harfst, Mieke
pfarr@dg.tu-darmstadt.de
Technische Universität Darmstadt

Münster, Sander
sander.muenster@tu-dresden.de
Technische Universität Dresden

Kuroczyński, Piotr
piotr.kuroczynski@herder-institut.de
Herder-Institut Marburg

Blümel, Ina
ina.bluemel@tib.uni-hannover.de
Technische Informationsbibliothek Hannover

Hauck, Oliver
oha@raumdarstellung.org
Institut für Raumdarstellung

Lutteroth, Jan
j.lutteroth@googlemail.com
Technische Universität Darmstadt

Die *Arbeitsgruppe Digitale Rekonstruktion* ging aus der *1. Jahrestagung der Digital Humanities im deutschsprachigen Raum* (25.-28.02.2014, Universität Passau) hervor und vereint zurzeit 43 Mitglieder aus 27 Forschungsinstitutionen aus Deutschland, Österreich und der Schweiz.

Das Panel *Pecha Kucha – Virtuelle Rekonstruktion* versammelte in Passau 2014 Kolleginnen und Kollegen, die sich dem Thema aus dem Blickwinkel der Architektur, Archäologie, Bau- und Kunstgeschichte sowie Computergraphik und Informatik verschrieben haben.

Die Gründungsmitglieder nutzten die Gelegenheit in Passau, um eine Plattform in Form der Arbeitsgruppe für einen engeren Austausch und eine feste Etablierung der digitalen hypothetischen 3D Rekonstruktion vom Kulturerbe innerhalb der Digital Humanities einzurichten.

Während der *2. Jahrestagung der Digital Humanities im deutschsprachigen Raum* an der Universität Graz wurde das Ziel einer gemeinsamen Publikation für Ende 2016 zum Thema der Digitalen 3D Rekonstruktion im Kontext der geisteswissenschaftlichen Forschung vorgestellt.

Beim 3. Arbeitstreffen der *Arbeitsgruppe Digitale Rekonstruktion* im Herder-Institut zu Marburg am 11.09.2015 wurde das Konzept und die inhaltliche Strukturierung intensiv besprochen. Die Publikation soll den heutigen Standpunkt der digitalen 3D Rekonstruktion beleuchten und richtungsweisende Perspektiven aufzeigen, wie die digitalen 3D-Datensätze sowie ihre vielfältigen Anwendungen wissenschaftlichen Standards entsprechen können.

Die *Arbeitsgruppe Digitale Rekonstruktion* möchte für die *3. Jahrestagung der Digital Humanities im deutschsprachigen Raum* in Leipzig ein weiteres Panel unter dem Titel *Der Modelle Tugend 2.0 – Vom digitalen 3D-Datensatz zum wissenschaftlichen Informationsmodell* anbieten.

Hierzu laden wir vier Vortragende jeweils zu vier eruierten und neu strukturierten Inhalten der geplanten Publikation ein. Damit wollen wir die lebendige Diskussion und die Zwischenergebnisse der Arbeit an der Publikation in die *Digital Humanities im deutschsprachigen Raum* hineintragen.

Was ist eine digitale 3D-Rekonstruktion?
Stephan Hoppe

Die dreidimensionalen Computermodelle im Kontext des materiellen und immateriellen kulturellen Erbes sind Wissensträger, in denen objektimmanentes Wissen meist in Bezug auf nicht mehr vorhandene Gebäude oder Siedlungsstrukturen gespeichert wird. Doch was ist eine digitale 3D-Rekonstruktion tatsächlich, wie kann man diese Modelle definieren und was befähigt sie, ein Träger des Wissens zu sein? Der Vortrag beleuchtet die Typologie digitaler 3D-Rekonstruktion als eigenständige Forschungs- und Vermittlungsmethodik im Kontext von Cultural Heritage und zeigt die damit verbundenen grundlegenden Herausforderungen auf. Ausgehend von der Einführung in die Kulturgeschichte dieser Disziplin und unter besonderer Betrachtung der Bild- und Technikgeschichte sowie der Projektgeschichte und den damit verbundenen Meilensteine wird eine Historie digitaler 3D-Modellierung skizziert. Grundlage ist dabei die generalistisch-theoretische Betrachtungsweise, die neben der Verortung dieser neuen Disziplin im Bereich

der Digital Humanities auch die Notwendigkeit von einheitlichen Glossaren oder Thesauri aufgreift. Ein weiterer Fokus des Vortrags liegt auf der Betrachtung der Eigenschaften digitaler Rekonstruktionen, den daraus resultierenden Potentialen und Anwendungsmöglichkeiten bezogen auf die Anwendungsfelder der 3D-Modelle. So sind die digitalen dreidimensionalen Modelle eine neue Ausdrucksform und ein neues Medium, um unser Wissen darzustellen. Dabei haben sich einerseits vielfältige Anwendungsgebiete und Verwendungskontexte in den letzten Jahren herausgebildet, andererseits fehlen noch immer einheitliche Standards und Strukturen.

Alle diese Themenfelder sind eng miteinander verwoben und müssen in Abhängigkeit miteinander betrachtet werden. Unter dem Aspekt, die dreidimensionalen Computermodelle als Wissensträger zu verstehen, zeigt der Vortrag den grundlegenden Handlungsbedarf systematisch auf und skizziert erste Ideen und Visionen in diesem Kontext.

Welche Vermittlungs- und Darstellungsformen bietet sie an?

Oliver Hauck

Im Spannungsfeld von Illusion und Immersion existiert ein breites Spektrum von Darstellungs- bzw. Publikationsformen digitaler 3D-Rekonstruktionen, die sich nicht nur in Ihrer Perzeption, sondern auch in ihren technischen Anforderungen sowie im Erstellungsaufwand massiv unterscheiden. Einleitend soll ein Überblick über die verschiedenen Darstellungsmöglichkeiten und die ihnen immanenten Eigenschaften und Anforderungen gegeben werden. Es soll der Frage nachgegangen werden, ob und in wie fern die Visualisierungsmethode Einfluss auf die Modellerstellung selbst hat und wo die Grenzen zwischen wissenschaftlich fundierbaren, nachvollziehbar zu machenden Entscheidungen und solchen rein künstlerischer Natur liegen. Eine Frage, die im wissenschaftlichhen Diskurs bisher erstaunlich geringe Beachtung fand, dreht sich der Diskurs momentan hauptsächlich um Fragen der Nachvollziehbarkeit der Modellierung (überwiegend dabei die Frage der Verknüpfung des Modells und seiner Quellen).

Zuletzt soll die Frage diskutiert werden, ob die allgemein als Desideratum angesehenen Standards für digitale 3D Rekonstruktionen überhaupt gebraucht werden, oder ob nicht die Forderung nach Werkzeugen zur wissenschaftlichen Diskurseinbindung, mithin also zur Herstellung der Zitierbarkeit der Modelle und ihrer Visualisierungen, insbesondere im Hinblick auf den rückwirkenden Umgang mit bestehenden 3D Modellen und Visualisierungen wichtiger ist, da dies den Rückgriff auf die existierenden Standards wissenschaftlicher Arbeit erlaubt.

Welche Methoden wendet sie an?

Mieke Pfarr-Harfst, Sander Münster

Die Frage nach der Wissenschaftlichkeit einer Disziplin ist unmittelbar mit dem Status Quo ihrer Methodologie in Bezug auf die vorhandenen Prozesse und Arbeitsweisen sowie den partizipierenden Disziplinen verbunden. Die derzeit existierende internationale wissenschaftliche Community zur digitalen 3D-Rekonstruktion im wissenschaftlichen Kontext greift vor allem Perspektiven der Archäologie sowie der Cultural Heritage Conservation auf. Dabei nutzen digitale 3D-Rekonstruktionen nicht nur Technologien aus der Informatik zur Bearbeitung geisteswissenschaftlicher Fragestellungen, sondern inkorporieren darüber hinaus eine Vielzahl unterschiedlicher disziplinärer Perspektiven und Verwendungskontexte. Der Impulsvortrag möchte die aktuelle Diskussion in der wissenschaftlichen Community in Bezug auf die angewendeten Methoden und die jeweils am Prozess beteiligten Disziplinen beleuchten. Vor diesem Hintergrund stellt sich die Frage: Welches sind Forschungs- und Nutzungsansätze digitaler Rekonstruktion und wie spiegeln sich diese in einer Methodologie wider? Nutzungsansätze umfassen neben der Darstellung historischer Gebäude und Objekte im Kontext der Wissensvermittlung auch die Erforschung von historischen Planungs- und Konstruktionsprozessen, die Kontextualisierung und Prüfung der Konsistenz von Quellen, die Klassifikation von Objekten und die Identifikation von Schemata.

Im Rahmen des Vortrags sollen zunächst generelle Typologien und Abläufe digitaler 3D-Rekonstruktion sowie damit verbundene wissenschaftsmethodische Anforderungen dargestellt werden. Darüber hinaus soll die wissenschaftliche Einbettung, also der Zweck, die Anwendungsmöglichkeiten, die Disziplinabhängigkeit, die Nutzungskultur, die Akzeptanz und Etablierung sowie die Wissenschaftlichkeit beleuchtet werden. Die Potentiale der digitalen Rekonstruktionen wie Generierung, Sicherung und Fusionierung von Wissen sollen im Kontext des methodologischen Vorgehens und typischer Arbeitsabläufe aufgezeigt werden

Wie soll das Wissen organisiert werden?

Piotr Kuroczyński, Ina Blümel

Semantic-Web-Technologien, z. B. eine aussagenbasierte Strukturierung von Daten in RDF-Tripeln und Anwendung von sogenannten Ontologien scheinen auf breiter Front einen Vormarsch in die Digital Humanities zu wagen. Die Potenziale von Linked Open Data, insbesondere der Verknüpfung von Wissen über Bereichsgrenzen hinweg zu einem globalen Graphen für die Wissensdomäne, ziehen immer mehr Forschungsprojekte in den Bann.

Was verbirgt sich hinter RDF, Ontologien und ihren NameSpaces? Welche Motivation steckt hinter der in Ferne auftauchenden Idee künstlicher Intelligenz? Wie wird das Wissen im Web 3.0 strukturiert und organisiert? Welche Forschungsinfrastrukturen werden der Wissenschaft in diesem Kontext einen angemessenen Dienst leisten?

Der Impulsvortrag möchte einen Einblick in diese Themen innerhalb der Domäne der digitalen 3D-Rekonstruktion mit ihren 3D-Datensätzen geben und auf die domänenspezifischen Möglichkeiten, Chancen

aber auch die Herausforderungen von Semantic-Web-Technologien eingehen.

Die Komplexität der Aufgaben stellt geisteswissenschaftlich grundierte Projekte vor neue Herausforderungen, unter anderem die Entwicklung und Anwendung von adäquaten Thesauri oder die Referenzierung auf den ISO-Standard 21127:2006 (CIDOC-CRM). Anhand beispielhafter laufender Projekte im Bereich der Rekonstruktion von barocken Schlössern und der Archivierung von Architekturmodellen soll ein Einblick in Strategien zur Wissensorganisation hinsichtlich der Interoperabilität, der Verfügbarkeit sowie zu den Applikationsontologien gegeben werden. Des Weiteren wird eine virtuelle Forschungsumgebung mit einem idealtypischen Interface skizziert, sowie ausgewählte Features und das Backend für die Erfassung der Quellen, der Aktivität einer 3D-Computerrekonstruktion und Annotation von 3D-Objekten sowie die Publikation der 3D-Daten mit ihren „digitalen Fußnoten“ (Meta- und Paradaten) konzipiert.

Die Rolle des Zeigens

Kepper, Johannes
kepper@edirom.de
Musikwissenschaftliches Seminar Detmold / Paderborn, Deutschland

Ein zentrales Merkmal der Digital Humanities ist die Verwendung von Codierungen, um verschiedenste fachwissenschaftliche Inhalte und Forschungsgegenstände digitalen Methoden und Fragestellungen zugänglich zu machen. Vor allem im Bereich Digitaler Editionen werden diese Codierungen häufig durch Faksimiles ergänzt, die einen Zugriff auf das „Original“ bieten sollen. Die hier vorgeschlagene Panel Session geht der Frage nach, in welchem Verhältnis Codierung und Faksimile in verschiedenen Disziplinen zueinander stehen.

Literaturwissenschaftliche Editionen haben i. d. R. den Vorteil, dass die einzelnen Zeichen weitgehend dem Zeichenvorrat der Schriftsprache entnommen werden und sich damit innerhalb des gleichen Zeichensystems als Codierung erfassen lassen. Diese Nähe zwischen Codierung und Codiertem veranlasste die (missverständliche) Rede von der "Rekonstruktion" der Handschrift durch deskriptive Apparate. Dennoch wurde in der Folge die Unersetzlichkeit von Faksimiles verfochten. Die jüngeren Entwicklungen in der digitalen Editorik haben die Möglichkeiten der Erschließungstiefe auf der Ebene der Textkodierung stark erweitert und zugleich den Faksimile-Editionen neuen Auftrieb gegeben, ohne beider Verhältnis konzeptionell zu klären.

Gegenstand der Altertumswissenschaft sind verschiedene Schriftsysteme, deren Bestandteile von rein bildhaften Schriftzeichen (z. B. Ägyptische Hieroglyphen) über abstraktere Zeichen, die heutigen Buchstaben ähnlich sind (z. B. phönizisches Alphabet), bis hin zu Buchstaben des griechischen und lateinischen Alphabets reichen. Für paläographische Fragen sind genaue Beschreibungen der Zeichen unerlässlich; in der Epigraphik spielt z. B. neben dem Material des Schriftträgers auch das gegenseitige Durchdringen von Bild und Text eine Rolle. Das Zeigen im Faksimile ist ein fester Bestandteil der Auseinandersetzung mit dem Text, da alle Phänomene der Schrift, des Schreibens und der Beschriftung allein in der Codierung nur schwer abgebildet werden könnten.

In der Musiknotation haben die geschriebenen Zeichen zwar überwiegend eine klar umrissene, d. h. codierbare Bedeutung. Allerdings finden sich vor allem in Skizzen und Entwürfen oft fragmentarische Zeichen, bei denen jede Zuweisung eines Namens bereits mehr in ein solches Zeichen hineininterpretiert, als dieses in einem solchen Stadium auszusagen vermag [1] . Eine mögliche Antwort auf die daraus resultierenden Probleme ist die bewusste Trennung von deutender Codierung und (vermeintlich) objektivem Faksimile, die dafür engmaschig aufeinander bezogen (d. h. verknüpft) werden.

In der Kunst, insbesondere der Malerei, gibt es keine direkte Entsprechung zum Einzelzeichen, und somit keine Transkription im eigentlichen Sinn. Der Fokus liegt hier entsprechend ganz auf dem Faksimile, welches direkt gezeigt wird. Allerdings gibt es häufig wiederkehrende Motive und Topoi, die markiert, verknüpft und kommentiert werden können. Damit einher geht eine andere Perspektive: Der Inhalt wird nicht transkribierend codiert, sondern durch Annotationen erschlossen.

In allen vorgestellten Wissenschaften ist ein Spektrum verschiedener Kombinationsmöglichkeiten von Faksimilie und Codierung zu beobachten. Diese Herangehensweisen sind dabei aber nicht rein fachspezifisch, sondern illustrieren fächerübergreifend eine Bandbreite unterschiedlicher Auffassungen zur Abgrenzung von Befund und Deutung. Im Rahmen der Panel Session soll zunächst durch Kurzvorträge in die Thematik und die Situation der Einzeldisziplinen eingeführt werden. Nach dieser ersten halben Stunde werden die Referenten zunächst für eine weitere halbe Stunde auf dem Podium die verschiedenen Ansätze erörtern, um Unterschiede und Gemeinsamkeiten herauszuarbeiten. Schließlich soll das Plenum in diese Diskussion einbezogen werden, um möglichst weitere Aspekte ggf. auch aus anderen Fächern beizutragen.

Vernetzung ist wichtig, Vernetzung ist gut - Aber wie vernetzt man richtig?

Pfeil, Patrick
ppfeil@uni-leipzig.de

Universität Leipzig, Deutschland

Aehnlich, Barbara

barbara.aehnlich@uni-jena.de
Friedrich--Schiller-Universität Jena

Das Panel hat das Ziel, verschiedene Formen der Vernetzung innerhalb der Digital Humanities zu diskutieren. Dabei soll besprochen werden, welche Netzwerke sinnvoll sind, welche Ziele Vernetzungen haben können und welcher Nutzen für die Entwicklung der Digital Humanities erzielt werden soll.

Zurzeit sind die Digital Humanities im deutschsprachigen Raum auf verschiedensten Ebenen vernetzt und es gibt unterschiedlichste Formen der Kommunikation der beteiligten Wissenschaftler_innen und Einrichtungen. Die folgende Auflistung stellt keine Gewichtung der einzelnen Ebenen dar – vielmehr sind alle Formen für sich von enormer Bedeutung für die Profilierung und Weiterentwicklung der Digital Humanities an Hochschulen und anderen Forschungseinrichtungen; von besonderem Wert ist der Austausch zwischen den Strukturen.

An erster Stelle zu nennen sind der Dachverband für die Digital Humanities im deutschsprachigen Raum (DHd) sowie die großen Infrastrukturdienstleister wie CLARIN-D oder DARIAH die anstreben, eine breite Kommunikationsbasis für alle Formen der Digital Humanities national sowie im internationalen Wettbewerb anzubieten. Dabei sollen möglichst alle Geisteswissenschaften sowie die Informationswissenschaften erreicht werden, um die verschiedenen Probleme und Fragestellungen in einem offenen und weit gefächerten Rahmen zu diskutieren und zu lösen. Darüber hinaus bieten diese Plattformen noch weitere Hilfestellungen: Möglichkeiten bei der Förderung von Projekten sowie Unterstützung bei Fragen zur Infrastruktur, zur nachhaltigen Datensicherung oder zu Lizenzierungsangelegenheiten. Auch das grundsätzliche wissenschaftliche Selbstverständnis der Digital Humanities wird diskutiert. Insbesondere in CLARIN-D haben die einschlägigen Fachwissenschaften eine eigene Plattform durch die Einrichtung von fachspezifischen Arbeitsgruppen erhalten, die ihnen Partizipation und Mitsprache an den überregionalen Entwicklungen sichert.

Zweitens existieren einzelne Verbände, die die Entwicklung der Digital Humanities in den jeweiligen Fachwissenschaften begleiten. Der Historikerverband hat einen Unterausschuss „Geschichte in der digitalen Welt“ und eine AG „Digitale Geschichtswissenschaft“ eingerichtet, die durch große Jahrestagungen hervortreten. Andere Fachverbände werden folgen. Hier findet man eine nationale und internationale Interessenvertretung der jeweiligen Fachwissenschaft in Verbindung mit den Digital Humanities als Schwerpunkt der Tätigkeiten. Daneben ist man bemüht, Projektvorhaben bzw. Projekte, die dem eigenen thematischen Rahmen entspringen, zu koordinieren und zu unterstützen. Als Beispiel für eine infrastrukturelle Gründung ist hier das Forschungszentrum Archäologie und Altertumswissenschaften IANUS zu nennen.

Drittens gibt es verschiedene Initiativen in einzelnen Bundesländern bzw. in länderübergreifenden Regionen, die ebenfalls eine Vernetzung der dort laufenden oder im Entstehen begriffenen Projekte und Initiativen zum Ziel haben. Als Beispiel kann hier der an der Universität Leipzig gegründete Lehrpraxis im Transfer-Facharbeitsarbeitskreis „ Digital Humanities in Sachsen “ gelten, der sich aber nicht auf Sachsen beschränkt, sondern auch eine Vernetzung mit den Hochschulen in Sachsen-Anhalt und Thüringen anstrebt, außerdem die Vernetzungsplattform Digital Humanities Forschungsverbund Niedersachsen (DHVF) und Digital Humanities Berlin . Auch in Thüringen befindet sich eine derartige Struktur im Aufbau. So gab es bereits ein erstes DH-Treffen in Erfurt, auf dem sich Wissenschaftler_innen verschiedenster Forschungseinrichtungen in Bezug auf die Zukunft der Digital Humanities und mögliche Kooperationen in Thüringen austauschten. In derartigen Projekten wird den regionalen Gegebenheiten, die sich unter anderem aus dem Föderalismus der deutschen Hochschullandschaft ergeben, in der eigenen Arbeit Rechnung getragen. Auf dieser Ebene erfolgen Vernetzungsgespräche und werden Initiativen erarbeitet, in deren Fokus die eigenen Voraussetzungen und Ziele stehen. Auch der nationale Wettbewerb im Vergleich mit anderen Bundesländern und Regionen spielt dabei eine Rolle. So werden zum Beispiel gemeinsame Antragsvorhaben initiiert, gemeinsame Projekte besprochen oder Fördervorschläge in die Politik getragen. Darüber hinaus werden aber ebenso die generellen Fragestellungen der Digital Humanities besprochen, die auch auf der Agenda der bundesweiten Initiativen zu finden sind.

Vernetzungsbewegungen an den einzelnen Hochschulen bzw. im Verbund mit den benachbarten Hochschulen bilden eine vierte Ebene. Hier stehen besonders die Interessen der einzelnen Institute und Einrichtungen im Verbund mit der Entwicklung der jeweilige(n) Hochschule(n) im Fokus der Arbeit. Dabei sollen die verschiedenen Projektideen, Projekte und Initiativen gebündelt, Absprachen getroffen und Synergieeffekte in den eigenen Bemühungen erreicht werden. Beispielgebend hierfür sei das DHnet Jena genannt, welches sich als Ansprechpartner für Fragen der Digital Humanities an der FSU Jena versteht und als interdisziplinäres Forschungsnetzwerk laufenden oder geplanten DH-Projekten ein Forum bietet. Ein Ziel dieser Vernetzung ist der Aufbau eines DH-Kompetenzzentrums, welches sich dem interdisziplinären und institutionenübergreifenden Methodentransfer verschreibt und zugleich eine Agenda liefert, um Fragen der technischen Ausrüstung und methodischen Umsetzung problemorientiert, nachhaltig und innovativ zu beantworten. Ähnlich stellen sich auch das Netzwerk für

Digitale Geisteswissenschaften an der Universität Erfurt und andere, hier nicht aufgeführte, aber ebenso wertvolle Initiativen dar.

Die verschiedenen Ebenen der Vernetzung innerhalb der Digital Humanities bieten damit zahlreiche Möglichkeiten der Kommunikation zwischen den Interessierten. Dies ist sicherlich gewinnbringend für alle Bemühungen um die Digital Humanities, zu hinterfragen ist der über den allgemeinen Austausch hinaus gehende Mehrwert. Im Panel soll daher diskutiert werden, wie die verschiedenen Vernetzungsebenen einzuschätzen sind, ob die dargestellte Struktur, wie sie zur Zeit existiert – also mit zahlreichen Überschneidungen der Vernetzungen, institutioneller, aber auch personeller Natur –, sinnvoll und gewinnbringend für alle Beteiligten ist, wie die Aufgaben der jeweiligen Strukturen angesehen werden und ob es Handlungsbedarf für Veränderungen im Bereich Vernetzung der Digital Humanities im deutschsprachigen Raum gibt.

Im Panel sollen Vertreter_innen aller vier aufgeführten Vernetzungsebenen in einem kurzen einleitenden Beitrag ihre jeweiligen Tätigkeitsfelder vorstellen. Dabei sollen die eigenen Aufgaben und der eigene Anspruch an die Vernetzung als Schwerpunkte der Kurzpräsentationen dargelegt werden. Dafür sind 40 Minuten vorgesehen. Damit stehen jedem/r Vortragenden etwa fünf Minuten Zeit für die Präsentation zur Verfügung.

Im Anschluss findet eine Diskussion unter den Vortragenden statt, für die 20 Minuten veranschlagt werden. Dabei soll auf die sechs folgenden Leitfragen eingegangen werden:

- Wie lassen sich Selbstverständnis, Ziele und Aufgaben der verschiedenen Vernetzungsebenen definieren?
- Wie können regionale und übergreifende Vernetzung verknüpft werden?
- Wo existieren Kompetenzprobleme, wie können diese vermieten werden und wie ist das Verhältnis der verschiedenen Vernetzungsebenen zueinander einzuschätzen?
- Wie ist das Verhältnis der Vernetzungsstrukturen zu den Geldgebern, wie BMBF, DFG oder Stiftungen, einzuschätzen? Was sollte diesbezüglich verändert werden?
- Wie ist die Vernetzungsstruktur im deutschsprachigen Raum im Vergleich zur internationalen Ebene zu bewerten und wie soll man sich in Zukunft in dieser Hinsicht aufstellen?
- Welche Rolle spielt der Verband DHd?

Zum Abschluss des Panels soll die Diskussion dieser Fragen ins Publikum getragen werden. Dabei ist auch möglich und gewünscht, dass die Zuhörer_innen weitere Themenfelder eröffnen. Für diese Diskussionsrunde sind 30 Minuten vorgesehen.

Die Panelleitung und Diskussionsmoderation übernimmt Dr. Andreas Christoph (Universität Jena – Ernst-Hackel-Haus). Als Diskutant_innen sind vorgesehen:

Vertreter_in DHd (seitens des Verbandes wird in Kürze eine Person benannt)

Prof. Dr. Heike Neuroth (Fachhochschule Potsdam – Bibliothekswissenschaften) (Bereitschaft zur Teilnahme per E-Mail zugesagt)

Prof. Dr. Cathleen Kantner (Universität Stuttgart – Sozialwissenschaften, Internationale Beziehungen und Europäische Integration; Stellvertretende Sprecherin der fachspezifischen Arbeitsgruppen bei CLARIN-D) (Bereitschaft zur Teilnahme per E-Mail zugesagt)

Dr. Leif Scheuermann (Universität Graz – Alte Geschichte und Altertumskunde; Universität Erfurt – Max-Weber-Kolleg, Interdisciplinary Center of E-Humanities in History and Social Sciences (ICE)) (Bereitschaft zur Teilnahme per E-Mail zugesagt)

Patrick Pfeil, M.A. (Universität Leipzig – Alte Geschichte; Koordinator des LiT-Facharbeitskreises „Digital Humanities in Sachsen“) (als Antragsteller zugesagt)

Dr. Barbara Aehnlich (Universität Jena – Geschichte der deutschen Sprache; Koordinatorin des DHnet Jena) (als Antragstellerin zugesagt)

Angesichts der wachsenden Bedeutung der Digital Humanities auf wissenschaftlicher, gesellschaftlicher und institutioneller Ebene entstanden in kürzester Zeit verschiedene Vernetzungsebenen, die laufenden oder geplanten Projekten ein Forum bieten, interdisziplinären Austausch und methodische Orientierung ermöglichen und damit als Ansprechpartner für entsprechende Fragestellungen dienen können. An vielen Stellen gibt es strukturelle und personelle Überschneidungen zwischen den Initiativen, die womöglich Vorteile mit sich bringen, aber auch das Risiko in sich bergen, sich „im Kreis zu drehen“ – fachspezifisch, methodisch, konzeptuell, institutionell und strukturell. Die Diskussion wird einen Einblick in diverse Vernetzungsprojekte bieten, den Mehrwert dieser Initiativen aber auch kritisch hinterfragen. Die große Frage hinter dem Panel wird sein: „Wie sinnvoll sind die Vernetzungen in ihren verschiedenen Ausprägungen in der praktischen Arbeit tatsächlich?“

Datenzentren für die nachhaltige Forschung in den Digital Humanities

Sahle, Patrick
sahle@uni-koeln.de
Data Center for the Humanities, Universität zu Köln

Trippel, Thorsten
thorsten.trippel@uni-tuebingen.de
CLARIN-D, Universität Tübingen

Neumann, Gerald
gneumann@bbaw.de
Berlin-Brandenburger Akademie der Wissenschaften

Engelhardt, Claudia
claudia.engelhardt@sub.uni-goettingen.de
Humanities Data Center, Staats- und Universitätsbibliothek Göttingen

Kurzawe, Daniel
kurzawe@sub.uni-goettingen.de
Humanities Data Center, Staats- und Universitätsbibliothek Göttingen

Schäfer, Felix
felix.schaefer@dainst.de
IANUS, Deutsches Archäologisches Institut

Wörner, Kai
kai.woerner@uni-hamburg.de
Geisteswissenschaftliche Infrastruktur für Nachhaltigkeit (gwin), Universität Hamburg

Durch die Digitalisierung der Forschung und die damit verbundenen gestiegenen Anforderungen an das Forschungsdatenmanagement ergibt sich ein zunehmender Bedarf an umfassender und forschungsorientierter Datenexpertise. Während die Geisteswissenschaften in diesem Prozess einen großen methodischen Fundus an Analysemethoden entwickeln, mangelt es noch an der Infrastruktur, den Institutionen und stabilen Praktiken, die notwendig sind, um Forschungsdaten für künftige Forschergenerationen zu bewahren und auf geeignete Weise zur Verfügung zu stellen. Datenzentren spielen im Idealfall nicht nur zum Zeitpunkt der Aufnahme abgeschlossener Projekte, also bei der Übernahme, Archivierung und ggf. Einbettung in eigene Repositorien eine wichtige Rolle. Vielmehr werden die Weichen für die optimale Archivfähigkeit und Nachnutzbarkeit von Forschungsdaten bereits vorher gestellt und erfordern ggf. eine Beratung von Anfang an. Auf der anderen Seite ist die Kuratierung von Daten und Anwendungen kein einmaliger Akt, sondern eine andauernde Aufgabe. Es stellt sich deshalb die Frage, welche Funktionen geisteswissenschaftliche Datenzentren und Forschungsinfrastrukturen für Forschende in den Digital Humanities insgesamt haben können und sollen. Wenn ein Mehrwert für die Forschung durch die dauerhafte Anschlussfähigkeit von Daten und Anwendungen an den ganzen Forschungsprozess entstehen soll, dann erfordert dies anscheinend einen ganzen Strauß an Diensten und Angeboten, die sich von der koordinierten Unterstützung durch Fachberatung, der Vermittlung von Hardware, der Beantragung von Mitteln, dem Betrieb von Repositorien bis zur anhaltenden Pflege von Daten und Anwendungen erstrecken.

Dieses Panel bringt Vertreter von geisteswissenschaftlichen Datenzentren und den zentralen DH-Infrastrukturprojekten zusammen. Gemeinsam wird diskutiert, wie die Forschung in ihrer Arbeit und in der Erzeugung und Nachnutzung von Daten unterstützt werden kann, wie erste Erfahrungen mit der Kuratierung von Daten aussehen und wie zielgerichtete zusätzliche Angebote gestaltet werden können.

Insbesondere soll dabei aufgezeigt werden, was heute bereits an Angeboten verfügbar ist und auf welche Weise diese in der Forschung genutzt werden. Für Forschende in den Digital Humanities bietet das Panel die Möglichkeit, über weiterhin bestehende Hürden und Herausforderungen zu diskutieren und so den Dialog zwischen der Forschung und den Infrastrukturen voranzutreiben. Die Themen reichen von konkreten Beispielen aus der Forschungspraxis bis zu Modellen und momentanen Fragestellungen zur institutionellen Organisation geisteswissenschaftlicher Forschungsdatenzentren.

Folgende Themenschwerpunkte und Fragestellungen bieten eine Orientierung zu möglichen Inhalten:

- Ausgangspunkte: Datengetriebene Forschung in den Geisteswissenschaften - was wird beforscht, welche Methoden werden angewendet, welche Daten entstehen und wie wird mit diesen Daten umgegangen? Welche Anforderungen ergeben sich daraus an die Langzeitarchivierung, Bereitstellung und Nachnutzung der Daten? Wie können die geisteswissenschaftliche Forschung und die langfristige Sicherung und Bereitstellung ihrer Ergebnisse durch die Datenzentren unterstützt werden? Welche Lücken in den momentanen Angeboten zeigen sich dabei? Konkrete Projekte, Beispiele oder Methoden können demonstriert und diskutiert werden.
- Zugänglichkeit: Wie lassen sich die bestehenden Angebote nutzen und wer ist für “meine Forschung” zuständig? Mit wem lassen sich Projekte planen, wer hostet mein Projekt und wer übernimmt die Daten und Anwendungen nach der Projektlaufzeit? Fragen dieser Art können direkt zwischen den Forschenden und Zentren diskutiert werden.
- Zielgruppen: An wen richten sich welche Angebote? Nicht ein Schuh passt allen. Wie finden Forscher zu den passenden Angeboten? Wie können Angebote für ein breites Anforderungsspektrum mit vielen heterogenen Forschungsfragen und Forschungsmethoden gestaltet werden? Wie stellen die Forschungsdatenzentren sicher, sich nicht an der Zielgruppe vorbei zu entwickeln?

- Zuständigkeit: Die aktuelle Landschaft an Forschungsdateninfrastrukturen lässt sich anhand mehrerer Dimensionen kartieren: regionaler bzw. standortgebundener gegenüber übergreifendem Auftrag; Spezialisierungen auf Datentypen, Disziplinen, Methoden; Art, Schichtung und Umfang der Dienstleistung: Beratung - umfassende Begleitung - technische Dienstleistungen, Übernahme von Daten, Anwendungskonservierung. Wo sind bestehende Initiativen zuzuordnen und wo ergeben sich Lücken? In welchem Verhältnis stehen die lokalen Datenzentren zu den übergreifenden Infrastrukturprojekten CLARIN-D und DARIAH-DE? Wie lassen sich die Aktivitäten der verschiedenen geisteswissenschaftlichen Datenzentren sinnvoll koordinieren und aufeinander abstimmen; welche Modelle der Kooperation sind denkbar?
- Beratung von Anfang an: Wie können z. B. fachspezifische Datenzentren IT-Empfehlungen zum nachhaltigen Umgang mit digitalen Daten in bestimmten Bereichen geben, um auf die Problematik der langfristigen Nachnutzbarkeit von digitalen Forschungsdaten aufmerksam zu machen und um Forschende mit praktischen Informationen zu Fragen des Forschungsdatenmanagements zu unterstützen?
- Kuratierung komplexer Forschungsdaten: Wie können Daten und Anwendungen aus beendeten Projekten übernommen und in übergreifende Strukturen der Langfristverfügbarkeit eingebracht werden? Hier liegen bei einigen Datenzentren inzwischen erste praktische Erfahrungen vor, über die zu diskutieren ist.
- Vertrauenswürdige Repositorien: Wie lassen sich projektübergreifende und stabile “generische” Lösungen für die Vorhaltung von Forschungsdaten aufbauen? Welchen Abdeckungsgrad können Repositorien hinsichtlich der Vielfalt der Forschungsprojekte erreichen? Wieso sind Repositorien nicht in allen Fällen die Lösung und was kann in solchen Fällen getan werden?
- Angepasste Dienste für die Fachgemeinschaft: Wie können Werkzeuge zur Kuratierung, Nutzung und Bearbeitung von Datenzentren oder größeren Forschungsinfrastrukturen entwickelt und dauerhaft vorgehalten werden, um die Reproduzierbarkeit von Forschungsergebnissen und eine kontinuierliche Nutzung und Weiterentwicklung zu gewährleisten?
- Verteilte Datenhaltung und virtuelle Kollektionen: Wie lässt sich eine verteilte Datenhaltung und übergreifende Nutzung in den DH organisieren? Wie kann die Zusammenarbeit von Datenzentren technisch durch Schnittstellen umgesetzt werden, die z. B. gemeinsame Datenmodelle, Formate und PID-Systeme verwenden? In welchem Maße erfordern technisch übergreifende Lösungen auch eine institutionelle Föderierung?
- Institutionelle Stabilität: Können nur Datenzentren als dauerhafte Einrichtungen die nachhaltige Bereitstellung von Forschungsdaten und Anwendungen gewährleisten oder gibt es andere Optionen?
- Fachbereichsspezifische Kompetenzen: Wie können an Datenzentren und Forschungsinfrastrukturen dauerhafte Kompetenzen und Strukturen zur Beratung von Wissenschaftlerinnen und Wissenschaftlern etabliert werden?

Fachwissenschaftliche Nutzungsszenarien der CLARIN-D Infrastruktur

Wiedemann, Gregor
gregor.wiedemann@uni-leipzig.de
Universität Leipzig

Gloning, Thomas
thomas.gloning@germanistik.uni-giessen.de
Universität Gießen

Blätte, Andreas
andreas.blaette@uni-due.de
Universität Duisburg/Essen

Keller, Maret
keller@gei.de
Georg-Eckert-Institut Braunschweig

Haaf, Susanne
haaf@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Würzner, Kay-Michel
wuerzner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

CLARIN-D: Eine Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften

Auffinden, Auswerten, Aufbewahren: Die Forschung mit digitalen Sprach- und Textdaten stellt die Sozial- und Geisteswissenschaften vor neue, fächerübergreifende Herausforderungen:

- Rohdaten und Ergebnisse sollen für Nachvollziehbarkeit und weitere Analysen zentral und einfach auffindbar sein,
- für die Auswertung können computergestützte Werkzeuge vielfältig eingesetzt werden – möglichst entlang methodischer Standards, und
- Forschungsergebnisse müssen nachvollziehbar und langfristig verfügbar gemacht werden.

Die Forschungsinfrastruktur CLARIN-D (Common Language Resources and Technology Infrastructure in Deutschland) unterstützt die Geistes- und Sozialwissenschaften dabei, ihre digitalen Ressourcen in nachhaltiger, offener und interoperabler Weise zur Verfügung zu stellen (www.clarin-d.de). Über neun CLARIN-D-Zentren mit unterschiedlichen Arbeitsschwerpunkten und einem breiten Angebot an Webservices stehen der deutschen Forschungslandschaft umfangreiche Angebote zur Forschung mit Sprachdaten zur Verfügung. Insofern bei den Angeboten die Perspektive von nicht computerlinguistisch vorgebildeten Fachwissenschaftler_innen als Zielgruppe im Vordergrund steht, sind die CLARIN-D Angebote für die Digital Humanities von besonderem Interesse, zielen sie doch darauf ab, die Arbeit mit digitalen Sprachdaten durch Vereinheitlichung von Standards und Bereitstellung von Werkzeugen und Webservices zu erleichtern.

Beim Auffinden von Ressourcen geht es darum, Zugang zu Daten zu erhalten, die der wissenschaftlichen Gemeinschaft zur Verfügung gestellt wurden. Diese Daten werden zitiert und können so gefunden und in anderen Forschungskontexten und zur Reproduktion von Ergebnissen verwendet werden.

Werkzeuge zur Analyse von Forschungsdaten sind über das Web zugänglich und können dadurch ohne zusätzlichen Aufwand verwendet werden. CLARIN-D macht Werkzeuge für die Geistes- und Sozialwissenschaften verfügbar, so dass unterschiedliche Werkzeuge für neue Forschungsfragestellungen zusammen verwendet werden können.

Daten, die in Forschungsprojekten entstehen oder die anderen Forschern zur Verfügung gestellt werden, können mit Hilfe der CLARIN-D-Zentren langfristig aufbewahrt und somit archiviert werden. Sie erhalten dabei eine eindeutige Referenz und können ähnlich wie Bücher und Artikel zitiert werden. Außerdem werden dadurch Anforderungen von Förderungsorganisationen zur Vorhaltung von Daten und Ergebnissen über Projektlaufzeiten hinaus gewährleistet.

Mit diesen Services ist CLARIN-D für eine Vielzahl von Fachdisziplinen, welche sich im Bereich der Digital Humanities bewegen, von großen Interesse. Um den Anforderungen und Bedürfnissen dieser Fachcommunities gerecht zu werden, haben sich (potenzielle) Nutzer_innen der Infrastruktur innerhalb von CLARIN-D in sogenannten Fach-Arbeitsgruppen (F-AGs) organisiert. In zehn F-AGs, welche von Germanistik und anderen Philologien über diverse Teilbereiche der Linguistik bis hin zu Sozial- und Geschichtswissenschaften reichen, sind mittlerweile ca. 200 Wissenschaftler_innen organisiert, welche sich über Möglichkeiten, Bedarfe und methodische Standards bei der Arbeit mit digitalen Sprachdaten austauschen. Zudem werden im Rahmen von sogenannten "Kurationsprojekten" wichtige fachwissenschaftliche Ressourcen aufbereitet und für die Forschungscommunity zugänglich gemacht.

Im geplanten Panel sollen exemplarisch zwei Use Cases dargestellt und dokumentiert werden (Abschnitt 2). Ein weiterer Beitrag befasst sich mit den Prinzipien und Möglichkeiten der Dokumentation von fachwissenschaftlichen Nutzungsszenarien von Clarin-D-Ressourcen (Abschnitt 3).

Fachwissenschaftliche Use Cases der Infrastruktur

Das Panel "Fachwissenschaftliche Nutzungsszenarien der CLARIN-D Infrastruktur" stellt zwei Use Cases aus unterschiedlichen fachwissenschaftlichen Perspektiven vor, bei denen Services bezüglich der drei zentralen Leistungen Auffinden, Auswerten und Aufbewahren zur Anwendung kommen. Über den Erkenntnisgewinn der Darstellung der einzelnen Use Cases hinaus wird so im Rahmen des Panels der Nutzen der Infrastruktur für unterschiedliche Disziplinen insgesamt sichtbar gemacht. Dazu stellen die Einzelvorträge die Generalisierbarkeit ihrer Ansätze an zentralen Punkten heraus um zu verdeutlichen, wie andere Forschungsfragen mit ähnlichen Methoden und Werkzeugen bearbeitet werden können. Zudem wird im Rahmen der Diskussionen zu den einzelnen Vorträgen vor allem auf Erfahrungen und Generalisierbarkeit der Ansätze fokussiert, um so den projekt- und fächerübergreifenden Mehrwert der Nutzung von Komponenten der Infrastruktur aufzuzeigen und die Anwendung für eigene Projekte zu ermutigen. Vorgestellt werden drei Nutzungsszenarien aus dem Bereich der Politikwissenschaften, der Neueren Geschichte und der Germanistik, wobei neben den Projekten vor allem die Reflexion des Einflusses der CLARIN-D Infrastruktur auf Forschungsmöglichkeiten und Dokumentation von Forschungsabläufen im Vordergrund steht.

Beitrag 1: Aufbereitung und Analyse von Parlamentsprotokollen als öffentliche Sprachressource der Demokratie

Im deutschsprachigen Raum besteht ein Mangel an frei verfügbaren, annotierten Korpora politischer Texte.

Dadurch wird der Einstieg in die DH-Forschung massiv gehemmt. Eine mögliche Lösung bieten Plenarprotokolle als öffentliches Archiv des politischen Zeitgeschehens einer Demokratie, welche im Rahmen der CLARIN-D Infrastruktur aufbereitet und verfügbar gemacht werden. Plenarprotokolle des Bundestags, der Landtage oder auch des Europaparlaments dokumentieren über große Zeiträume das gesamte Spektrum politischer Aktivität. Dies ist zugleich, ohne eine thematische Klassifikation von Debatten, ein Nachteil: Plenarprotokolle decken, wenn nicht Subkorpora nach inhaltlichen Kriterien gebildet werden können, das politische Geschehen für viele Auswertungszwecke zu undifferenziert ab. Für eine Vielzahl sozialwissenschaftlicher Fragestellungen ist es relevant, dass themen- bzw. politikfeldspezifische Subkorpora gebildet werden können. Dafür ist eine Annotation von Debatten und deren Klassifikation erforderlich. Eine entsprechende Aufwertung der Ressource "Plenarprotokollkorpus" wird im Kurationsprojekt der F-AG 8 vorgenommen.

Im Rahmen des Panel-Beitrags wird zuerst der Workflow dargestellt, wie auf Basis der bei Parlamenten verfügbaren Plenarprotokolle (im txt- und pdf-Format) XML-Dokumente aufbereitet werden, die TEI-Standards entsprechen. Anschließend zeigen wir, wie mit CLARIN-Tools Annotationen für die Korpusaufbereitung vorgenommen werden. Redebeiträge sollen auf Basis der manuellen Annotation automatisch in bestimmte Themenkategorien klassifiziert werden. Im Beitrag wird auch dargestellt, wie die Qualitätssicherung der (halb-)automatischen Aufbereitung durch Intercoderreliabilitäten und Qualitätskriterien maschineller Sprachverarbeitung von Seiten der klassischen Politikwissenschaft als auch von Seiten der Informatik realisiert werden kann. Versionierung der Daten und Issue Tracking werden dabei gleichermaßen realisiert. Im Ergebnis steht der sozialwissenschaftlichen Community (und selbstverständlich auch anderen Wissenschaftler_innen) ein nach Themengebieten differenziert auswertbares Korpus zur Verfügung.

Beitrag 2: CLARIN-kompatible Aufwertung OCR-erfasster Texte aus der Lehrbuchsammlung des GEIs und deren Integration in die CLARIN-D-Infrastruktur: Ein Fazit aus fachwissenschaftlicher Sicht

Welchen Aufwand bringt eine Integration historischer Quellen in die CLARIN-D Infrastruktur mit sich? Welche Mehrwerte ergeben sich daraus für die historische Forschung? Diese Fragen sollte das Projekt „ Quellen des Neuen: Realkundliches und naturwissenschaftliches Wissen für Dilettanten und Experten zwischen Aufklärung und Moderne" der 2014 gegründeten CLARIN-D-Facharbeitsgruppe „Neuere Geschichte" klären. Zu diesem Zweck sollten im Projekt Korpora aus verschiedenen Projektkontexten über die CLARIN-Infrastruktur miteinander verknüpft und interoperabel gemacht werden. Im Blick waren dabei (1) das digitale Schulbuchkorpus des GEI (Georg-Eckert-Institut Braunschweig; GEI), (2) das Korpus des Deutschen Textarchivs (Berlin-Brandenburgischen Akademie der Wissenschaften; BBAW) sowie (3) die gedruckten Publikationen des Universitätsgelehrten Johann Friedrich Blumenbach (1752–1840) (Akademie der Wissenschaften Mainz). Durch Verknüpfung dieser Korpora sollten Untersuchungen zum Zusammenhang und Verhältnis von schulischer Lehre mit der Wissensproduktion und -vermittlung im universitären Umfeld ermöglicht werden. Die CLARIN-Infrastruktur ermöglicht zudem den Rückgriff auf gegebene Tools für die vergleichende Analyse der gegebenen Korpora.

Die Herausforderung des Projekts bestand darin, die in Qualität und Form heterogenen Daten der verschiedenen Korpora CLARIN-konform aufzubereiten und miteinander interoperabel zu machen. Von mehreren möglichen Integrationsszenarien wurde in Zusammenarbeit mit dem CLARIN-D-Servicezentrum an der BBAW eine sehr genaue und vergleichsweise aufwändige Methode realisiert, mit dem Ziel, adäquate Forschungsdaten für die Wissensgeschichte und andere historisch arbeitende Disziplinen zu erhalten. Unter Nachnutzung bzw. Anpassung bestehender Workflows der BBAW wurden exemplarisch Schulbücher verschiedener Fächer und Zeiträume dem Auszeichnungslevel der DTA-Korpora und der bereits integrierten „Blumenbach-Online"-Ressourcen angepasst. In verschiedenen 'Stadien' der Textaufbereitung wurde zudem die Anwendbarkeit von CLARIN-D-Tools auf Teilmengen der verfügbaren Daten getestet.

Der Panel-Beitrag erläutert den Workflow zur CLARIN-konformen Aufbereitung der GEI-digital-Ressourcen und führt den Mehrwert dieser Arbeiten für die Forschung beispielhaft vor.

Beitrag 3: Möglichkeiten und Prinzipien der Dokumentation von fachlichen Nutzungsszenarien von Clarin-D-Ressourcen

Für die fachliche Nutzung der Ressourcen (Daten, Werkzeuge), die in einer Infrastruktur angeboten werden, ist die Frage von zentraler Bedeutung, wie man typische wissenschaftliche Anwendungsszenarien, bei denen Daten und Werkzeuge für eine spezifische fachliche Aufgabenstellung genutzt werden, so darstellen kann, dass die Dokumentationen für unterschiedliche Zielgruppen hilfreich und im besten Fall auch stimulierend sind.

Dabei sind einerseits unterschiedliche Nutzertypen (z. B. Doktorand_innen, Studierende, erfahrene und weniger erfahrene Forscher_innen), andererseits die ganz unterschiedlichen fachlichen Fragestellungen in verschiedenen Disziplinen (z. B. Germanistik) und Unterdisziplinen (z. B. Syntax, Wortschatzforschung) zu berücksichtigen. Gegenstand des Panel-Beitrags sind zunächst drei Typen der Dokumentation, die jeweils von einer fachlichen Fragestellung ausgeht, sodann die Ermittlung von Daten und Werkzeugen umfasst, die zielorientierte Anwendung von Daten und Werkzeugen beschreibt und mit dem Hinweis auf ein fachliches Resultat endet, das auf die ursprüngliche fachliche Fragestellung rückzubeziehen ist. Die drei Typen sind:

- gedruckte und bebilderte Anleitungen;
- Screencasts, bei denen eine fachliche Anwendung digitaler Ressourcen von einer Stimme aus dem Off kommentiert wird;
- Experten-Interviews, die vor Ort oder als virtuelles Video-Meeting aufgezeichnet werden und dann als Filme / Podcasts dauerhaft zugänglich sind. Hier vertritt ein/e Interviewer/in die Nutzerinteressen, eine Expertin oder ein Experte stellt die Ressourcen und ihre Anwendung dar.

Der Beitrag wird zunächst die Prototypen vorstellen und dann allgemeine Überlegungen zu Zielen, Verfahrensweisen, Prinzipien (z. B. Usability, Zielgruppenorientierung), Darstellungsformen und zum Repertoire von Darstellungsmitteln anschließen.

Fächerübergreifende Perspektiven

Wir versprechen uns durch die multiperspektivische Reflexion der hier vorgestellten Nutzungsszenarien nicht nur eine Verbesserung des Verständnisses für den Nutzen einer Forschungsinfrastruktur, sondern hegen insbesondere die Hoffnung, einen wichtigen Beitrag zur Methodendiskussion in den Digital Humanities zu liefern. In der gemeinsamen Nutzung und Weiterentwicklung von Technologien zum Auffinden, Auswerten und Aufbewahren digitaler Sprachdaten liegt der Schlüssel zur Etablierung von Standards, zum Herabsetzen von Zugangsschwellen und damit letztlich zur Ermöglichung eines breiten Austausches in diesem noch vergleichsweise jungen Forschungsfeld.

Ablauf

Das Panel wird moderiert von Gregor Wiedemann, Koordinator der Fach-AGs im CLARIN-D Projekt. Der inhaltliche Ablauf wird wie folgt gestaltet:

- Einführung zur CLARIN-D Infrastruktur : Gregor Wiedemann (Universität Leipzig, Koordinator für die CLARIN-D Fach-AGs), Vortragszeit: 5 Minuten
- Beitrag 1, Vortragender: Prof. Dr. Andreas Blätte (Universität Duisburg/Essen, Lehrstuhl für Public Policy und Landeskunde), Vortragszeit: 10 Minuten
- Beitrag 2, Vortragende: Dr. Maret Keller (Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung), Susanne Haaf (Berlin-Brandenburgische Akademie der Wissenschaften), Kay-Michel Würzner (Berlin-Brandenburgische Akademie der Wissenschaften), Vortragszeit: 10 Minuten
- Beitrag 3, Vortragender: Prof. Dr. Thomas Gloning (Universität Gießen, Professur für Germanistische Sprachwissenschaft), Vortragszeit: 10 Minuten
- Panel-Diskussion zu fächerübergreifenden Problemen bei der Arbeit mit digitalen Sprachdaten und Lösungsansätzen im Rahmen der CLARIN-D Infrastruktur, Diskussionszeit: 25 Minuten
- Publikumsdiskussion Möglichkeiten, Perspektiven aber auch (Einstiegs-)Hürden bei der Anwendung virtueller Forschungsinfrastrukturen. Neben Verständnisfragen sollen vor allem Bedarfe seitens der (potenziellen) Anwender_innen und Möglichkeiten zur Generalisierung der präsentierte Forschungsabläufe diskutiert werden. Diskussionszeit: 30 Minuten

Sektionen

Argumentanalyse in digitalen Textkorpora

Butt, Miriam
miriam.butt@uni-konstanz.de
Universität Konstanz, Deutschland

Heyer, Gerhard
heyerasv@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

Holzinger, Katharina
katharina.holzinger@uni-konstanz.de
Universität Konstanz, Deutschland

Kantner, Cathleen
cathleen.kantner@sowi.uni-stuttgart.de
Universität Stuttgard, Deutschland

Keim, Daniel A.
daniel.keim@uni-konstanz.de
Universität Konstanz, Deutschland

Kuhn, Jonas
jonas.kuhn@ims.uni-stuttgart.de
Universität Stuttgard, Deutschland

Schaal, Gary
gschaal@hsu-hh.de
Helmut-Schmidt-Universität, Universität der Bundeswehr, Hamburg

Blessing, André
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgard, Deutschland

Dumm, Sebastian
sebastian.dumm@hsu-hh.de
Helmut-Schmidt-Universität, Universität der Bundeswehr, Hamburg

El-Assady, Mennatallah
mennatallah.el-assady@uni-konstanz.de
Universität Konstanz, Deutschland

Gold, Valentin
valentin.gold@uni-konstanz.de
Universität Konstanz, Deutschland

Hautli-Janisz, Annette
annette.hautli@uni-konstanz.de
Universität Konstanz, Deutschland

Lemke, Matthias
lemkem@hsu-hh.de
Helmut-Schmidt-Universität, Universität der Bundeswehr, Hamburg

Müller, Maike
maike.mueller@uni-konstanz.de
Universität Konstantz, Deutschland

Niekler, Andreas
aniekler@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

Overbeck, Maximilian
maximilian.overbeck@sowi.uni-stuttgart.de
Universität Stuttgard, Deutschland

Wiedemann, Gregor
gregor.wiedemann@uni-leipzig.de
Universität Leipzig, Deutschland

Zusammenfassung der Sektion

Valentin Gold, Annette Hautli-Janisz, Andreas Niekler, Maximilian Overbeck und Gregor Wiedemann

Die Extrahierung und Annotation von Argumentationsstrukturen hat im Bereich der automatischen Diskursanalyse in den letzten Jahren an Bedeutung gewonnen, sei es in juristischen Dokumenten (Mochales / Moens 2011; Bach et al. 2013), wissenschaftlichen Texten (Kirschner et al. 2015), Zeitungsartikeln (Feng / Hirst 2011) oder Online-Diskussionen (Bex et al. 2013, 2014; Oraby et al. 2015). Vor diesem Hintergrund haben sich in den vergangenen Jahren die drei interdisziplinären Projekte *e-Identity*, *ePol* und *VisArgue* im Rahmen der eHumanities-Förderlinie des BMBF mit der semi-automatischen Identifikation und Analyse von Argumenten auseinandergesetzt.

Die Herausforderung, die allen Projekten gemein ist, ist die, dass die jeweilige Fragestellung über den eigentlichen Prozess der Argumentationsanalyse hinausgeht: Im Falle von *VisArgue* soll die Deliberativität des Diskurses approximiert werden, bei *ePol* geht um Ökonomisierungstechniken neoliberalen Sprechens, Begründens und Argumentierens in der politischen Öffentlichkeit und bei *e-Identity* um die Mobilisierung unterschiedlicher kollektiver Identitäten in politischen Debatten zu bewaffneten Konflikten und humanitären militärischen Interventionen. Daher sind diese Projekte

beispielhaft für die Anforderung der eHumanities: Trotz des gemeinsamen Zieles der Argumentationsextraktion wird der Begriff des Arguments und dessen Rolle in den einzelnen Projekten konzeptionell sehr verschieden gefasst und muss daher im Hinblick auf die jeweilige inhaltliche Fragestellung und die zu untersuchende Datenbasis unterschiedlich operationalisiert werden.

Den Kern im Projekt *VisArgue* bildet die Extrahierung von kausalen und adversativen Argumentstrukturen (Bögel et al. 2014), um Instanzen von Begründungen, Schlussfolgerungen und Gegenargumenten im Diskurs herausfinden zu können. Dies geschieht mithilfe eines linguistisch motivierten, regelbasierten Systems, das explizite Diskurskonnektoren automatisch disambiguiert und die Teile des Arguments im Diskurs verlässlich annotiert. Diese Annotationen dienen als Basis für die Visualisierung von deliberativen Mustern über den Diskurs hinweg und die damit einhergehende Interpretation desselben. Im Gegensatz zur regelbasierten Extrahierung werden im Projekt *ePol* maschinelle Lernverfahren angewandt, die jene Abschnitte in Zeitungstexten für eine inhaltsanalytische Auswertung identifizieren, die sprachliche Muster ökonomisierter Begründungen für Politik enthalten. Allerdings finden sich in Zeitungstexten nur sehr wenige explizite Argumentstrukturen, die einer formalen Anforderung expliziter Formulierung von beispielsweise Prämisse, Schlussregel und Schlussfolgerung genügen. Muster der hier eher implizit enthaltenen Begründungsstrukturen können anhand einer Menge von annotierten Beispielargumenten gelernt und zur Identifikation ähnlicher Textabschnitte angewendet werden, ohne dass eine bestimmte Form der Argumente explizit vorgegeben wird. Im *e-Identity* Projekt wurden die Potentiale für computer- und korpuslinguistische Methoden erschlossen, die eine interaktive und flexible Tiefenanalyse der Mobilisierung unterschiedlicher kollektiver Identitäten in öffentlichen Debatten über Krieg und Frieden ermöglichen. Maschinelle Lernverfahren kamen dabei sowohl bei der inhaltlichen Bereinigung der mehrsprachigen Textkorpora sowie bei der halb-automatischen Identifikation der unterschiedlichsten kollektiven Identitäten zum Einsatz.

In dieser Sektion wird daher der Frage nachgegangen, wie unterschiedliche theoretische und methodische Ansätze für die (semi-)automatische Identifikation und Analyse von Argumenten eingesetzt werden. In den Vorträgen werden die heterogenen Ansätze vor dem Hintergrund der jeweiligen Fragestellungen und daraus resultierender Anforderungen einzelnen eHumanities-Projekte im Detail vorgestellt. Dabei liegt der Fokus der Vorträge auf den Anwendungen, Ergebnissen und auf Perspektiven für die Evaluation. Insbesondere der Gütekontrolle räumen die Vorträge mehr Raum ein, um die Leistungsfähigkeit unterschiedlicher Ansätze und die Auswirkung auf Ergebnisse transparent darzustellen. Als prototypische Anwendungen von Argumentanalysen in den Humanities zeigen die Vorträge methodische Perspektiven und Ideen für Verwendungsmöglichkeiten jenseits der vorgestellten Projekte.

Vortrag 1: Deliberation in politischen Verhandlungen: Eine linguistisch-motivierte visuelle Analyse

Mennatallah El-Assady, Valentin Gold, Annette Hautli-Janisz, Maike Müller, Miriam Butt, Katharina Holzinger, Daniel A. Keim

Einleitung

Das *VisArgue* Projekt hat zum Ziel, automatisch zu erfassen, ob Verhandlungsteilnehmer deliberativ agieren, d. h. ob sie ihre Positionen u. a. respektvoll und rational begründen und sich schlussendlich dem besten Argument fügen. Die Datenbasis sind dabei transkribierte reale Verhandlungen, wie zum Beispiel die Schlichtungsgespräche zu Stuttgart 21. Zusätzlich zur Erfassung von Argumentationsmustern spielen bei der Argumentanalyse auch noch andere Faktoren eine Rolle, insbesondere die Beziehung des Sprechers zum Gesagten, die Beziehungen der Sprecher untereinander und die Struktur der Diskussion insgesamt. Mithilfe eines innovativen Visualisierungssystems werden diese vielschichtigen Muster aufgearbeitet, damit die einzelnen Faktoren von Argumentation, aber auch die Beziehungen der einzelnen Faktoren untereinander, interpretierbar gemacht werden können.

In diesem Beitrag wird am Beispiel der automatischen Erfassung von Argumentationsmustern aufgezeigt, wie das Projekt mit den generellen Herausforderungen der eHumanities umgeht: Das Konzept der Deliberation ist (computer)linguistisch gesehen eher abstrakt und bedarf einer konkreten Operationalisierung, damit das Konzept in den Daten fassbar gemacht wird. Die Zusammenhänge zwischen den verschiedenen Faktoren, die den Diskurs bestimmen, werden dann mithilfe eines Visualisierungssystems interpretierbar gemacht.

Im Folgenden werden die verschiedenen Dimensionen von Deliberation vorgestellt, gefolgt von einer Beschreibung der automatischen Argumentationsextraktion und der Annotation anderer deliberationsrelevanter Merkmale. Abschließend wird anhand eines konkreten Beispiels gezeigt, wie das Visualisierungssystem die Interpretation von Argumentationsmustern im Diskurs erlaubt.

Die Operationalisierung des Konzeptes der Deliberation

Das Konzept der Deliberation wird, wie in der folgenden Abbildung gezeigt, operationalisiert durch vier Dimensionen, die für die automatische Extraktion deliberativer Muster im Text relevant sind: Teilnahme (Participation), Atmosphäre und Respekt (Atmosphere & Respect), Argumentation und Rechtfertigung (Argumentation & Justification) und Entgegenkommen (Accomodation) (Gold / Holzinger 2015). In der Dimension 'Argumentation & Justification' werden unter anderem kausale Argumentationsketten annotiert, die darauf hindeuten, dass die Teilnehmer im Prozess der Entscheidungsfindung sind und Argumente austauschen ('Reason-giving'). In der Subdimension 'Information Certainty' wird auf der Basis von Ausdrücken epistemischer Modalität wie 'mit Sicherheit', 'wahrscheinlich' etc. annotiert, wie sicher sich die Sprecher des Gesagten sind. In der Dimension 'Accomodation' werden solche Einheiten im Diskurs annotiert, die entweder auf eine Einigung in der Verhandlung abzielen oder eine Uneinigkeit bekräftigen. Informationen, ob Sprecherbeiträge emotional oder sachlich sind, ob Sprecher andere Redner unterbrechen oder ob sie sich höflich verhalten, werden in der Dimension 'Atmosphere & Respect' gebündelt.

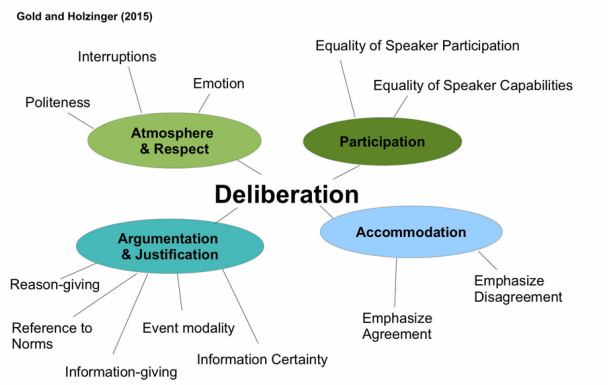


Abb. 1: Dimensionen der Deliberation

Auf Basis dieser Konkretisierung des Begriffs der Deliberation wird im Folgenden anhand der Dimensionen 'Argumentation & Justification' und 'Accomodation' gezeigt, wie die verschiedenen Ebenen innerhalb des Diskurses konkret annotiert werden. Zusammengenommen dienen diese Annotationen als Basis für die Visualisierung, um die Muster im Diskurs im Sinne der Deliberation interpretieren zu können.

Argumenterfassung

Als Datenbasis dienen transkribierte Verhandlungen, die entweder in projektinternen Verhandlungssimulationen gewonnen wurden oder von realen politischen Verhandlungen stammen, wie z. B. dem Schlichtungsverfahren von Stuttgart 21. Diese Daten werden in ein XML Schema übertragen, auf dessen Basis der Diskurs annotiert wird. Dazu werden die Äußerungen der Teilnehmer in Sätze aufgeteilt, die wiederum in kleinere Einheiten, sogenannte "elementary discourse units (EDUs)" eingeteilt werden, unter der Annahme, dass jede dieser Diskurseinheiten ein Event darstellt (Polanyi et al. 2004).

Ein Modul in der Annotation ist die Extrahierung von kausalen Argumentstrukturen (Bögel et al. 2014), was mithilfe eines linguistisch motivierten, regelbasierten Systems geschieht, das explizite Diskurskonnektoren automatisch disambiguiert und die einzelnen Teile eines Arguments im Diskurs verlässlich annotiert. Kausale Diskurskonnektoren wie 'weil', 'da' und 'denn' etc. leiten die Begründung einer Schlussfolgerung ein und geben so Hinweise auf argumentative Phasen in der Diskussion. Diese Informationen sind Teil der Ebene 'Reason-giving' in der Dimension 'Argumentation & Justification'. Im Gegensatz dazu stehen adversative Konnektoren wie 'aber', 'allerdings', 'jedoch' etc., die eine gegensätzliche Aussage zum Hauptsatz zum Ausdruck bringen und eine Ablehnung des Sprechers indizieren. Diese Äußerungen sind Teil der Subdimesion 'Emphasize Disagreement' in der Dimension 'Accomodation'.

Für die automatische Annotation derjenigen EDUs, die Teil der kausalen oder adversativen Einheiten bilden, werden den EDUs verschiedene Werte des XML Attributes 'discrel' zugeordnet, zum Beispiel discrel="reason" und discrel="conclusion" für kausale Argumentationsketten und discrel="opposition" für adversative Strukturen.

Zusätzlich zu der Information, dass die Teilnehmer Argumente austauschen oder sich zustimmend oder ablehnend in einer Diskussion verhalten, wird in der Unterdimension 'Information Certainty' in 'Argumentation & Justificiation' herausgearbeitet, wie sicher sich der Sprecher mit dem Inhalt seines Beitrages ist, d. h. welchen Kenntnisstand er vorgibt zu haben. Dies wird sichtbar durch Ausdrücke epistemischer Modalität, wie zum Beispiel 'wahrscheinlich', 'vielleicht' oder 'mit Sicherheit'. Um deren Bedeutung messbar zu machen, wird die Skala von Lassiter (2010), der die sogenannten "modes of knowing" von 0 (unmöglich – impossible) bis 1 (mit Sicherheit – certain) quantifiziert, herangezogen, und entsprechend annotiert: Der epistemische Ausdruck wird auf der Lexem-Ebene identifiziert und seine Bedeutung auf der Ebene der EDU mit dem XML-Attribut 'epistemic_value' versehen.

Ein weiterer Faktor, der für Deliberation relevant ist, ist die Haltung des Sprechers zum Gesagten. Dabei bleibt der Wahrheitsgehalt der Aussage unberührt, aber der Sprecher zeigt, wie er sich im Diskurs positioniert. Diese pragmatisch-relevante Ebene, die aus theoretisch-linguistischer Sicht schon vielseitig analysiert wurde, wird insbesondere von Partikeln wie 'ja', 'halt' und 'doch' ausgelöst (u. a. Kratzer 1999; Karagjosova 2004;

Zimmermann 2011) und ist linguistisch gesehen eine konventionelle Implikatur ('conventional implicature') (Potts 2012). Eine Herausforderung ist die Ambiguität der Partikel in der gesprochenen Sprache. Beispielsweise wird 'ja' häufig dazu verwendet, das gemeinsame Wissen der Diskussionteilnehmer zu betonen, auch verstanden als 'common ground' ("Sie wissen ja, dass …."). Allerdings kann 'ja' auch noch Zustimmung oder Ungeduld ("ja ja…") signalisieren, oder aber Hinhaltetechnik sein ("ja [Pause] ja"). Mithilfe eines regelbasierten Systems, das den Kontext vor und nach den Partikeln untersucht, werden die unterschiedlichen Bedeutungen herausgefiltert und als konventionelle Implikatur (CI) annotiert.

Diese Ebenen, die die klassische Argumentationsstruktur komplettieren, sind hochrelevant für die Analyse im Sinne der Deliberation: Neben der Frage, ob und wann argumentiert wird, ist auch noch relevant, WIE argumentiert wird: Argumentiert der Sprecher auf der Basis gemeinsamen Wissens (common ground), oder ist er sich seiner Schlussfolgerung sicher? Die Visualisierung muss daher die verschiedenen Bedeutungsebenen, die für die Herausarbeitung deliberativer Muster relevant sind, einzeln, aber auch im Zusammenspiel darstellen. Dazu wird im Folgenden das VisArgue Visualisierungssystem vorgestellt und gezeigt, wie Muster von Argumentationsstrukturen und Sprecherhaltung visuell über den Diskurs hinweg dargestellt werden können.

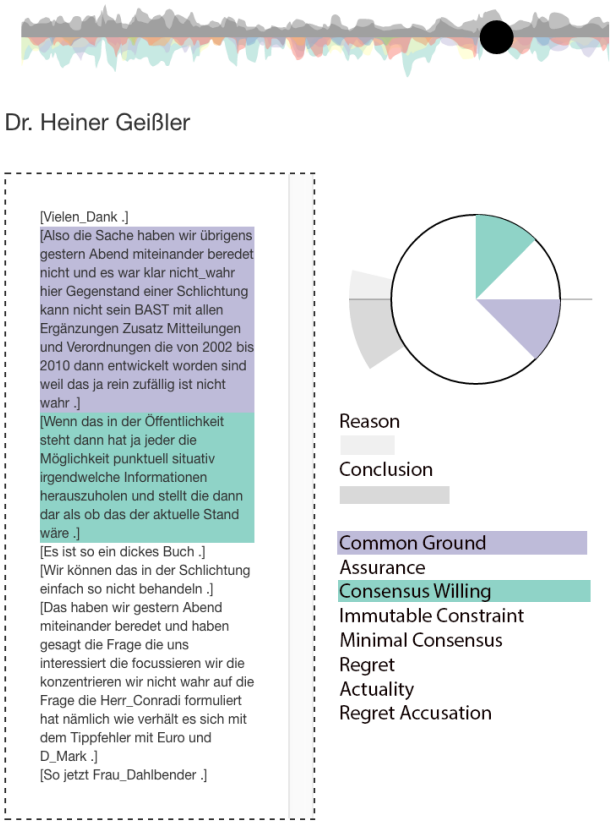


Abb. 2: Detailansicht Glyph

Visualisierung

Neben der Visualisierung von thematischen Blöcken in politischen Verhandlungen (Gold / Rohrdantz et al. 2015; Gold / El-Assady et al. 2015), ist ein Ziel der Visualisierung, Muster von Deliberation über den Diskurs hinweg, aber auch aggregiert für einzelne Sprecher so darzustellen, dass die zugrundeliegenden Daten, aber auch das große Ganze sichtbar wird. Eine Herausforderung ist hierbei die Mehrdimensionalität der Information, da zum einen die Ebene OB argumentiert wird, zum anderen aber auch die Information WIE argumentiert wird, visuell dargestellt werden soll. Dazu wird beispielsweise die Argumentationsdichte mit den Partikeln gemeinsam visualisiert: Jede Äußerung eines Sprechers wird als Glyph (Abbildung 2) dargestellt, wobei die Größe des Glyphen bestimmt wird durch die Länge der Aussage. Innerhalb des Glyphen sind die verschiedenen Werte der konventionellen Implikaturen abgetragen. Die zwei äußeren Ringe um den Glyphen zeigen Argumentationsmuster von 'reason' und 'conclusion' in einer Äußerung an; je größer die Teilringe, desto mehr EDUs sind Teil einer kausalen Argumentation. Die zugrundeliegende Äußerung kann mit einem Doppelklick auf den Glyph eingesehen werden (Abbildung 3).

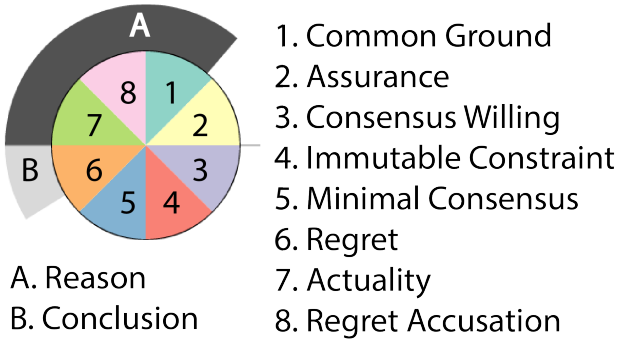


Abb. 3: Glyphdarstellung von Argumenten

Diese Glyphen werden für jede Äußerung über den Diskurs hinweg erstellt, wie in Abbildung 4 gezeigt. Durch die Interaktivität lässt sich ein Überblick über die Äußerungen der einzelnen Sprecher herstellen, und dabei Rückschlüsse ziehen, welche Rolle ein Sprecher in der Verhandlung eingenommen hat. Abbildung 4 zeigt einen Verhandlungstag der Schlichtungen zu Stuttgart 21: In der obersten Zeile findet sich der Mediator der Schlichtung, Dr. Heiner Geißler, wieder, dessen Beiträge von einem hohen Maß an 'consensus willing' und 'common ground' geprägt sind, wobei relativ wenig Argumente angeführt werden. Im Gegensatz dazu zeigt einer der Befürworter des Projektes, Dr. Volker Kefer, ein anderes Muster auf, nämlich einen deutlich höheren Anteil an argumentativen Redebeiträgen, die geprägt sind von Zusagen ('assurance')

und unabänderlichen Vorgaben ('immutable constraint'). Der hohe Grad an kausaler Argumentation findet sich auch bei einem Gegner des Projektes, Boris Palmer, der sich in seinen Beiträgen mehrheitlich auf den 'common ground', d. h. das gemeinsame Wissen der Verhandlungsteilnehmer, beruft.

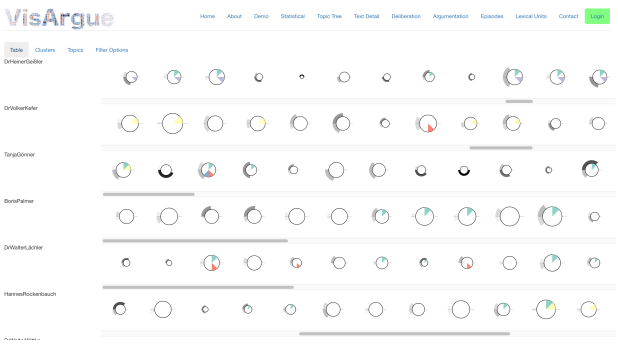


Abb. 4: Glyphen pro Sprecher

Im Sinne der Deliberation sind diese Muster relevant, weil sie zeigen, dass Verhandlungsteilnehmer verschieden argumentieren und sich damit unterschiedlich in der Verhandlung positionieren. Diese Muster tragen wesentlich dazu bei, den Verlauf und den Ausgang der Verhandlung zu erklären und Segmente von intensiven deliberativen Debatten zu identifizieren.

Zukünftige Arbeiten werden sich insbesondere mit dem Thema befassen, wie die weiteren Dimensionen der Deliberation in die Glyphenstruktur eingearbeitet werden können und inwiefern die linguistische Analyse weitere Anhaltspunkte von Argumentation und ihre Ausprägung aus dem Text extrahieren kann.

Zusammenfassung

Das *VisArgue*-Projekt zeigt am Beispiel der Argumentationserfassung, wie ein Ziel der Digital Humanities erreicht werden kann, nämlich der interdisziplinäre Austausch von Konzepten und Methoden: Durch die Kooperation von Politikwissenschaft, Linguistik und Informatik werden regelbasierte Analyse und visuelle Darstellung kombiniert und dadurch eine valide Basis für die Interpretation von Deliberation in politischen Verhandlungen möglich.

Vortrag 2: (Semi)-automatische Klassifikation für die Analyse neo-liberaler Begründungen und Argumentationen in großen Nachrichtenkorpora

Sebastian Dumm, Matthias Lemke, Andreas Niekler, Gregor Wiedemann, Gerhard Heyer, Gary S. Schaal,

Für die Analyse großer Mengen qualitativer Textdaten stehen den Sozialwissenschaften unterschiedliche konventionelle und innovative Methoden der Inhalts- und Diskursanalyse zur Verfügung. Die klassische sozialwissenschaftliche Inhaltsanalyse kann methodisch mit Verfahren des überwachten maschinellen Lernens verbunden werden (Scharkow 2012). Zur effizienten Generierung von Trainingsbeispielen kann eine solche (semi-)automatische Textklassifikation zu einem Active Learning Prozess erweitert werden (Dumm / Niekler 2015). Dabei werden schrittweise vom Computer vorgeschlagene Textbeispiele als Kandidaten für eine inhaltsanalytische Kategorie manuell evaluiert, und der Klassifikationsprozess mit den manuell bewerteten Beispielen erneut ausgeführt. Auf diese Weise können schnell mehrere hundert repräsentative Beispiele für eine Kategorie in großen Textkollektionen identifiziert werden. Ein solches Untersuchungsdesign ist im Rahmen des Projekts „ePol - Postdemokratie und Neoliberalismus“ methodologisch entworfen und technisch umgesetzt worden (Wiedemann et al. 2013). Der Vortrag beschreibt Ergebnisse und Lessons Learned aus diesem Projekt.

Das Projekt *ePol* greift die politiktheoretische Diskussion um die Erscheinungsformen gegenwärtiger westlicher Demokratien auf, welchen mit dem Konzeptbegriff Postdemokratie unter anderem eine Ökonomisierung des Poltischen unterstellt wird. [1] Die Ökonomisierung in den Begründungen politischer Entscheidungen untersuchen wir anhand von Sprachgebrauchsmustern in der politischen Öffentlichkeit, speziell in einem Korpus aus 3,5 Millionen Artikeln deutscher Tages- und Wochenzeitungen im Zeitraum von 1949 bis 2011. Unter neoliberalem Plausibilisieren verstehen wir dabei „Ökonomisierungstechniken“, die Argumente, Behauptungen und Metaphern zur Legitimation von politischem Output einsetzen und somit zum öffentlichen Sprachspiel der Politik gerechnet werden können. Den Gebrauch solcher qualitativer Begründungsmuster quantitativ im Zeitverlauf zu verfolgen und dessen Zu- oder Abnahme in Bezug auf bestimmte Randbedingungen zu testen (z. B. Zeitung oder Politikfeld) ist Ziel des Projekts. Dazu wurde ein modulares Forschungsdesign in drei Schritten umgesetzt:

- Selektion relevanter Artikel aus dem Korpus von 3,5 Millionen Artikeln, welche eine hohe dichte an neoliberalen Begründungsmustern erwarten lassen,
- Manuelle Annotation von Textstellen, welche neoliberale Begründungsmuster enthalten. Unterschieden werden zwei Kategorien von Ökonomisierungstechniken, die des Argumentierens und die des Behauptens.
- Automatische Klassifikation der beiden Kategorien auf dem Gesamtdatenbestand zur Identifikation

von Trends im Sprachgebrauch ökonomisierter Begründungen.

Selektion relevanter Artikel

In einem ersten Schritt wird eine Dokument-Retrieval-Strategie auf das gesamte Korpus angewendet, um Artikel mit (potenziell) möglichst hoher Dichte an neoliberalem Sprachgebrauch und Begründungsmustern zu identifizieren. Die Dokumente werden mit Hilfe eines einfachen Wörterbuches von 127 Argumentmarkern (Dumm / Lemke 2013) und eines kontextualisierten Wörterbuches (Wiedemann / Niekler 2014) nach Relevanz bewertet. Das kontextualisierte Wörterbuch enthält typischen Sprachgebrauch, der aus 36 in deutscher Sprache verfügbaren Schriften der Mitglieder des neoliberalen Think Tanks „Mont Pélerin Society“ extrahiert wurde. Dies umfasst eine Liste mit 500 Schlüsselbegriffen (z. B. Markt, Freiheit, Preis) sowie Statistiken über deren typische Kontexte (z. B. persönliche Freiheit, unternehmerische Freiheit). Die Berechnung eines Ähnlichkeitsmaßes des Sprachgebrauchs in diesem Vergleichskorpus mit den Artikeln aus unserem Zeitungskorpus hinsichtlich neoliberaler Sprachgebrauchsmuster und Argumentmarker führen zu einer sortierten Liste von Artikeln, welche als Ausgangspunkt für den Prozess der (semi-)automatischen Kodierung dient. Die 10.000 höchst bewerteten Dokumente werden für die Folgeschritte selektiert.

(Semi-)autiomatische Kodierung als Active Learning

Nachrichtenartikel enthalten für gewöhnlich nur wenige detaillierte und elaborierte argumentative Strukturen, welche den formalen Anforderungen einer vollständigen Argumentation folgen. Aus diesem Grund betrachten wir zwei Kategorien von Begründungsmustern: Argumente und Plausibilisierungen in neoliberalen Begründungszusammenhängen. Diese Kategorien werden in einem theoretisch begründeten Codebuch formal definiert. Im Gegensatz zu Argumenten, welche die Vollständigkeit von Argumentationsmustern durch Vorhandensein von Prämisse, Kausalmarker und Schlussfolgerung voraussetzen, sind Plausibilisierungen durch Behauptungen und idiomatische Referenzen auf vermeintlich akzeptiertes Wissen gekennzeichnet (z. B. „Tatsache ist ...“, „selbstverständlich“). Anschließend werden in den 100 relevantesten Artikeln aus Schritt 1 Textstellen annotiert, die den Codebuch-Definitionen entsprechen. Zur Überprüfung der Qualität der Codebuch-Definitionen und der Arbeit der Kodierer kann die Intercoder-Reliabilität bestimmt werden – ein Maß, welches die (zufallsbereinigte) Übereinstimmung zweier Kodierer auf demselben Text angibt. Insofern es sich bei den in unserem Projekt verwendeten Kategorien um zwei recht abstrakte Konzepte handelt, sind die Übereinstimmungsmaße eher am unteren Ende der akzeptablen Werte für eine verlässliche Kodierung angesiedelt. Im Gegensatz zu typischen Kodes wie Thema oder Affektposition wird hier die Schwierigkeit bei der Operationalisierung komplexer politiktheoretischer Konzepte deutlich. Insbesondere die Kategorie des Behauptens zeichnet sich durch eine große sprachliche Varianz aus, welche sowohl manuelle als auch automatische Kodiermethoden vor große Probleme stellt. Insofern es uns aber eher um die Bestimmung von Kategorieproportionen und Trends in sehr großen Datenmengen geht, als um die exakte Bestimmung von Einzelereignissen in den Daten, sind diese Ungenauigkeiten hinnehmbar. In diesem initialen Annotationsprozess wurden 218 Absätze mit Argumentationszusammenhang und 135 Absätze mit Plausibilisierungszusammenhang in den 100 relevantesten Artikeln annotiert.

Diese initiale Trainingsmenge muss für eine valide Trendbestimmung mit Hilfe automatischer Textklassifikation noch deutlich erweitert werden. Um effizient mehr gute, das heißt die Kategorien gut beschreibende, Textbeispiele zu finden, wird ein Active-Learning-Ansatz angewendet. Dazu wird ein maschineller Lernalgorithmus auf Basis der aktuell annotierten Textbeispiele trainiert und auf die noch nicht annotierten Dokumente aus den 10.000 zuvor selektierten, potenziell relevanten Dokumenten angewendet. Auf der technologischen Ebene nutzen wir eine Support Vector Machine (SVM) mit einem linearen Kernel. Wir extrahieren eine große Vielfalt von Texteigenschaften (Features) aus den Trainingsbeispielen, um den Klassifikationsprozess auch generisch für andere Probleme nutzen zu können. Die extrahierten Feature-Strukturen beinhalten Wort-N-Gramme, Part-of-Speech-N-Gramme und binäre Features über das Vorhandensein von Begriffen in unseren zwei initial erstellten Diktionären (neoliberaler Sprachgebrauch und Argumentmarker). Wir wenden eine Chi-Square Feature-Selektion an, um für die eigentliche Klassifikation nur Kategorie-relevante Features zu verwenden und übergeben die so vorverarbeitete Trainingsmenge an den Klassifikator. Der Klassifikator liefert eine Menge an Absätze aus den bislang ungesehenen Zeitungsartikeln zurück, welche eine hinreichende Ähnlichkeit in Bezug auf die Merkmalsstrukturen der bereits annotierten Artikel aufweisen. Die Kodierer sind nun gefragt, eine Auswahl dieser Textbeispiele manuell zu evaluieren und so der Trainingsmenge neue Positiv- bzw. Negativ-Beispiele für die zwei Kategorien hinzuzufügen. In je zehn Iterationen dieses Prozesses, bei denen jeweils 200 gefundene Textstellen evaluiert wurden, wurde die initiale Trainingsmenge um 515 Absätze mit Argumentationszusammenhang und 540 Absätze mit Plausibilisierung erweitert.

Evaluation und automatische Kodierung

Analog zu den Gütekriterien der Sozialforschung werden für Ansätze des Text Mining bzw. des maschinellen Lernens Methoden zur Qualitätssicherung eingesetzt. Die Güte einer Textklassifikation wird in der Regel mit Hilfe der k-fachen Kreuzvalidierung bewertet, für die k mal auf k-1 Teilen der Trainingsdaten ein Klassifikationsmodell trainiert und auf dem verbliebenen ein Teil der Trainingsdaten getestet wird (Dumm / Niekler 2015). Dazu werden Qualitätskennzahlen wie Precision, Recall und ihr gewichtetes Mittel, der F1-Wert, zur Beurteilung der Güte des Verfahrens berechnet. Diese Maße sind verwandt mit den Reliabilitätsmaßen aus den klassischen Methoden der Sozialwissenschaften wie beispielsweise Cohens Kappa. Idealerweise werden F1-Werte um 0,7 analog zu reliablen menschlichen Kodierern angestrebt. Für den oben beschriebenen Active-Learning Prozess lässt sich feststellen, dass die F1-Werte ausgehend von sehr niedrigen Werten um 0,25 im Zuge weiterer Iterationen zunächst schrittweise auf höhere Werte ansteigen, nach ca. sieben Iterationen jedoch kaum noch eine Verbesserung stattfindet. Die Sammlung von Trainingsbeispielen für die Kategorie kann in diesem Fall als weitgehend gesättigt betrachtet werden, insofern das Hinzufügen von neuen Beispielen die Performance nicht mehr allzustark verändert. Gleichzeitig sind nach ca. 7 bis 10 Iterationen genug Trainingsbeispiele vorhanden, um eine valide Klassifikation des Gesamtkorpus aller 3,5 Mio. Dokumente vorzunehmen.

Für die finalen Trainingsmengen werden die folgenden F1-Werte erreicht: $F1_{Argument} = 0{,}608$ und $F1_{Plausibilisierung} = 0{,}491$. Für eine individuelle Klassifikation, welche darauf bedacht ist möglichst genau Einzelereignisse in einer Datenmenge korrekt zu bestimmen, können diese Qualitätswerte nur bedingt zufrieden stellen. Unser Klassifikator liefert bei relativ hohem Recall auch viele Textstellen zurück, die bei manueller Evaluation nicht in die entsprechende Kategorie einsortiert werden können. Unser Analyseziel liegt jedoch, wie häufig in den Sozialwissenschaften, nicht in der Vorhersage von Einzelereignissen, sondern in der validen Bestimmung von Proportionen und Trends (Hopkins / King 2010). Für diesen Fall kann die Performance des Klassifikators als ausreichend betrachtet werden, da durch die systematische Überschätzung des wahren Anteils an Textbeispielen für eine Kategorie die Änderungen im Verhältnis der Kategorieproportionen zueinander an unterschiedlichen Zeitpunkten des diachronen Korpus nicht verfälscht werden. Auch wenn die Anteile insgesamt durch den Klassifikator als zu hoch eingeschätzt werden mögen, reflektieren die Messungen der Kategorieanteile im Korpus in unterschiedlichen Zeitabschnitten die Zu- bzw. Abnahme der Häufigkeit des Gebrauchs von neoliberalen Argumentations- bzw. Plausibilisierungsmustern korrekt. Im Gesamtkorpus des *ePol*-Projekts werden im Zuge der finalen Klassifikation 105.740 Argumentansätze und 753.653 Plausibilisierungsabsätze identifiziert.

Für die Bestimmung von Trends werden die Dokumente gezählt, in denen eine der beiden Kategorien vorkommt. Daraus lassen sich wiederum Dokumentfrequenzen für bestimmte Zeiträume aggregieren und mit dem Gesamtdatenbestand in diesen Zeiträumen normalisieren. Damit können Zeitverläufe der Kategorien sichtbar gemacht werden, die wiederum politikwissenschaftlich interpretiert werden können (siehe Abbildung 5).

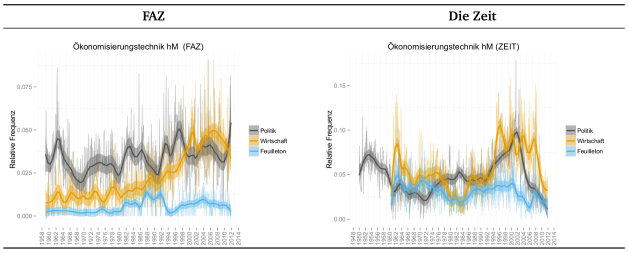


Abb. 5: Relative Frequenzen von Dokumenten in Die Zeit und FAZ welche neoliberale Argumentzusammenhänge enthalten, getrennt nach drei Zeitungsressorts (Politik, Wirtschaft, Kultur)

Verallgemeinerung der Ergebnisse

Der Ansatz des Active Learning im *ePol*-Projekt zur Messung abstrakter Kategorien, welche bislang lediglich qualitativ beschrieben worden sind, kann zu einem Ansatz von semi-automatischer Inhaltsanalyse verallgemeinert werden, bei dem die Schritte 1. Dokumentidentifikation, 2. manuelle Kodierung und 3. automatische Kodierung in der beschriebenen Weise miteinander kombiniert werden. Für die Auswertung sehr großer Datenmengen erlaubt der Ansatz nicht nur die Beobachtung von komplexen Kategorien im Zeitverlauf, sondern auch, in Erweiterung des *ePol*-Ansatzes mit einem größeren Kategorienschema, die Beobachtung des gemeinsamen Auftretens von Kategorien für inhaltliche Schlussfolgerungen auf sich gegenseitig bedingende Inhalte. Zusätzlich bietet die beliebige Facettierung der automatischen Analyse einer Vollerhebung Vorteile gegenüber manuellen Analysen, die auf vorab festgelegte Sampling-Strategien beschränkt sind.

Vortrag 3: Die Anwendung computer- und korpuslinguistischer Methoden für eine interaktive und flexible Tiefenanalyse der Mobilisierung kollektiver

Identitäten in öffentlichen Debatten über Krieg und Frieden – e-Identity

Cathleen Kantner, Jonas Kuhn, André Blessing und Maximilian Overbeck

Internationale Krisenereignisse wie Kriege und humanitäre militärische Interventionen lösen heftige öffentliche Kontroversen aus. Die Menschen machen sich Sorgen und fragen: Welche Effekte hat der Konflikt für unser eigenes Land, für Europa und für die Welt? Wer sind die Opfer, wer die Täter im Krisenland? Soll unser Land Truppen entsenden, verstärken oder ihren Einsatz zum wiederholten Male verlängern? Und falls ja, mit welchem Mandat sollen „unsere" Truppen agieren – verteilen sie Lebensmittel oder setzten sie Waffen ein? Wie gehen „wir" (in unserem Land, in Europa, im Westen, …) damit um, wenn Zivilisten oder „unsere" Soldaten dabei das Leben verlieren?

In öffentlichen Debatten zu kontroversen politischen Themen werden unterschiedlichste politische Positionen oftmals mit Rekurs auf das kollektive Selbstverständnis einer Wir-Gemeinschaft begründet. Die Mobilisierung unterschiedlichster kollektiver – europäischer, nationaler, religiöser usw. – Identitäten stellt somit eine zentrale Argumentationsfigur in der politischen Öffentlichkeit dar. Politische Sprecher begründen ihre Beteiligung an einem militärischen Einsatz oder ihre Enthaltung mit Rekurs auf das kollektive Selbstverständnis einer Wir-Gemeinschaft.

Ein Beispiel: In der europäischen, öffentlichen Debatte über die militärische Intervention in Libyen 2011 wurde auch über das kollektive Selbstverständnis der Europäer verhandelt. Die europäische Identität wurde teils als Problemlösungsgemeinschaft, teils aber auch als Gemeinschaft mit einem normativen Selbstverständnis diskutiert, die sich der Verteidigung der Menschenrechte verpflichtet habe.

Im *e-Identity* Projekt wurden die Potentiale für computer- und korpuslinguistische Methoden erschlossen, die eine interaktive und flexible Tiefenanalyse der Mobilisierung dieser unterschiedlichsten Formen kollektiver Identitäten in öffentlichen Debatten über Krieg und Frieden ermöglichen. [2] Zur methodischen Umsetzung der Forschungsfragen und Überprüfung der Hypothesen untersuchten wir internationale Diskussionen über Kriege und humanitäre militärische Interventionen seit dem Ende des Kalten Krieges 1990. Dabei ging es uns vor allem darum, das komplexe Geflecht von Identitätsdiskursen in diesen Kontroversen genauer zu analysieren. Im Prozess der Anwendung und Analyse wurden zwei computer- und korpuslinguistische Tools entwickelt, der *Complex Concept Builder* und eine *Explorationswerkbank*.

Eine Explorationswerkbank zur Korpuserstellung, -erschließung und -bearbeitung wurde entwickelt, um Sozialwissenschaftlern auch über das Projektende hinaus als flexibles Bindeglied zu vorhandenen Infrastrukturen (z. B. CLARIN) zu dienen. Sie lässt sich unterschiedlichsten individuellen Forschungsfragen und Textmaterialien anpassen und bildet insbesondere auch die technische Basis für den Complex Concept Builder (Kliche et al. 2014; Mahlow et al. 2014). Im *e-Identity* Projekt wurde somit ein bereinigtes, mehrsprachiges Korpus von 460.917 [3]

Zeitungsartikeln aus sechs Ländern (Deutschland, Österreich, Frankreich, UK, Irland, USA) generiert, das den Zeitraum von Januar 1990 bis Dezember 2012 abdeckt.

Um in Korpora Textbelege zu finden, in denen Sprecher sich auf eine kollektive Identität beziehen, sind gängige stichwortbasierte Suchtechnologien nicht ausreichend, weil solch ein komplexer Begriff sehr unterschiedlich lexikalisiert und in seiner Interpretation in hohem Maße kontextuell bestimmt sein kann. Gesucht waren daher neue Methoden zur interaktiven inhaltlichen Korpuserschließung.

Um der Vielschichtigkeit der im Korpusmaterial zu untersuchenden Indikatoren ebenso Rechnung zu tragen wie dem erheblichen Korpusumfang und dem Nebeneinander von deutsch-, englisch und französischsprachigen Texten, wurde ein transparenter, vom jeweiligen Forschungsteam individuell nutzbarer *Complex Concept Builder* entwickelt, der sprachtechnologische Werkzeuge und Methoden anbietet, die in den Sozialwissenschaften bislang nur in Ausnahmefällen Anwendung fanden (Blessing et al. 2013). Maschinelle Lernverfahren kamen dabei sowohl bei der inhaltlichen Bereinigung der mehrsprachigen Textkorpora sowie bei der halbautomatischen Identifikation der verschiedenen Identitätstypen zum Einsatz. Komplexe fachwissenschaftliche Begriffe (wie der Identitätsbegriff inklusive der feinen Unterschiede und Nuancen zwischen verschiedenen kollektiven Identitäten) können innerhalb des *Complex Concept Builder* für die Anwendung an alltagssprachlichem Textmaterial operationalisiert werden.

Explorationswerkbank und Complex Concept Builder (CCB) werden im Verlauf dieses Jahres über einen CLARIN Server zugänglich gemacht. Beide Tools erlauben den Export ihrer aggregierten Ergebnisse (z. B. Artikelanzahl, Anzahl der identifizierten Textstellen) zur anschließenden statistischen Analyse. Die für die Fachwissenschaftler transparente Reflexion der Ergebnisse bleibt dabei weiterhin gewährleistet, indem beispielsweise ein Aufsplitten der quantitativen Analysen in die einzelnen qualitativen Analysen möglich ist.

Im Folgenden wird in Kürze angerissen, in welchen Bereichen computer- und korpuslinguistische Methoden sowie Ansätze des maschinellen Lernens Anwendung fanden, um die Analyse kollektiver Identitäten innerhalb der umfangreichen Zeitungstextkorpora durchzuführen. Die Verbindung quantitativer und qualitativer Analyseschritte ermöglichte es, eine komplexe sozialwissenschaftliche Fragestellung auf einer großen Textmenge zu untersuchen und

zugleich die Einhaltung der sozialwissenschaftlichen Forschungsstandards der Validität und Reliabilität zu gewährleisten. Im Rahmen unseres Vortrags auf der DHd-Jahrestagung 2016 sollen die folgenden Verfahren genauer präsentiert werden.

- *Inhaltliche Bereinigung der Zeitungstextkorpora von Sampling-Fehlern*: Der Complex-Concept-Builder (CCB) wurde entwickelt, um große mehrsprachige Textmassen nach sozialwissenschaftlich relevanten Aspekten „vorzusortieren“ und er erwies sich bereits bei der Samplebereinigung unter inhaltlichen Gesichtspunkten als äußerst produktiv (Blessing et al. 2015). Mithilfe einer Topic Modellierung wurde eine optimale Vorauswahl einer Trainingsmenge von Texten zur inhaltlichen Dokumentenbereinigung möglich. [4] Die manuelle Annotation erlaubte beispielsweise den sofortigen Ausschluss eines ‚Topics‘ wie Buchrezensionen, die für unsere politikwissenschaftliche Fragestellung nicht relevant sind. Andererseits konnte das schwierige 'Thema' Sport, das sowohl reine Sportberichterstattung mit militärischen Metaphern als auch politische Berichte über Militäreinsätze mit sportlichen Metaphern und Referenzen enthält, detailliert annotiert werden. Maschinelles Lernen setzten wir dann bei der Klassifikation der Dokumente des gesamten Korpus in gute versus off-topic Texte ein.

Die folgende Abbildung (Abb. 6) zeigt im oberen Teil eine herkömmliche Klassifizierung per Zufallsauswahl der manuell-annotierten Trainingsdaten. Unten ist unser Verfahren abgebildet: Topics helfen, die optimale Trainingsmenge zu bestimmen, wobei mindestens aus jedem Topic ein Dokument manuell annotiert wird und damit eine breite Abdeckung gewährt ist. Dadurch wird das Ergebnis des neuen Klassifikators besser (er findet nun z. B. auch Artikel zum Ruanda-Konflikt).

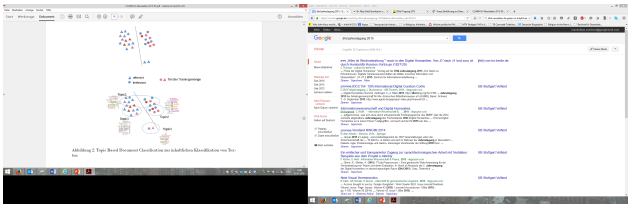


Abb. 6: Topic Based Document Classification zur inhaltlichen Bereinigung von Texten

Im Anschluss an die Bewertung einer bestimmten Anzahl an Zeitungsartikel wird über maschinelles Lernen die Bewertung auf die Gesamtmenge der Zeitungsartikel angewendet. Wir folgen der Idee von *Dualist*, einem interaktiven Klassifikationsmechanismus (Settles 2011; Settles / Zhu 2012). Die Architektur von Dualist basiert auf MALLET (McCallum 2002) und konnte leicht in unsere Architektur integriert werden. Die Zeitungsartikel, die durch den Computer automatisch aussortiert werden, können in weiteren iterativen Schritten erneut manuell bewertet werden, um den Klassifikator weiter zu optimieren. Eine weitere Abbildung (Abb. 7) zeigt den inhaltlichen Vorgang der Bereinigung im Complex Concept Builder. Die rot markierten Artikel wurden nach einer qualitativen Kodierung automatisch als Sampling Errors identifiziert, während die grün markierten Artikel automatisch dem Issue "Kriege und Humanitäre Militärische Interventionen" zugeordnet wurden.

Abb. 7: Halbautomatisierte inhaltliche Bereinigung von Sampling Fehlern im *Complex Concept Builder*

Im Fall des *e-Identity* Korpus blieben von insgesamt 766.452 Zeitungsartikeln, die ursprünglich Teil des unbereinigten Korpus waren, lediglich 460.917 Zeitungsartikel übrig (siehe Abbildung 8).

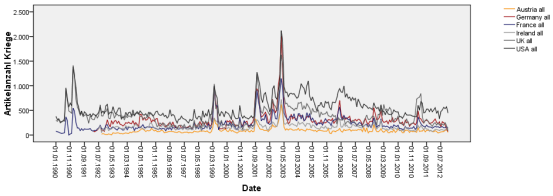


Abb. 8: Das bereinigte Issue-Cycle für das Thema Kriege und Humanitäre Militärische Interventionen (nach Monaten aggregiert, N=460.917)

Dieses Verfahren eignet sich darüber hinaus zum Aufspüren und Erstellen inhaltlicher Subkollektionen von Texten für spezifische Fragestellungen. Es bildet somit einen der methodischen Ausgangspunkte für das von Prof. Dr. Andreas Blätte an der Universität Duisburg-Essen geleitete Kurationsprojekt der FAG-8, in dem es um die thematische Strukturierung deutscher Parlamentsprotokolle geht.

- *Korpuslinguistische Verfahren für die semi-automatische Identifikation kollektiver Identitäten in den Zeitungstextkorpora:* Es wurden semantische Felder für die unterschiedlichen Identitätsebenen (z. B. nationale, europäische oder transatlantische Identitäten) mitsamt unterschiedlicher Begründungsfiguren (z. B. kulturelle Identität vs. Interessengeleitete Zweckgemeinschaft) generiert. Relevante Terme wurden über komplexe Diktionäre operationalisiert, die sowohl Lemmatisierungen als auch variable Äußerungen der interessierenden Terme innerhalb eines Satzes

berücksichtigen können. Die finalen Diktionäre wurden auf die mehrsprachigen Zeitungskorpora angewendet und in Form von Zeitreihen-Plots visualisiert und ausgewertet.

- *Manuelle Kodierung und die halbautomatische Identifikation von der Äußerung kollektiver Identitäten in bewaffneten Konflikten, unterstützt durch maschinelles Lernen*: Aus dem 460.917 Zeitungsartikel umfassenden Gesamtkorpus wurde ein Teilsample gezogen, das die wissenschaftlich üblichen Kriterien der Repräsentativität erfüllt. Auf diesem Teilsample wurde die manuelle Kodierung von insgesamt 5.000 Zeitungsartikeln durchgeführt. Die Unterstützung durch die *Complex Concept Builder*-Oberfläche ermöglichte die gleichzeitige und kontinuierliche Supervision und Datenauswertung der Kodierungen. Die manuell kodierten Textpassagen dienten im Anschluss als Datengrundlage für das Machine Learning Verfahren. Es wurde ein Klassifikator für die halbautomatische Identifikation der Äußerung kollektiver Identitäten trainiert und anschließend auf den Gesamtkorpus von 460.917 Zeitungsartikeln angewendet.

Zusammenfassung:

Die aus sozialwissenschaftlicher Perspektive interessanten und forschungsleitenden Konzepte sind nicht standardisierbar. Im *e-Identity* Projekt vertreten wir daher den Ansatz, dass computer- und korpuslinguistische Ansätze den Forscher dabei unterstützen sollten, ihre auf individuelle Fragestellungen gemünzten Korpora effizient zu managen und zu bereinigen. Sie sollten dem einzelnen Forscherteam Raum für seine eigene Operationalisierung lassen und dabei z. B. im Wechselspiel von manueller und automatischer Annotation in 'lernenden' Anwendungen die Vorteile beider Zugänge intelligent kombinieren. Dies schließt natürlich nicht aus, dass bewährte Operationalisierungen für die im Umfeld dieser komplexen fachlichen Konzepte ausgedrückten Sachverhalte, Bewertungen und Beziehungen usw. wie üblich analysiert werden können. Transparente und flexible CLARIN-Tools, die sich zu Workflows zusammenbinden lassen, die für eine spezifische fachwissenschaftliche Forschungsfrage sensibel bleiben, werden Sozialwissenschaftlern viele kreative Möglichkeiten bieten, interdisziplinären Austausch stimulieren und Spaß bei der Arbeit machen!

Notes

1. Ausführliche Informationen zum Projekthintergrund auf http://www.epol-projekt.de. Die hier vorgestellten Analysen wurden mit dem Leipzig Corpus Miner, einer webbasierten Analyseinfrastruktur, durchgeführt (Niekler et al. 2014; Wiedemann / Niekler 2015).
2. Für weitere Details zum e-Identity Projekt siehe http://www.uni-stuttgart.de/soz/ib/forschung/Forschungsprojekte/eIdentity.html . Für sozialwissenschaftliche Studien, in denen die Tools und Korpora des *e-Identity* Projekts bereits angewendet wurden, siehe Kantner 2015, Kantner et al. (erscheint in Kürze), Overbeck 2015 (im Druck).
3. Die unbereinigte Textmenge betrug 902.029 Zeitungsartikel. Der umfangreiche und innovative Bereinigungsprozess des Datensatzes von Dubletten und Samplingfehlern war ein zentraler Bestandteil des *e-Identity* Projekts.
4. Der *Complex Concept Builder* bietet ein Verfahren, um auf der Grundlage von insgesamt 50, 100 oder 200 automatisch erstellten Topics, die auf Grundlage der "Latent Dirichlet Allocation" (LDA) – Methode generiert werden (Blei et al. 2003; Niekler / Jähnichen 2012), inhaltliche Samplingfehler zu identifizieren. Die Visualisierung von Wortwolken einer automatischen Topicanalyse erleichtert die Identifikation von inhaltlichen Samplingfehlern.

Bibliographie

Bach, Ngo Xuan / Nguyen Le Minh / Tran Thi Oanh / Akira Shimazu (2013): "A Two-Phase Framework for Learning Logical Structures of Paragraphs in Legal Articles", in: *ACM Transactions on Asian Language Information Processing (TALIP)* 12, 1: Nr. 3.

Bex, Floris / Lawrence, John / Snaith, Mark / Reed, Chris (2013): "Implementing the Argument Web", in: *Communications of the ACM* 56, 10: 66–73.

Bex, Floris / Snaith, Mark / Lawrence, John / Reed, Chris (2014): "ArguBlogging: An Application for the Argument Web", in: *Journal of Web Semantics* 25: 9–15.

Blei, David M. / Ng, Andrew Y. / Jordan, Michael I. (2003): "Latent dirichlet allocation", in: *Journal of machine Learning research* 3: 993-1022.

Blessing, Andre / Sonntag, Jonathan / Kliche, Fritz / Heid, Ulrich / Kuhn, Jonas / Stede, Manfred (2013): "Towards a tool for interactive concept building for large scale analysis in the humanities", in: *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Sofia.

Blessing, Andre / Kliche, Fritz / Heid, Ulrich / Kantner, Cathleen / Kuhn, Jonas (2015): "Die Exploration großer Textsammlungen in den Sozialwissenschaften", in: *CLARIN Newsletter* 8: 17-20.

Bögel, Tina / Hautli-Janisz, Annette / Sulger, Sebastian / Butt, Miriam (2014): "Automatic Detection of Causal Relations in German Multilogs", in: *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)* 20–27.

Dumm, Sebastian / Lemke, Matthias (2013): "Argumentmarker. Definition, Generierung und Anwendung im Rahmen eines semi-automatischen Dokument-Retrieval-Verfahrens", in: *Schriftenreihe des Verbundprojekts „ePol – Postdemokratie und Neoliberalismus“*, Discussion-Paper 3 .

Dumm, Sebastian / Niekler, Andreas (2015): "Methoden, Qualitätssicherung und

Forschungs design. Diskurs- und Inhaltsanalyse zwischen Sozialwissenschaften und automatischer Sprachverarbeitung", in: Lemke, Matthias / Wiedemann, Gregor (eds.): *Text Mining in den Sozialwissenschaften*. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse. Wiesbaden: Springer VS 89-116.

Feng, Vanessa Wei / Hirst, Graeme (2011): "Classifying Arguments by Scheme", in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* 987–996.

Gold, Valentin / Holzinger, Katharina (2015): *An Automated Text-Analysis Approach to Measuring the Quality of Deliberative Communication*. Paper presented at the 73th Annual Meeting of the Midwest Political Science Association (MPSA), San Francisco.

Gold, Valentin / Rohrdantz, Christian / El-Assady, Mennatallah (2015): "Exploratory Text Analysis using Lexical Episode Plots", in: The Eurographics Association (ed.): *EuroVisShort2015* 85-89 http://dx.doi.org/10.2312/eurovisshort.20151130.

Gold, Valentin / El-Assady, Mennatallah / Bögel, Tina / Rohrdantz, Christian / Butt, Miriam / Holzinger, Katharina / Keim, Daniel (2015): "Visual Linguistic Analysis of Political Discussions: Measuring Deliberative Quality", in: *Digital Scholarship in the Humanities* http://dx.doi.org/10.1093/llc/fqv033 .

Hopkins, Daniel / King, Gary (2010): "A Method of Automated Nonparametric Content Analysis for Social Science", in: *American Journal of Political Science* 54, 229–247.

Kantner, Cathleen (2015): *War and Intervention in the Transnational Public Sphere: Problem-solving and European identity-formation*. London: Routledge.

Kantner, Cathleen / Overbeck, Maximilian / Sangar, Eric (erscheint in Kürze): "Die Analyse ‚weicher' Konzepte mit ‚harten' korpuslinguistischen Methoden: Multiple kollektive Identitäten", in: Behnke, Joachim / Blaette, Andreas / Schnapp, Kai-Uwe / Wagemann, Claudius (eds.): *Big Data: Große Möglichkeiten oder große Probleme?* Baden-Baden: Nomos Verlag.

Karagjosova, Elena (2004): *The Meaning and Function of German Modal Particles* (= Saarabrücken Dissertations in Computational Linguistics and Language Technology 18). Saarbrücken: Computational Linguistics & Phonetics, Universität des Saarlandes.

Kirschner, Christian / Eckle-Kohler, Judith / Gurevych, Iryna (2015): "Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications", in: *Proceedings of the 2nd Workshop on Argumentation Mining (ARG-MINING 2015)* 1-11.

Kliche, Fritz / Blessing, Andre / Sonntag, Jonathan / Heid, Ulrich (2014): *"The e-identity exploration workbench", in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik.

Kratzer, Angelika (1999): *Beyond "oops" and "ouch". How descriptive and expressive meaning interact*. Paper presented at the Cornell Conference on Theories of Context Dependency.

Lassiter, Daniel (2010): "Gradable epistemic modals, probability, and scale structure", in: *Proceedings of the 20th conference on Semantics and Linguistic Theory (SALT 20)* 197-215.

Mahlow, Cerstin / Eckart, Kerstin / Stegmann, Jens / Blessing, Andre / Thiele, Gregor / Gärtner, Markus / Kuhn, Jonas (2014): "Resources, Tools, and Applications at the CLARIN Center Stuttgart", in: *Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014)* 11-21.

McCallum, Andrew K. (2002): "MALLET: MAchine Learning for LanguagE Toolkit" http://mallet.cs.umass.edu/about.php .

Mochales Palau, Raquel / Moens, Marie-Francine(2011): "Argument Mining", in: *Artificial Intelligence and Law* 19, 1: 1-22.

Niekler, Andreas / Jähnichen, Patrick (2012): "Matching results of latent dirichlet allocation for text", in: *Proceedings of ICCM 2012, 11th International Conference on Cognitive Modeling* 317-322.

Niekler, Andreas / Wiedemann, Gregor / Heyer, Gerhard (2014): "Leipzig Corpus Miner - A Text Mining Infrastructure for Qualitative Data Analysis", in: *Proceedings of the Conference on Terminology and Knowledge Engineering 2014*, Berlin.

Oraby, Shereen / Reed, Lena / Compton, Ryan / Riloff, Ellen / Walker, Marilyn / Whittaker, Steve (2015): "And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue", in: *Proceedings of the 2nd Workshop on Argumentation Mining (ARG-MINING 2015*) 116-126.

Overbeck, Maximilian (im Druck): "Observers turning into Participants: Shifting perspectives on Religion and Armed Conflicts in Western News Coverage", in: *La revue Tocqueville* 36, 2.

Polanyi, Livia / Culy, Chris / van den Berg, Martin / Thione, Gian Lorenzo / Ahn, David (2004): "Sentential structure and discourse parsing", in: *Proceedings of the 2004 ACL Workshop on Discourse Annotation* 80–87.

Potts, Christopher(2012): "Conventional implicature and expressive content", in: Maienborn, Claudia / von Heusinger, Klaus / Portner, Paul (eds.): *Semantics* 3 (= Handbücher zur Sprach- und Kommuninikationswissenschaft 33, 3). Berlin: de Gruyter Mouton 2516–2536.

Scharkow, Michael (2012): *Automatische Inhaltsanalyse und maschinelles Lernen*. Berlin: epubli.

Settles, Burr (2011): "Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances", in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* 1467-1478.

Settles, Burr / Zhu, Xiaojin (2012): "Behavioral factors in interactive training of text classifiers", in: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 563-567.

Wiedemann, Gregor / Lemke, Matthias / Niekler, Andreas (2013): "Postdemokratie und Neoliberalismus. Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949-2011", in: *Zeitschrift für Politische Theorie* 4, 1: 99-115.

Wiedemann, Gregor / Niekler, Andreas (2014): "Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries", in: *Proceedings of the Conference on Terminology and Knowledge Engineering 2014*, Berlin.

Wiedemann, Gregor / Niekler, Andreas (2015): "Analyse qualitativer Daten mit dem 'Leipzig Corpus Miner'", in: Lemke, Matthias / Wiedemann, Gregor (eds.):*Text Mining in den Sozialwissenschaften*. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse. Wiesbaden: Springer VS 63-88.

Zimmermann, Malte (2011): "Discourse particles", in: von Heusinger, Klaus / Maienborn, Claudia / Portner, Paul (eds.): *Semantics 2*(= Handbücher zur Sprach- und Kommunikationswissenschaft 33, 2). Berlin: Mouton de Gruyter 2011-2038.

"Delta" in der stilometrischen Autorschaftsattribution

Evert, Stefan
stefan.evert@fau.de
Universität Erlangen-Nürnberg, Deutschland

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Dimpel, Friedrich Michael
friedrich.m.dimpel@fau.de
Universität Erlangen-Nürnberg, Deutschland

Schöch, Christof
christof.schoech@uni-wuerzburg.de
Universität Würzburg, Deutschland

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
Universität Würzburg, Deutschland

Vitt, Thorsten
thorsten.vitt@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reger, Isabella
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Büttner, Andreas
andreas.buettner@uni-wuerzburg.de
Universität Würzburg, Deutschland

Proisl, Thomas
thomas.proisl@fau.de
Universität Erlangen-Nürnberg, Deutschland

Die Sektion

Stilometrische Verfahren der Autorschaftsattribution haben eine lange Tradition in den digitalen Geisteswissenschaften: Mit der Analyse der *Federalist Papers* durch Mosteller und Wallace (1963) konnten schon Anfang der 1960er Jahre Erfolge verzeichnet werden. Überblicksbeiträge von Patrick Juola (2006) und Efstathios Stamatatos (2009) belegen die Vielfältigkeit der Bestrebungen, stilometrische Verfahren für die Autorschaftsattribution einzusetzen und weiterzuentwickeln.

Ein jüngerer Meilenstein der stilometrischen Autorschaftsattribution ist ohne Zweifel das von John Burrows (2002) vorgeschlagene “Delta”-Maß zur Bestimmung der stilistischen Ähnlichkeit zwischen Texten. Die beeindruckend gute Performance von Delta in verschiedenen Sprachen und Gattungen sollte allerdings nicht darüber hinwegtäuschen, dass die theoretischen Hintergründe weitgehend unverstanden geblieben sind (Argamon 2008). Anders ausgedrückt: Wir wissen, dass Delta funktioniert, aber nicht, warum es funktioniert. In diesem Kontext möchte die hier vorgeschlagene Sektion den aktuellen Stand der Forschung in der stilometrischen Autorschaftsattribution mit Delta vorstellen und neueste Entwicklungen anhand konkreter, eigener Untersuchungen demonstrieren. Jeder der drei Vorträge der Sektion leistet hierzu einen Beitrag:

- Der Beitrag von Stefan Evert, Thomas Proisl, Fotis Jannidis, Steffen Pielström, Isabella Reger, Christof Schöch und Thorsten Vitt “Burrows Delta verstehen” (vgl. 2.), gibt einen Überlick über den Forschungsstand rund um Delta und analysiert, warum die Veränderung von Delta durch Verwendung des Kosinus-Abstands zwischen den Vektoren (Smith / Aldridge 2012) eine so deutliche Verbesserung der Ergebnisse erbracht hat (Jannidis

et al. 2015). Am Beispiel einer Sammlung deutscher Romane aus dem 19. und 20. Jahrhundert zeigt der Beitrag, wie sich verschiedene Strategien der Normalisierung oder anderweitigen Behandlung des Merkmalsvektors (hier: Wortformen und ihre Häufigkeiten) auf die Attributionsqualität auswirken und inwiefern dies Einblick darin erlaubt, wie sich Information über Autorschaft im Merkmalsvektor manifestiert - was auch einen Aspekt der Leistungsfähigkeit des klassischen Delta erklärt.

- Der Vortrag von Friedrich Michael Dimpel, "Burrows Delta im Mittelalter: Wilde Graphien und metrische Analysedaten" (vgl. 3.), beleuchtet den Einsatz unterschiedlicher Merkmalstypen für die Ähnlichkeitsbestimmung von Texten mit Delta. Er zeigt am Beispiel einer Sammlung mittelhochdeutscher Texte, dass nicht nur die äußerst häufigen Funktionswörter, sondern auch metrische Eigenschaften für die Autorschaftsattribution eingesetzt werden können. Zugleich thematisiert er ein Problem, das immer dann auftritt, wenn Texte älterer Sprachstufen stilometrisch analysiert werden: das der nicht normierten, d. h. variablen Schreibweisen von Wörtern.
- Der Beitrag von Andreas Büttner und Thomas Proisl, "Stilometrie interdisziplinär: Merkmalsselektion zur Differenzierung zwischen Übersetzer- und Fachvokabular" (vgl. 4.), behandelt am Beispiel der Übersetzerattribution bei arabisch-lateinischen Übersetzungen philosophischer Texte die Manipulation des Merkmalsvektors nicht durch verschiedene Normalisierungsstrategien, sondern durch gezielte, selektive Merkmalseliminierung. Das Verfahren verbessert nicht nur die Attributionsqualität, sondern erlaubt auch die Isolierung des Autorsignals einerseits, des disziplinenbezogenen Signals andererseits und gibt einen Einblick darin, welche Einzelmerkmale für das Autorschaftssignal statistisch gesehen entscheidend sind.

Die drei Beiträge demonstrieren auf diese Weise verschiedene aktuelle Entwicklungen in der stilometrischen Autorschaftsattribution mit Delta und seinen Varianten. Sie zeigen, wie bei der Anwendung stilometrischer Distanzmaße auf ganz unterschiedliche Gegenstandsbereiche ähnliche methodische Fragen zu berücksichtigen sind. Und sie partizipieren direkt an aktuellsten, internationalen Entwicklungen bei der Verwendung von Distanzmaßen wie Delta für die stilometrische Autorschaftsattribution.

Burrows Delta verstehen

Überblick zum Forschungsstand

Burrows Delta ist einer der erfolgreichsten Algorithmen der Computational Stylistics (Burrows 2002). In einer ganzen Reihe von Studien wurde seine Brauchbarkeit nachgewiesen (z. B. Hoover 2004, Rybicki / Eder 2011). Im ersten Schritt bei der Berechnung von Delta werden in einer nach Häufigkeit sortierten Token-Dokument-Matrix alle Werte normalisiert, indem ihre relative Häufigkeit im Dokument berechnet wird, um Textlängenunterschiede auszugleichen. Im zweiten Schritt werden alle Werte durch eine z-Transformation standardisiert:

z i D = f i D - μ i σ i

wobei *f* i(*D*) die relative Häufigkeit des Wortes *i* in einem Dokument, μ i der Mittelwert über die relativen Häufigkeiten des Wortes *i* in allen Dokumenten ist und σ i die Standardabweichung. Durch diese Standardisierung tragen alle Worte in gleichem Maße zum Differenzprofil, das im dritten Schritt berechnet wird, bei. In einem dritten Schritt werden die Abstände aller Texte voneinander berechnet: Für jedes Wort wird die Differenz zwischen dem z-Score für das Wort in dem einen Text und dem anderen Text ermittelt. Die Absolutbeträge der Differenzen werden für alle ausgewählten Wörter aufaddiert:

Δ B = ∑ i = 1 m z i D 1 - z i (D 2)

m steht für die Anzahl der häufigsten Wörter (MFW - *most frequent words*), die für die Untersuchung herangezogen werden. Diese Summe ergibt den Abstand zwischen zwei Texten; je kleiner der Wert ist, desto ähnlicher – so die gängige Interpretation – sind sich die Texte stilistisch, und desto höher ist die Wahrscheinlichkeit, dass sie vom selben Autor verfasst wurden.

Trotz seiner Einfachheit und seiner praktischen Nützlichkeit mangelt es bislang allerdings an einer Erklärung für die Funktionsweise des Algorithmus. Argamon (2008) zeigt, dass der dritte Schritt in Burrows Delta sich als Berechnung des *Manhattan*-Abstands zwischen zwei Punkten in einem mehrdimensionalen Raum verstehen lässt, wobei in jeder Dimension die Häufigkeit eines bestimmten Wortes eingetragen ist. Er schlägt vor, stattdessen den Euklidischen Abstand, also die Länge der direkten Linie zwischen den Punkten, zu nehmen, weil dieser „possibly more natural" (Argamon 2008: 134) sei und zudem eine wahrscheinlichkeitstheoretische Interpretation der standardisierten *z*-Werte erlaubt. Bei einer empirischen Prüfung zeigte sich, dass keiner der Vorschläge eine Verbesserung bringt (Jannidis et al. 2015).

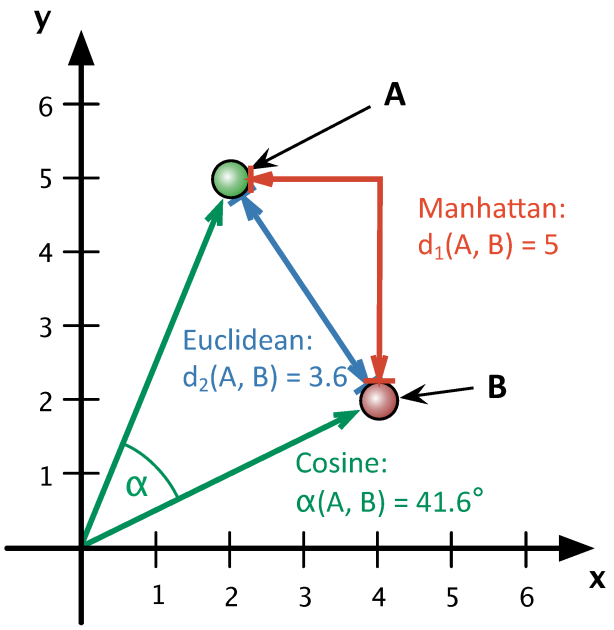


Abb. 1: Darstellung des Abstands zwischen zwei Texten, die nur aus zwei Worten bestehen. Burrows Delta verwendet die Manhattan-Distanz. Argamons Vorschlag, die Euklidische Distanz zu verwenden, sein *Quadratic-Delta*, brachte eine Verschlechterung der Clustering Ergebnisse, während der Vorschlag von Smith und Aldrige, den Cosinus-Abstand bzw. Winkel zwischen den Vektoren zu verwenden, eine deutliche Verbesserung erbrachte.

Smith und Aldrige (2011) schlagen vor, wie im Information Retrieval üblich (Baeza-Yates / Ribeiro-Neto 1999: 27), den Cosinus des Winkels zwischen den Dokumentenverktoren zu verwenden. Die Cosinus-Variante von Delta übertrifft Burrows Delta fast immer an Leistungsfähigkeit und weist, im Gegensatz zu den anderen Varianten, auch bei der Verwendung sehr vieler MFWs keine Verschlechterung auf (Jannidis et. al. 2015). Es stellt sich die Frage, warum Delta $_{Cos}$ besser ist als Delta $_{Bur}$ und ob auf diese Weise erklärt werden kann, warum Delta $_{Bur}$ so überraschend leistungsfähig ist.

Entscheidend für unsere weitere Analyse war die Erkenntnis, dass man die Verwendung des Cosinus-Abstands als eine Vektor-Normalisierung verstehen kann, da für die Berechnung des Winkels – anders als bei Manhattan- und Euklidischem Abstand – die Länge der Vektoren keine Rolle spielt (vgl. Abb. 1). Experimente haben gezeigt, dass eine explizite Vektor-Normalisierung auch die Ergebnisse der anderen Deltamaße erheblich verbessert und Leistungsunterschiede zwischen den Delta-Varianten weitgehend neutralisiert (Evert et al. 2015).

Daraus wurden zwei Hypothesen abgeleitet:

- (H1) Verantwortlich für die Leistungsunterschiede sind vor allem einzelne Extremwerte („Ausreißer“), d. h. besonders große (positive oder negative) *z*-Werte, die nicht für Autoren, sondern nur für einzelne Texte spezifisch sind. Da das Euklidische Abstandsmaß besonders stark von solchen Ausreißern beeinflusst wird, stellen sie eine nahe liegende Erklärung für das schlechte Abschneiden von Argamons „Quadratic Delta“ Delta $_{Q}$. Der positive Effekt der Vektor-Normalisierung wäre dann so zu deuten, dass durch die Vereinheitlichung der Vektorlängen der Betrag der *z*-Werte von textspezifischen Ausreißern deutlich reduziert wird (Ausreißer-Hypothese).
- (H2) Das charakteristische stilistische Profil eines Autors findet sich eher in der qualitativen Kombination bestimmter Wortpräferenzen, also im grundsätzlichen Muster von über- bzw. unterdurchschnittlich häufigem Gebrauch der Wörter, als in der Amplitude dieser Abweichungen. Ein Textabstandsmaß ist vor allem dann erfolgreich, wenn es strukturelle Unterschiede der Vorlieben eines Autors erfasst, ohne sich davon beeinflussen zu lassen, wie stark das Autorenprofil in einem bestimmten Text ausgeprägt ist (Schlüsselprofil-Hypothese). Diese Hypothese erklärt unmittelbar, warum die Vektor-Normalisierung zu einer so eindrucksvollen Verbesserung führt: durch sie wird die Amplitude des Autorenprofils in verschiedenen Texten vereinheitlicht.

Neue Erkenntnisse

Korpora

Für die hier präsentierten Untersuchungen verwenden wir drei vergleichbar aufgebaute Korpora in Deutsch, Englisch und Französisch. Jedes Korpus enthält je 3 Romane von 25 verschiedenen Autoren, insgesamt also jeweils 75 Texte. Die deutschen Romane aus dem 19. und dem Anfang des 20. Jahrhunderts stammen aus der Digitalen Bibliothek von TextGrid . Die englischen Texte aus den Jahren 1838 bis 1921 kommen von Project Gutenberg und die französischen Romane von Ebooks libres et gratuits umfassen den Zeitraum von 1827 bis 1934. Im folgenden Abschnitt stellen wir aus Platzgründen nur unsere Beobachtungen für das deutsche Romankorpus vor. Die Ergebnisse mit Texten in den beiden anderen Sprachen bestätigen – mit kleinen Abweichungen – unseren Befund.

Experimente

Um die Rolle von Ausreißern und damit die Plausibilität von H1 näher zu untersuchen, ergänzen wir Delta $_{Bur}$ und Delta $_{Q}$ um weitere Delta-Varianten, die auf dem allgemeinen Minkowski-Abstand basieren:

$$\Delta p = \sum i = 1 m z i D 1 - z i (D 2) p 1 / p \text{ für } p \geq 1.$$

Wir bezeichnen diese Abstandsmaße allgemein als L p-Delta. Der Spezialfall $p = 1$ entspricht dem Manhattan-

Abstand (also L_1-Delta = Delta$_{Bur}$), der Spezialfall $p = 2$ dem Euklidischen Abstand (also L_2-Delta = Delta$_Q$). Je größer p gewählt wird, desto stärker wird L p-Delta von einzelnen Ausreißerwerten beeinflusst.

Abbildung 2 vergleicht vier unterschiedliche L p-Abstandsmaße (für p = 1 , 2 , 2 , 4) mit Delta$_{Cos}$. Wir übernehmen dabei den methodologischen Ansatz von Evert et al. (2015): die 75 Texte werden auf Basis der jeweiligen Delta-Abstände automatisch in 25 Cluster gruppiert; anschließend wird die Güte der Autorenschaftszuschreibung mit Hilfe des *adjusted Rand index* (ARI) bestimmt. Ein ARI-Wert von 100% entspricht dabei einer perfekten Erkennung der Autoren, ein Wert von 0% einem rein zufälligen Clustering. Offensichtlich nimmt die Leistung von L p-Delta mit zunehmendem p ab; zudem lässt die Robustheit der Maße gegenüber der Anzahl von MFW erheblich nach.

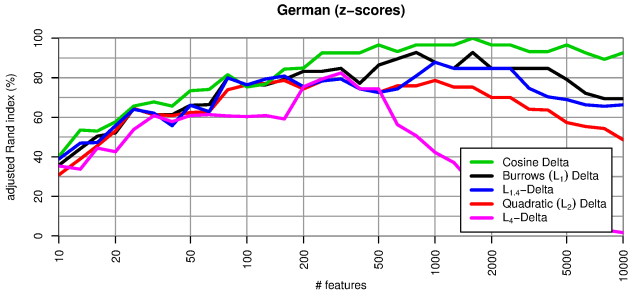


Abb. 2: Clustering-Qualität verschiedener Delta-Maße in Abhängigkeit von der Anzahl von MFW, die als Merkmale verwendet werden. Wie bereits von Janndis et al. (2015) und Evert et al. (2015) festgestellt wurde, liefert Delta$_{Bur}$ (L_1) durchgänig bessere Ergebnisse als Argamons Delta$_Q$ (L_2). Delta$_Q$ erweist sich als besonders anfällig gegenüber einer zu großen Anzahl von MFW. Delta$_{Cos}$ ist in dieser Hinsicht robuster als alle anderen Delta-Varianten und erreicht über einen weiten Wertebereich eine nahezu perfekte Autorenschaftszuschreibung (ARI > 90%).

Eine Vektor-Normalisierung verbessert die Qualität aller Delta-Maße erheblich (vgl. Abb. 3). Argamons Delta$_Q$ ist in diesem Fall identisch zu Delta$_{Cos}$: die rote Kurve wird von der grünen vollständig überdeckt. Aber auch andere Delta-Maße (Delta$_{Bur}$, $L_{1.4}$-Delta) erzielen praktisch dieselbe Qualität wie Delta$_{Cos}$. Einzig das für Ausreißer besonders anfällige L_4-Delta fällt noch deutlich gegenüber den anderen Maßen ab. Diese Ergebnisse scheinen zunächst H1 zu bestätigen.

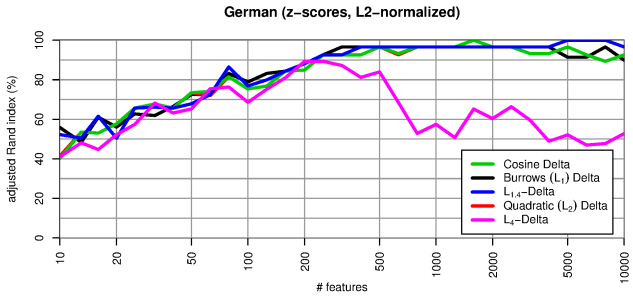


Abb. 3: Clustering-Qualität verschiedener Delta-Maße mit Längen-Normalisierung der Vektoren. In diesem Experiment wurde die euklidische Länge der Vektoren vor Anwendung der Abstandsmaße auf den Standardwert 1 vereinheitlicht.

Ein anderer Ansatz zur Abmilderung von Ausreißern besteht darin, besonders extreme z-Werte „abzuschneiden“. Wir setzen dazu alle $|z| > 2$ (ein übliches Ausreißerkriterium) je nach Vorzeichen auf den Wert +1 oder –1. Abbildung 4 zeigt, wie sich unterschiedliche Maßnahmen auf die Verteilung der Merkmalswerte auswirken. Die Vektor-Normalisierung (links unten) führt nur zu minimalen Änderungen und reduziert die Anzahl von Ausreißern praktisch nicht. Abschneiden großer z-Werte wirkt sich nur auf überdurchschnittlich häufige Wörter aus (rechts oben). Wie in Abbildung 5 zu sehen ist, wird durch diese Maßnahme ebenfalls die Qualität aller L p-Deltamaße deutlich verbessert. Der positive Effekt fällt aber merklich geringer aus als bei der Vektor-Normalisierung.

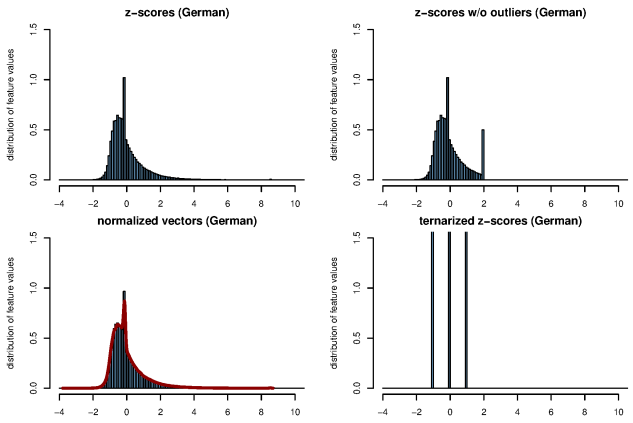


Abb. 4: Verteilung von Merkmalswerten über alle 75 Texte bei Vektoren mit 5000 MFW. Gezeigt wird die Verteilung der ursprünglichen z-Werte (links oben), die Verteilung nach einer Längen-Normalisierung (links unten), die Verteilung beim Abschneiden von Ausreißern mit $|z| > 2$ (rechts oben) sowie eine ternäre Quantisierung in Werte –1, 0 und +1 (rechts unten). Im linken unteren Bild gibt die rote Kurve die Verteilung der z-Werte ohne Vektor-Normalisierung wieder; im direkten Vergleich ist deutlich zu erkennen, dass die Normalisierung nur einen minimalen Einfluß hat und Ausreißer kaum reduziert.

Grenzwerte für die ternäre Quantisierung sind $z < -0.43$ (–1), $-0.43 \leq z \leq 0.43$ (0) und $z > 0.43$ (+1). Diese Grenzwerte sind so gewählt, dass bei einer idealen Normalverteilung jeweils ein Drittel aller Merkmalswerte in die Klassen –1, 0 und +1 eingeteilt würde.

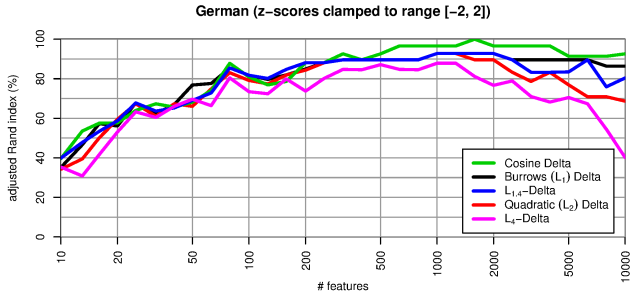


Abb. 5: Clustering-Qualität nach „Abschneiden“ von Ausreißern, bei dem Merkmalswerte $|z| > 2$ je nach Vorzeichen durch die festen Werte –2 bzw. +2 ersetzt wurden.

Insgesamt erweist sich Hypothese H1 somit als nicht haltbar. H2 wird durch das gute Ergebnis der Vektor-Normalisierung unterstützt, kann aber nicht unmittelbar erklären, warum auch das Abschneiden von Ausreißern zu einer deutlichen Verbesserung führt. Um diese Hypothese weiter zu untersuchen, wurden reine „Schlüsselprofil“-Vektoren erstellt, die nur noch zwischen überdurchschnittlicher (+1), unauffälliger (0) und unterdurchschnittlicher (–1) Häufigkeit der Wörter unterscheiden (vgl. Abb. 4, rechts unten).

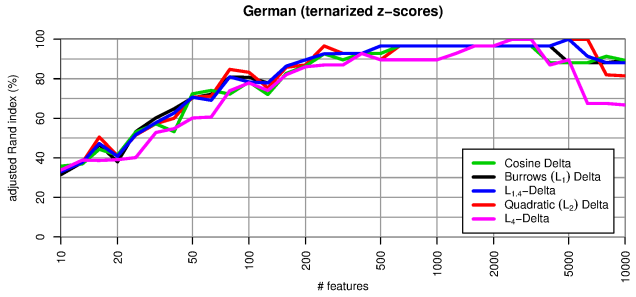


Abb. 6: Clustering-Qualität bei ternärer Quantisierung der Vektoren in überdurchschnittliche (+1, bei $z > 0.43$), unauffällige (0, bei $-0.43 < z < 0.43$) und unterdurchschnittliche (–1, bei $z < -0.43$) Häufigkeit der Wörter.

Abbildung 6 zeigt, dass solche Profil-Vektoren hervorragende Ergebnisse erzielen, die der Vektor-Normalisierung praktisch ebenbürtig sind. Selbst das besonders anfällige L_4-Deltamaß erzielt eine weitgehend robuste Clustering-Qualität von über 90%. Wir interpretieren diese Beobachtung als eine deutliche Bestätigung der Hypothese H2.

Diskussion und Ausblick

H1, die Ausreißerhypothese, konnte widerlegt werden, da die Vektor-Normalisierung die Anzahl von Extremwerten kaum verringert und dennoch die Qualität aller L p-Maße deutlich verbessert wird. H2, die Schlüsselprofil-Hypothese, konnte dagegen bestätigt werden. Die ternäre Quantisierung der Vektoren zeigt deutlich, dass nicht das Maß der Abweichung bzw. die Größe der Amplitude wichtig ist, sondern das Profil der Abweichung über die MFW hinweg. Auffällig ist das unterschiedliche Verhalten der Maße, wenn mehr als 2000 MFW verwendet werden. Fast alle Varianten zeigen bei sehr vielen Features eine Verschlechterung, aber sie unterscheiden sich darin, wann dieser Verfall einsetzt. Wir vermuten, dass das Vokabular in diesem Bereich weniger spezifisch für den Autor, und eher für Themen und Inhalte ist. Die Klärung dieser Fragen wird zusätzliche Experimente erfordern.

Burrows’ Delta im Mittelalter: Wilde Graphien und metrische Analysedaten

Einleitung

Burrows’ Delta (Burrows 2002) hat sich in Autorschaftsfragen etabliert; viele Studien zeigen, dass Delta für germanische Sprachen ausgezeichnet funktioniert (Hoover 2004b; Eder / Rybicki 2011; Eder 2013a; Eder 2013b; für das Neuhochdeutsche zuletzt Jannidis / Lauer 2014; Evert et al. 2015). Beim Mittelhochdeutschen ist jedoch die Schreibung nicht normiert: Das Wort „und“ kann als „unde“, „unt“ oder „vnt“ verschriftet sein. Ein Teil dieser Varianz wird zwar in normalisierten Ausgaben ausgeglichen, jedoch nicht vollständig. Viehhauser (2015) hat in einer ersten Delta-Studie zum Mittelhochdeutschen diese Probleme diskutiert: Wolfram von Eschenbach benutzt zum Wort „kommen“ die Präteritalform „kom“, Hartmann von Aue verwendet „kam“, eine Form, die eher in den südwestdeutschen Raum gehört. Die Bedingungen für den Einsatz von Delta auf der Basis der *most frequent words* erscheinen auf den ersten Blick also als denkbar ungünstig; Viehhauser war skeptisch, inwieweit Autor, Herausgeber, Schreibereinflüsse oder Dialekt erfasst werden, auch wenn seine Ergebnisse zeigen, dass Delta Texte von gleichen Autoren korrekt sortiert.

Normalisierte Texte sind besser für Autorschaftsstudien geeignet, da hier die Zufälligkeiten von Schreibergraphien reduziert sind; Längenzeichen stellen dort meist weitere lexikalische Informationen zur Verfügung – etwa zur Differenzierung von „sin“ („Sinn“) versus „sîn“ („sein“; allerdings ohne Disambiguierung von „sîn“ als verbum substantivum oder Pronomen). In diplomatischen Transkriptionen sind dagegen etwa „u-e“ Superskripte und andere diakritische Zeichen enthalten;

die gleiche Flexionsform des gleichen Wortes kann in verschiedenen Graphien erscheinen.

Anlass zu vorsichtigem Optimismus bietet allerdings eine Studie von Eder (2013a), die den Einfluss von Noise (wie z. B. Schreibervarianten) analysiert – mit dem Ergebnis (u. a.) für das Neuhochdeutsche, dass ein zufälliger Buchstabentausch von 12% bei 100-400 MFWs die Ergebnisse kaum beeinträchtigt; bei einer mäßig randomisierten Manipulation der MFWs-Frequenzen verschlechtert sich die Quote der korrekten Attributionen bei 200-400 MFWs ebenfalls kaum. Ersetzt man im Autortext Passagen durch zufällig gewählte Passagen anderer Autoren, ergibt sich bei der Quote lediglich ein „gentle decrease of performance"; im Lateinischen bleibt die Quote gut, selbst nachdem 40% des Originalvokabulars ausgetauscht wurden.

Während die 17 Texte, die Viehhauser analysiert hat, in normalisierten Ausgaben vorliegen, habe ich zunächst 37 heterogene Texte von sieben Autoren mit Stylo-R (Eder / Kestemont et al. 2015) getestet sowie drei Texte mit fraglicher Autorzuschreibung zu Konrad von Würzburg. Ein Teil ist normalisiert (Hartmann, Wolfram, Gottfried, Ulrich, Wirnt, Konrad), andere liegen zum Teil in diplomatischen Transkriptionen vor: Bei Rudolf von Ems sind ‚Gerhard', ‚Alexander' und ‚Barlaam' normalisiert, nicht normalisiert sind ‚Willehalm' und ‚Weltchronik' (hier etwa „ubir" statt „über"). Beim Stricker ist lediglich der ‚Pfaffe Amis' normalisiert.

Per Skript wurden Längenzeichen eliminiert, damit nicht Texte mit und ohne Längenzeichen auseinander sortiert werden. Tustep-Kodierungen etwa für Superskripte habe ich in konventionelle Buchstaben transformiert. Dennoch bleiben große Unterschiede: Die Genitivform zu „Gott" lautet teils „gotes", teils „gotis", so dass eigentlich eine primäre Sortierung entlang der Unterscheidung normalisiert–nicht-normalisiert zu erwarten wäre. Das Ergebnis ist jedoch frappierend: Auf der Basis von 200 MFWs (diesen Parameter verwenden auch Eder 2013b und Viehhauser 2015) gelingt stylo-R ohne Pronomina und bei Culling=50% eine fehlerfreie Sortierung nach Autorschaft; Delta ordnet Rudolf zu Rudolf – ob normalisiert oder nicht.

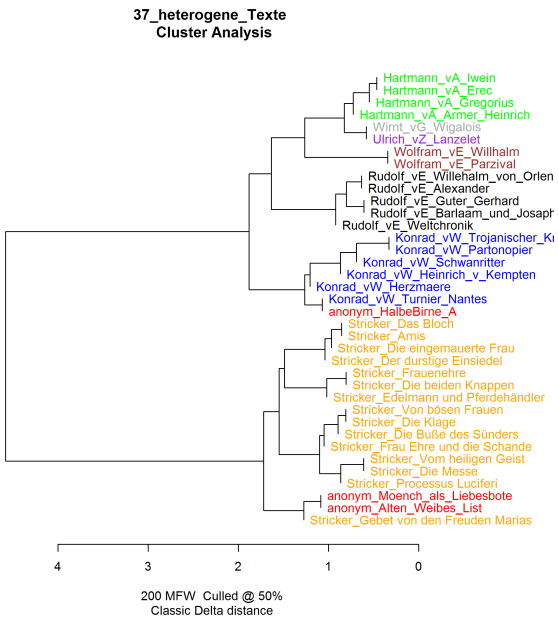


Abb. 7: Clusteranalyse

Validierungstests

Dieser Befund ist Anlass für eine Serie an automatisierten Tests in Anlehnung an Eder (2013b): Bei welchem Vektor und ab welcher Textlänge liefert Delta zuverlässige Ergebnisse? Wie wirkt sich das Einbringen von Noise aus?

Vektorlänge

Per Perlskript wurde ein Delta-Test implementiert, der in einer großen Zahl an Iterationen (13.425 Delta-Berechnungen) verschiedene „Ratetexte" mit bekannter Autorschaft gegen ein Validierungskorpus mit bekannter Autorschaft jeweils daraufhin prüft, ob für jeden Text im Ratekorpus tatsächlich der niedrigste Delta-Wert bei einem Text des gleichen Autors im Validierungskorpus herauskommt. Gegen ein heterogenes Validierungskorpus mit 18 Texten wurden 19 normalisierte Ratetexte getestet; gegen ein heterogenes Validierungskorpus mit 15 Texten wurden 13 nicht-normalisierte Ratetexte getestet. Ermittelt wurde der Prozentsatz der richtig erkannten Autoren für jeweils eine Vektorlänge; die Vektorlänge wurde in 100er Schritten bis auf 2.500 MFWs erhöht. Pronomina wurden beseitigt. Bei den normalisierten Ratetexten ist die Erkennungsquote sehr gut bis 200–900 MFWs, bei den nicht-normalisierten sehr gut für 100–600 MFWs.

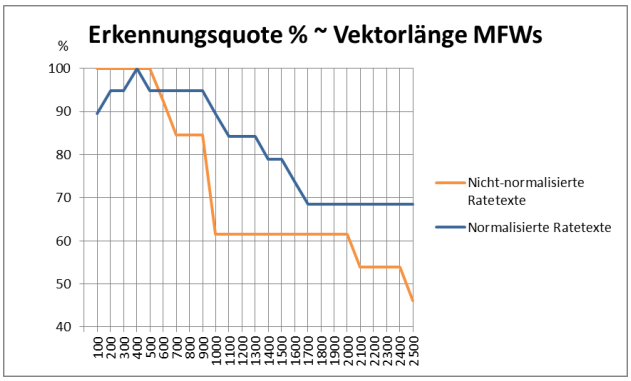


Abb. 8: Vektorlänge

Interessante Fehlattributionen – etwa Strickers ‚Pfaffe Amis' und Konrads ‚Herzmäre' – machen weitere Validierungsläufe nötig: Der normalisierte ‚Pfaffe Amis' wurde gegen einen nicht-normalisierten Stricker-Text getestet; das ‚Herzmäre' ist kurz (2991 Wörter). Während Burrows (2002) davon ausgeht, dass Delta ab einer Textlänge von 1.500 Wörtern anwendbar ist, zeigt Eder (2013b), dass Delta im Englischen ab 5.000 Wörtern sehr gute und unter 3.000 Wörtern teils desaströse Ergebnisse liefert; nur im Lateinischen werden ab 2.500 Wörtern gute Ergebnisse erreicht.

Korrelation Vektorlänge und Textlänge in konventionellen Segmentierungen

Hier wurde die Textlänge linear begrenzt, die Texte wurden nach 1000, 2000 Wörtern usw. abgeschnitten. Das Korpus ist kleiner als zuvor, da zu kurze Texte herausgenommen wurden (normalisierte: 16 Texte Validierungskorpus, 15 Ratekorpus; nicht-normalisierte 14 Validierungskorpus, 6-7 Ratekorpus; 10.056 Delta-Berechnungen).

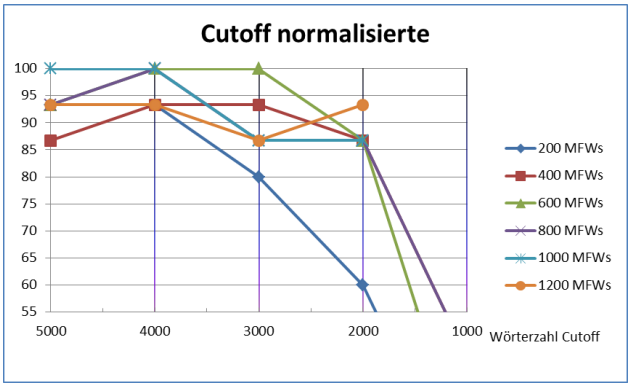


Abb. 9: Cutoff normalisierte

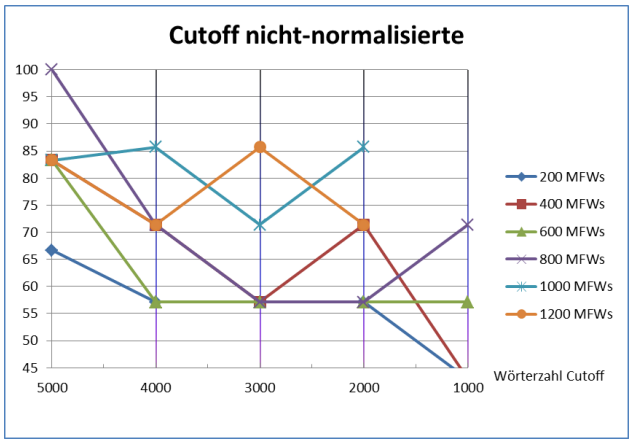


Abb. 10: Cutoff nicht-normalisierte

Korrelation Vektorlänge und Textlänge bei randomisierter Wortauswahl (‚bag-of-words'; vgl. Eder 2013b)

Gleiches Korpus wie zuvor; 167.600 Delta-Berechnungen. Da die bag-of-words randomisiert zusammengestellt wird, schwankt die Erkennungsquote etwas, daher wurde jeder Test pro Textlänge und Wortlistenlänge 25x durchgeführt und der Mittelwert dieser 25 Erfolgsquoten verwendet.

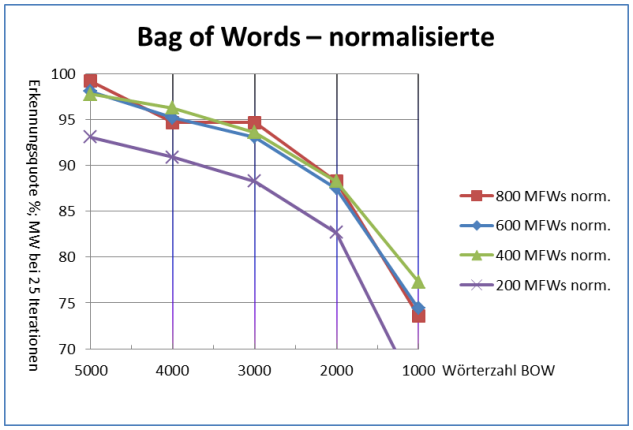


Abb. 11: bag-of-words normalisierte

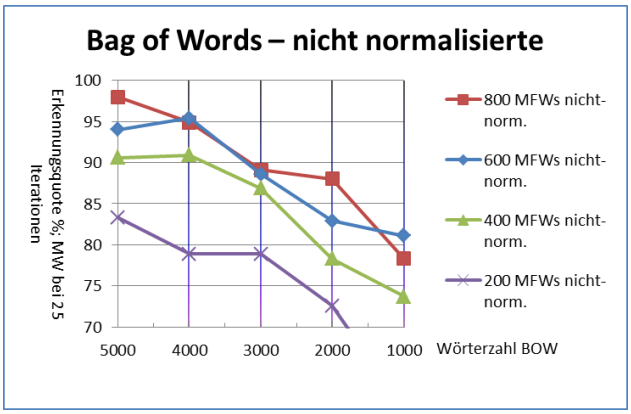


Abb. 12: bag-of-words nicht-normalisierte

Auswirkung bei der Eliminierung von Pronomina

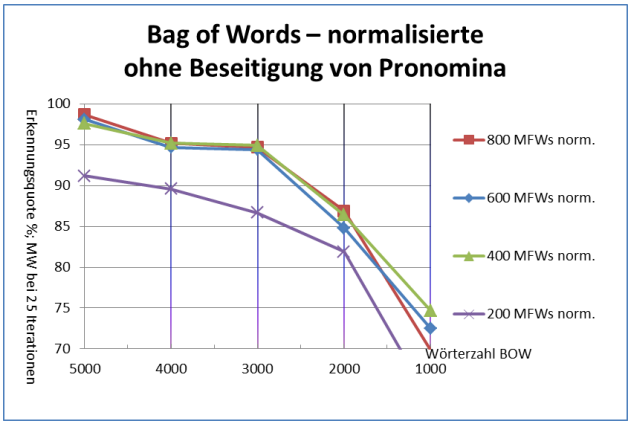


Abb. 13: bag-of-words normalisierte, ohne Beseitigung der Pronomina

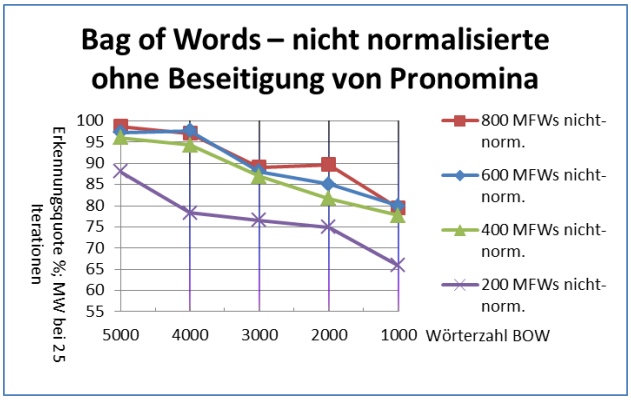


Abb. 14: bag-of-words nicht-normalisierte, ohne Beseitigung der Pronomina

Auswirkungen beim Hinzufügen von Noise

Aus einer Noise-Datei mit >18.000 mittelhochdeutschen und altfranzösischen Wortformen ohne Duplikate werden die Ratetexte prozentual aufsteigend randomisiert: Teile der bag-of-words werden gegen fremdes Sprachmaterial ausgetauscht, um Fehler in der Überlieferungskette zu simulieren. Die Kurve verläuft nicht konstant linear, da für jede bag-of-words-Berechnung erneut Noise randomisiert hinzugefügt wird (hier 10 Iterationen pro Einzelwert; 1.179.360 Delta-Berechnungen).

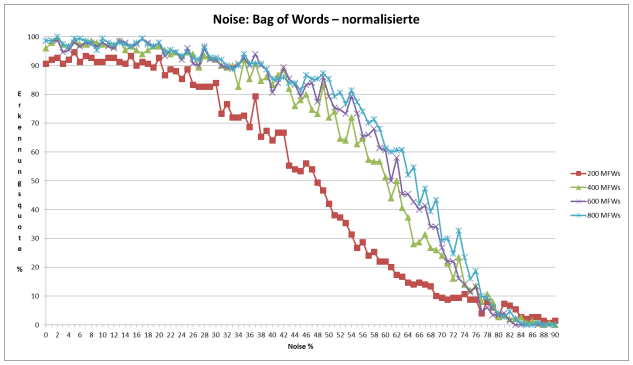


Abb. 15: Noise bei normalisierten

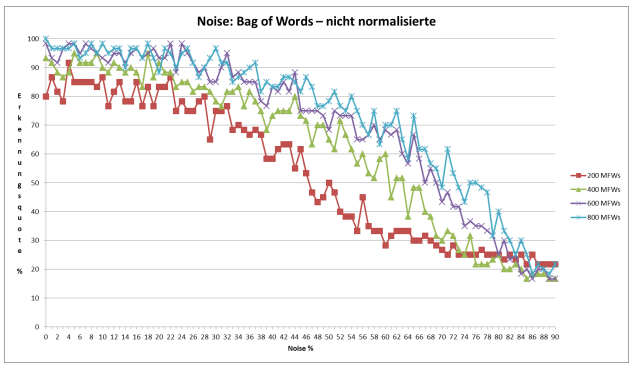


Abb. 16: Noise bei nicht-normalisierten

Beim Test der Vektorlänge (vgl. 3.2.1.) bleiben die Erkennungsquoten bei normalisierten Ratetexten sehr gut bis 200–900 MFWs. Bei den nicht-normalisierten Texten sind die Quoten nur für einen kleineren Bereich sehr gut: für 100–600 MFWs. Bei einer Begrenzung der Textlänge (Cutoff; vgl. 3.2.2.) bleiben die Ergebnisse bei normalisierten Texten nur ab 4000 Wörtern Textlänge weitgehend gut bis sehr gut. Schlecht sieht es bei den nicht-normalisierten Texten aus: Sehr gut ist die Quote nur bei 800 MFWs und 5000 Wörtern, ansonsten weithin desaströs. Bag-of-words (vgl. 3.2.3.) bieten dagegen stabilere Ergebnisse: Bei den normalisierten Texten sind bei einer Textlänge von 5000 die Quoten sehr gut bei

400-800 MFWs. Bei den nicht-normalisierten Texten ist die Quote wiederum nur bei Textlänge 5000 und 800 MFWs sehr gut. Bei kürzeren Texten und anderen Frequenzen verschlechtern sich die Quoten massiv, allerdings bleiben sie noch deutlich besser als beim Cutoff-Test. Bei normalisierten Texten werden durch das Eliminieren von Pronomina geringfügig bessere Quoten erreicht (vgl. 3.2.4.), bei nicht-normalisierten Texten etwas schlechtere Quoten.

Stabil bleiben die Quoten bei normalisierten Texten nach dem Einbringen von Noise (vgl. 3.2.5.): Solange nicht mehr als 17% des Vokabulars ausgetauscht wurden, werden die Erkennungsquoten nur etwas schlechter. 600-800 MFWs liefern sehr gute Erkennungsquoten bis 20%. Auch die Quoten bei nicht-normalisierten Texte sind einigermaßen stabil, solange nicht mehr als 20% Noise eingebracht werden: Der Bereich von 600-800 MFWs liefert bis 9% Noise noch sehr gute und bis 22% noch gute Ergebnisse.

Die Stabilität der Erkennungsquoten gibt Grund zum Optimismus für eine Anwendbarkeit bei normalisierten mittelhochdeutschen Texten. Am besten geeignet ist der Vektorbereich von 400-800 MFWs bei langen Texten mittels bag-of-words. Auch wenn die Ergebnisse für nicht-normalisierte Texte etwas zurückfallen, hat mich angesichts der wilden mittelhochdeutschen Graphien doch überrascht, dass die Delta-Performanz derart robust bleibt. Während jedoch etwa Eder Validierungsstudien mit über 60 Texten durchführen konnte, ist es um die digitale Verfügbarkeit von längeren mittelhochdeutschen Texten, von denen mindestens zwei Texte vom gleichen Autor verfasst wurden, derzeit noch deutlich schlechter bestellt. Die Aussagekraft der vorliegenden Studien wird daher durch die Korpusgröße v. a. bei den nicht-normalisierten Texten limitiert.

Noise-Reduktion: Metrik-Delta

Bei einem weiteren Versuch geht es darum, die Einflüsse von Schreibergraphie und Normalisierungsart zu reduzieren, indem nicht der Wortschatz, sondern abstraktere Daten verwendet werden: Nach Hirst und Feiguina (2007) erzielen Tests auf der Basis von Part-of-Speech–Bigrammen gute Ergebnisse; für das Mittelhochdeutsche ist jedoch noch kein Part-of-Speech-Tagger in Sicht.

2014 habe ich das Metrik-Modul aus meiner Dissertation grundlegend überarbeitet, die Fehlerquote reduziert (nun unter 2%) und es ins Internet zur freien Benutzung eingestellt (Dimpel 2015). Dieses Modul gibt Kadenzen aus (etwa „weiblich klingend“). Die metrische Struktur wird mit „0“ (unbetonte Silben) und „1“ (betonte Silben) ausgegeben; der dritte ‚Parzival‘-Vers hat das Muster „01010011“. Anstatt mit MFWs habe ich Metrikmuster und Kadenzinformationen verwendet und so die Metrikdaten als „Worte“ testen lassen. Da ein weniger variationsreiches Ausgangsmaterial verwendet wird, habe ich wie Hirst und Feiguina (2007) mit Bigrammen gearbeitet.

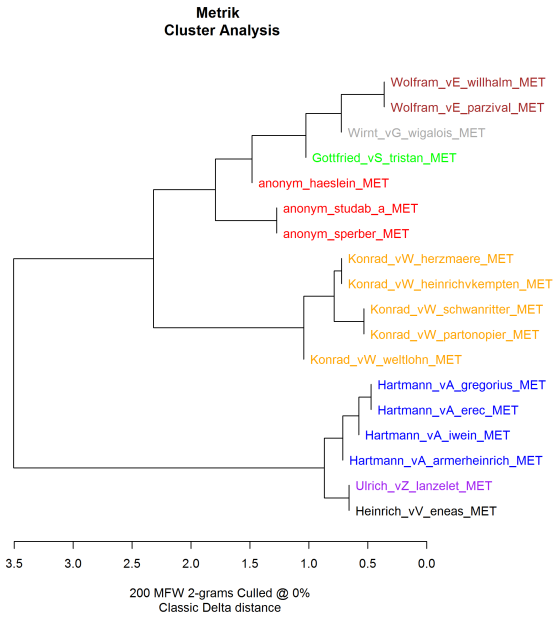


Abb. 17: Clusteranalyse auf Basis von Metrik-Bigrammen

Auch ein Metrik-Delta-Plot mit Stylo-R clustert Autoren hier fehlerlos. Validierungstests sind bislang nur mit einem kleineren Korpus möglich, da das Metrikmodul Längenzeichen benötigt und nur für Texte mit vierhebigen Reimpaarversen konstruiert ist. Bei 13 Ratedateien und 11 Validierungsdateien ergibt sich bei 250–300 „MFWs“ eine Erkennungsquote von 92,3%. Ein erfreuliches Ergebnis: (1) Bei Tests auf Grundlage von Metrik-Daten ist eine etwas geringere Abhängigkeit von Schreibergraphie und von Normalisierungsgewohnheiten gegeben. Zwar hat es mitunter metrische Eingriffe der Herausgeber gegeben, aber längst nicht immer. Wenn ein Herausgeber aus metrischen Gründen lieber das Wort „unde“ statt „unt“ verwendet, dann geht in den Metrik-Delta ein ähnlicher Fehler wie in den konventionellen Delta-Test ein. Immerhin immunisieren Metrik-Daten gegen Graphie-Varianten wie „und“ oder „unt“. (2) Zudem kann Autorschaft offenbar nicht nur mit dem vergleichsweise einfachen Parameter MFWs dargestellt werden: Nicht nur eine pure Wortstatistik führt zum Ziel, vielmehr erweist sich auch die Kompetenz zum philologischen Programmieren und zur filigranen Textanalyse als fruchtbar. (3) Bei der metrischen Struktur handelt es sich um ein Stilmerkmal, das Autoren oft intentional kunstvoll gestalten. Während es als communis opinio gilt, dass vor allem die unbewussten Textmerkmale wie MFWs autorspezifisch sind, gelingt es nun auch über ein wohl oft bewusst gestaltetes Stilmerkmal, Autorschaft zu unterscheiden. Man muss also den Dichtern nicht nur einen unbewussten stilistischen Fingerabdruck zutrauen, vielmehr lässt sich Autorschaft zumindest hier über ein

Merkmal erfassen, das dem bewussten künstlerischen Zugriff unterliegen kann.

Stilometrie interdisziplinär: Merkmalsselektion zur Differenzierung zwischen Übersetzer- und Fachvokabular

Einleitung

Stilometrie ist der Versuch, sprachliche Besonderheiten durch statistische Methoden herauszustellen und zu vergleichen, um damit unter anderem Rückschlüsse auf die Urheberschaft eines Textes ziehen zu können. Als probates Mittel bei der Autorschaftsattribuierung hat sich die Analyse der Verwendung der häufigsten Wörter bewährt. Insbesondere Varianten des von Burrows (2002) vorgeschlagenen Deltamaßes haben sich als sehr erfolgreich erwiesen (Hoover 2004a; Eder / Rybicki 2011). Faktoren der Zusammensetzung des Textkorpus, die sich negativ auf die Qualität der Ergebnisse auswirken können, sind unter anderem zu kurze Texte (Eder 2015), unterschiedliche Genres der Texte (Schöch 2013) und eine Überlagerung von Autor- und Übersetzerstilen (Rybicki 2012). Gerade inhaltliche Unterschiede zwischen Texten stellen ein Hindernis bei der Erkennung der Autoren dar, das nur mit erheblichem technischen Aufwand überwunden werden kann (Stamatatos et al. 2000; Kestemont et al. 2012).

In unserem Beitrag verwenden wir Deltamaße zur Identifikation von Übersetzern. Textgrundlage ist eine Sammlung von im 12. Jahrhundert entstandenen arabisch-lateinischen Übersetzungen wissenschaftlicher Texte aus verschiedenen Disziplinen. Wir zeigen eine Möglichkeit auf, wie die aus den oben genannten Faktoren resultierenden Limitierungen durch den Einsatz maschineller Lernverfahren kompensiert werden können. Gleichzeitig eröffnet sich dadurch eine Möglichkeit, unter den häufigsten Wörtern solche zu identifizieren, die eher Informationen zum Übersetzer oder eher zur Disziplin tragen.

Das Korpus

Die hier verwendete Textsammlung wurde mit dem philologischen Ziel angelegt, die Übersetzer zu identifizieren, die im 12. Jahrhundert eine Vielzahl von Texten aus dem Arabischen ins Lateinische übertragen und damit in den verschiedensten Disziplinen die weitere Entwicklung der europäischen Wissenschaften nachhaltig beeinflusst haben (Hasse / Büttner in Vorbereitung)[1] . Es handelt sich dabei um Texte unterschiedlicher Autoren aus den Bereichen Philosophie, Mathematik, Astronomie, Astrologie, Medizin, Geologie und Meteorologie, aber auch um religiöse, magische und alchemistische Traktate, wobei einzelne Texte nicht eindeutig einer Disziplin zugeordnet werden können. Elf der Übersetzer sind namentlich bekannt, fast die Hälfte der Texte ist jedoch nur anonym überliefert.

Für die Experimente wird ein Testkorpus so zusammengestellt, dass von jedem Übersetzer und aus jeder Disziplin mindestens drei Texte zur Verfügung stehen. Dieses besteht aus insgesamt 37 Texten von 5 Übersetzern, wobei die Texte aus 4 Disziplinen stammen (siehe Abb. 18). Das daraus resultierende Textkorpus ist nicht balanciert: Die Anzahl der Texte pro Übersetzer ist ungleich verteilt, die Länge der Texte liegt zwischen 500 und fast 200000 Wörtern; insgesamt sind die Texte auch deutlich kürzer als diejenigen der oft verwendeten Romankorpora (vgl. etwa Jannidis et al. 2015).

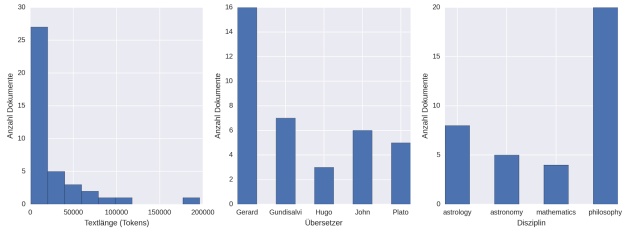


Abb. 18: Verteilung der Textlängen, Übersetzer und Disziplinen im verwendeten Teilkorpus

Weitere, die Analyse erschwerende Faktoren sind Doppelübersetzungen desselben Originaltextes durch zwei Übersetzer und die – historisch nicht völlig klar belegte – Zusammenarbeit einiger Übersetzer. Auf der anderen Seite sind die unterschiedlichen Disziplinen prinzipiell klarer und eindeutiger unterscheidbar als literarische Subgenres in Romankorpora.

Methoden

Delta-Maße

Ausgehend von Burrows ursprünglichem Deltamaß (Burrows 2002) wurde eine ganze Reihe von Deltamaßen für die Autorschaftszuschreibung vorgeschlagen (bspw. Hoover 2004b, Argamon 2008, Smith / Aldridge 2011, Eder et al. 2013). Alle Maße operieren auf einer Term-Dokument-Matrix der *n* häufigsten Terme im Korpus, die die relativen Häufigkeiten der Terme in den einzelnen Dokumenten enthält. In einem ersten Schritt werden die relativen Häufigkeiten der Terme standardisiert (üblicherweise durch eine *z*-Transformation) um die Größenordnungsunterschiede, die sich durch die Zipfsche Verteilung der Worthäufigkeiten ergeben, zu beseitigen. Im optionalen zweiten Schritt können die Dokumentvektoren normalisiert, d. h. auf Länge 1 gebracht werden. Im dritten Schritt wird die

Ähnlichkeit zwischen zwei Dokumentvektoren durch ein Ähnlichkeits- oder Abstandsmaß bestimmt (bei Burrows Delta wird bspw. die Manhattan-Distanz verwendet, bei Kosinus-Delta der Kosinus des Winkels zwischen den beiden Dokumentvektoren). Auf Basis der so erhaltenen Ähnlichkeitswerte können die Dokumente dann geclustert werden, wobei idealerweise Texte desselben Autors im selben Cluster landen.

Für die folgenden Experimente verwenden wir Kosinus-Delta, das sich unter anderem bei Jannidi, Pielström, Schöch und Vitt (2015) sowie Evert, Proisel und Jannidis et al. (2015) als das robusteste Mitglied der Delta-Familie erwiesen hat.

Rekursive Merkmalseliminierung

Rekursive Merkmalseliminierung (recursive feature elimination, RFE) ist eine von Guyon, Weston, Barnhill und Vapnik (2002) vorgeschlagene Methode zur Selektion einer möglichst kleinen Teilmenge von Merkmalen, mit der trotzdem möglichst optimale Ergebnisse mit einem überwachten maschinellen Lernverfahren erzielt werden können. Evert, Proisel und Jannidis et al. (2015) experimentieren zur Autorschaftszuschreibung mit durch RFE ermittelten Termen als Alternative zu den üblichen *n* häufigsten Termen.

Da RFE auf einem überwachten Lernverfahren (üblicherweise einem *Support Vector Classifier*) basiert, müssen zumindest für eine Teilmenge der Dokumente die wahren Autoren bzw. Übersetzer bekannt sein. Das rekursive Verfahren trainiert zunächst den Klassifikator auf allen Merkmalen (Termen), wobei den einzelnen Merkmalen Gewichte zugeordnet werden. Anschließend werden die *k* Merkmale mit den niedrigsten absoluten Gewichten entfernt (*pruning*). Die Schritte Training und *pruning* werden nun auf den verbleibenden Merkmalen so lange wiederholt, bis die gewünschte Anzahl von Merkmalen übrigbleibt. Alternativ kann durch Kreuzvalidierung die optimale Merkmalsmenge bestimmt werden.

In den folgenden Experimenten kombinieren wir beide Varianten und verkleinern die Merkmalsmenge (also die Menge der verwendeten Wörter) zunächst schrittweise auf die 500 besten Merkmale, um anschließend die optimale Merkmalsmenge zu bestimmen.

Experimente

Zunächst führen wir mit dem Testkorpus einige Versuche zur Anpassung der stilometrischen Methoden durch. Als Maß der Qualität des Clusterings dient dabei der *Adjusted Rand Index (ARI)*, der zwischen -1 und 1 liegen kann. Ein vollständig korrektes Clustering erhält einen ARI von 1, eine zufällige Gruppierung der Elemente einen ARI um 0, und negative Werte weisen auf ein Clustering hin, das schlechter als zufällig ist. Wie in Abbildung 19 dargestellt, wird bei Verwendung von Kosinus-Delta der höchste ARI für das Clustering der Übersetzer bereits bei etwa 300–400 der häufigsten Wörter erreicht, bei über 1000 Wörtern fällt das Qualitätsmaß stark ab. Ein Clustering nach Disziplinen hingegen erreicht bei ca. 500–700 Wörtern die besten Ergebnisse. Es fällt auf, dass zum einen die besten Ergebnisse mit viel kleineren Wortmengen erreicht werden als bei Studien zur Autorschaftszuschreibung, und dass zum anderen die Ergebnisse deutlich schlechter sind.

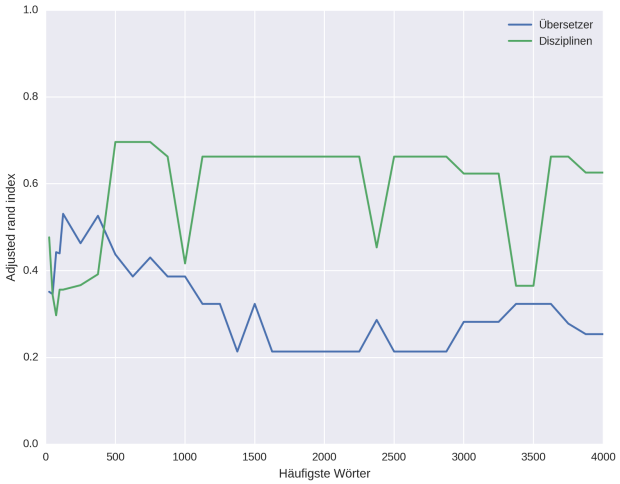


Abb. 19: Clusteringqualität in Abhängigkeit von der Anzahl der häufigsten Wörter

Da das Hauptziel eine korrekte Zuordnung der Übersetzer ist, soll die Menge der 500 häufigsten Wörter (im Folgenden *MFW500*), mit der ein ARI $_{Ü}$ von 0,437 erreicht wird, als Vergleichsmaßstab für die weiteren Versuche dienen. Für ein Clustering nach Disziplinen wird mit diesen Wörtern ein ARI $_{D}$ von 0.696 erreicht.

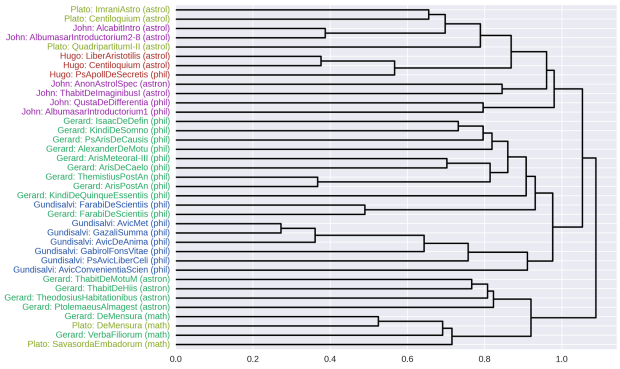


Abb. 20: Dendrogramm für das Clustering mit MFW500, Einfärbung nach Übersetzern

Durch RFE wählen wir aus der Gesamtmenge weniger als 500 Wörter aus. Mit 483 Wörtern ist eine perfekte Klassifikation nach Übersetzern möglich. Wenig überraschend erzielen wir mit diesen Wörtern auch ein perfektes Clustering der Texte nach Übersetzern

(ARI Ü=1,0). Auch für die Disziplinen lässt sich eine Menge von 475 Wörtern finden, bei der die Texte sich perfekt aufteilen lassen (ARI D=1,0). Da die durch RFE bestimmten Wörter teilweise sehr spezifisch sind und dadurch zu befürchten ist, dass Merkmale selektiert werden, die jeweils nur zwei Texte aneinander binden oder voneinander trennen, wählen wir aus den für die Übersetzer RFE-selektierten Merkmalen diejenigen aus, die auch in MFW500 enthalten sind. Mit diesen 68 Wörtern ist immer noch eine sehr gute, wenn auch nicht perfekte Unterscheidung der Übersetzer möglich (ARI Ü=0,910). Disziplinen lassen sich mit diesen Merkmalen nur sehr schlecht unterscheiden (ARI D=0,162).

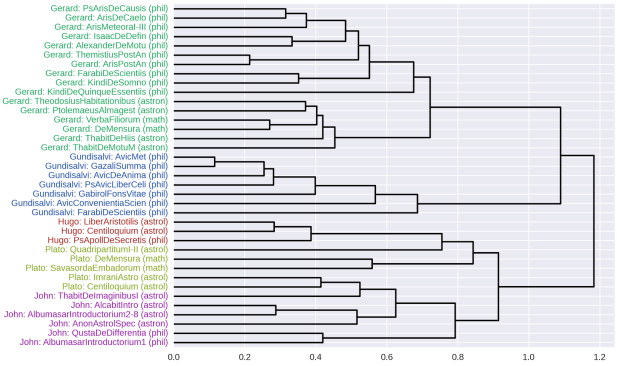


Abb. 21: Dendrogramm für das Clustering mit der Schnittmenge aus RFE und MFW500, Einfärbung nach Übersetzern

Die Analyse der *z*-Werte dieser Wörter zeigt, dass diese überwiegend bei nur einem einzigen Übersetzer besonders häufig sind. Sie lassen sich deshalb zu dem Übersetzer gruppieren, in dessen Texten der Mittelwert dieser z-Werte am höchsten ist, wodurch sich für jeden Übersetzer eine Liste von spezifischen bevorzugten Wörtern ergibt.

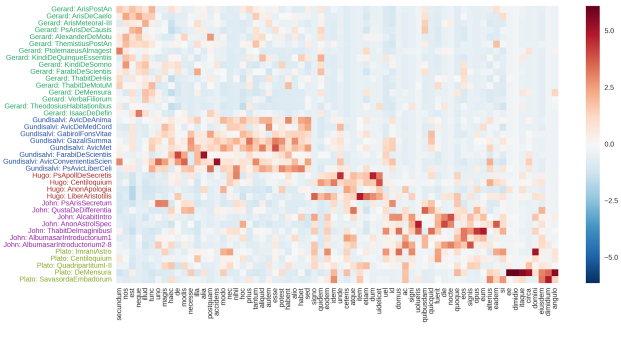


Abb. 22: Heatmap der *z*-Werte aus der Schnittmenge von RFE und MFW500

Die 432 Wörter aus MFW500, die in der Menge der RFE-selektierten Wörter nicht enthalten sind, unterscheiden, wie erwartet, deutlich schlechter zwischen Übersetzern (ARI Ü=0,222), dafür aber sehr gut zwischen Disziplinen (ARI D=0,727) – überraschenderweise sogar deutlich besser als alle 500 Wörter aus MFW500.

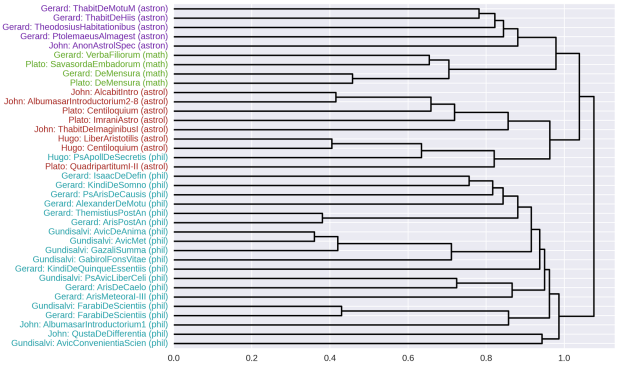


Abb. 23: Dendrogramm für das Clustering mit der Differenzmenge aus MFW500 und RFE, Einfärbung nach Disziplinen

Bei den Disziplinen erzielt die Schnittmenge der dafür mit RFE ausgewählten Wörter mit MFW500 sogar perfekte Ergebnisse (Anzahl der Merkmale: 109, ARI=1,0). Die Differenzmenge zeigt hier allerdings nicht den oben beschriebenen Effekt. Zwar ist die Clusteringqualität nach Disziplinen deutlich schlechter als der mit MFW500 erzielte Wert (ARI D=0,384), die nach Übersetzern aber ebenfalls (ARI Ü=0,198).

Um die Robustheit der Ergebnisse zu prüfen und insbesondere gegen ein Overfitting durch das RFE-Verfahren abzusichern, kann das bisher Beschriebene mit einem in ein Trainingsset und ein Testset aufgeteilten Korpus wiederholt werden, wobei die RFE-selektierten Wörter aus dem Trainingsset bestimmt und im Testset getestet werden. Dabei lassen sich die mit dem Gesamtkorpus beschriebenen Effekte reproduzieren, wenn auch – aufgrund der dann sehr kleinen Textanzahl – in schwächerer Ausprägung.

Ergebnisse

Durch die Experimente wurde gezeigt, dass sich die Menge der *n* häufigsten Wörter, die üblicherweise zur Autorschaftszuschreibung verwendet wird, so in zwei Teilmengen partitionieren lässt, dass die eine die Identifikation der Übersetzer der Texte besser ermöglicht als die Gesamtmenge, während die Wörter aus der anderen Teilmenge zur Identifizierung von Disziplinen verwendet werden können. Die rekursive Merkmalseliminierung erwies sich dabei als wirksames Mittel zur Differenzierung zwischen den zur Bestimmung des Verfassers relevanten und den durch die unterschiedlichen Inhalte der Texte bedingten Merkmalen. Darüber hinaus bietet eine solche Kondensierung der Wortliste die Chance, von einer aus philologischer Sicht undurchschaubaren statistischen

Maschinerie zu tatsächlich durch den Leser der Texte intuitiv nachvollziehbaren Kriterien zu gelangen.

Weitere Experimente in diesem Kontext werden dem Versuch dienen, die unterscheidenden Wörter besser zu charakterisieren, sodass idealerweise auch ohne maschinelles Lernen eine Auswahl der Merkmale möglich wird. Zudem steht eine Anwendung der Methode auf andere Textkorpora aus.

Notes

1. Siehe hierzu auch die Projekthomepage des Digital Humanities-Zentrums KALLIMACHOS der Universität Würzburg http://kallimachos.de/project/doku.php/kallimachos:identifizierunguebersetzer:start

Bibliographie

Argamon, Shlomo (2008): "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations", in: *Literary and Linguistic Computing* 23, 2: 131–47. 10.1093/llc/fqn003 .

Baeza-Yates, Ricardo / Ribeiro-Neto, Berthier (1999): *Modern Information Retrieval*. Harlow: Addison-Wesley.

Burrows, John (2002): "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship", in: *Literary and Linguistic Computing* 17, 3: 267–87. 10.1093/llc/17.3.267 .

Dimpel, Friedrich Michael (2015): "Automatische Mittelhochdeutsche Metrik 2.0", in: *Philologie im Netz* 73: 1–26 http://web.fu-berlin.de/phin/phin73/p73i.htm [letzter Zugriff 26. Januar 2016].

Eder, Maciej (2013a): "Mind Your Corpus: systematic errors in authorship attribution", in: *Literary and Linguistic Computing* 28: 603-614. 10.1093/llc/fqt039 .

Eder, Maciej (2013b): "Does size matter? Authorship attribution, small samples, big problem", in: *Literary and Linguistic Computing Advanced Access* 29: 1-16. 10.1093/llc/fqt066 .

Eder, Maciej (2015): "Does size matter? Authorship attribution, small samples, big problem", in: *Digital Scholarship Humanities* 30, 2: 167–182. 10.1093/llc/fqt066 .

Eder, Maciej / Kestemont, Mike / Rybicki, Jan (2013): "Stylometry with R: a suite of tools", in: *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska 487–489 http://dh2013.unl.edu/abstracts/ab-136.html [letzter Zugriff 26. Januar 2016].

Eder, Maciej / Kestemont, Mike / Rybicki, Jan(2015): "stylo R package" https://sites.google.com/site/computational-stylistics/stylo [letzter Zugriff 20. März 2015].

Eder, Maciej / Rybicki, Jan (2011): "Deeper Delta across genres and languages: do we really need the most frequent words?", in: *Literary and Linguistic Computing* 26, 3: 315–321. 10.1093/llc/fqr031 .

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): "Towards a better understanding of Burrows's Delta in literary authorship attribution", in: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver 79–88. 10.5281/zenodo.18177 . http://www.aclweb.org/anthology/W/W15/W15-0709.pdf [letzter Zugriff 20. August 2015].

Guyon, Isabelle / Weston, Jason / Barnhill, Stephen / Vapnik, Vladimir (2002): "Gene Selection for Cancer Classification using Support Vector Machines", in: *Machine Learning* 46, 1: 389–422. 10.1023/A:1012487302797 .

Hasse, Dag Nikolaus / Büttner, Andreas (in Vorbereitung): "Notes on the Identity of the Latin Translator of Avicenna's Physics and on Further Anonymous Translations in Twelfth-Century Spain." Vorabversion: https://go.uniwue.de/hassevigoni [letzter Zugriff 17. Februar 2016].

Hirst, Graeme / Feiguina, Ol'ga (2007): "Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts", in: *Literary and Linguistic Computing Advance Access* 22: 1–13. 10.1093/llc/fqm023 .

Hoover, David L. (2004a): "Testing Burrows's Delta", in: *Literary and Linguistic Computing* 19, 4: 453–475. 10.1093/llc/19.4.453 .

Hoover, David L. (2004b): "Delta Prime?", in: *Literary and Linguistic Computing* 19, 4: 477–495. 10.1093/llc/19.4.477 .

Jannidis, Fotis / Lauer, Gerhard (2014): "Burrows's Delta and Its Use in German Literary History", in: Erlin, Matt / Tatlock, Lynne (eds.): *Distant Readings.* Topologies of German Culture in the Long Nineteenth Century. Rochester / New York: Camden House 29–54.

Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): "Improving Burrows' Delta - An Empirical Evaluation of Text Distance Measures", in: *Digital Humanities Conference 2015*, Sydney http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html [letzter Zugriff 26. Januar 2016].

Juola, Patrick (2006): "Authorship Attribution", in: *Foundations and Trends in Information Retrieval* 1, 3: 233–334.

Kestemont, Mike / Luyckx, Kim / Daelemans, Walter / Crombez, Thomas (2012): "Cross-Genre Authorship Verification Using Unmasking", in: *English Studies* 93, 3: 340–356. 10.1080/0013838X.2012.668793 .

Mosteller, Frederick / Wallace, David L. (1963): "Inference in an Authorship Problem", in: *Journal of the American Statistical Association* 58, 302: 275–309. 10.2307/2283270 .

Rybicki, Jan (2012): "The great mystery of the (almost) invisible translator: stylometry in translation", in: Oakley, Michael P. / Ji, Meng (eds.):

Quantitative Methods in Corpus-Based Translation Studies. Amsterdam: John Benjamins 231–248 https://sites.google.com/site/computationalstylistics/preprints/Rybicki%20Great%20Mystery.pdf [letzter Zugriff 26. Januar 2016].

Schöch, Christof (2013): "Fine-Tuning our Stylometric Tools: Investigating Authorship and Genre in French Classical Drama", in: *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska 383–386 http://dh2013.unl.edu/abstracts/ab-270.html [letzter Zugriff 26. Januar 2016].

Smith, Peter W. H. / Aldridge, W. (2011): "Improving Authorship Attribution: Optimizing Burrows' Delta Method*", in: *Journal of Quantitative Linguistics* 18, 1: 63–88. 10.1080/09296174.2011.533591 .

Stamatatos, Efstathios (2009): "A Survey of Modern Authorship Attribution Methods", in: *Journal of the Association for Information Science and Technology* 60, 3: 538–56. 10.1002/asi.v60:3 .

Stamatatos, Efstathios / Fakotakis, Nikos / Kokkinakis, George (2000): "Automatic Text Categorization in Terms of Genre and Author", in: *Computational Linguistics* 26, 4: 471–497. 10.1162/089120100750105920 .

TextGrid Konsortium (2006–2015): *TextGrid*. Virtuelle Forschungsumgebung für die Geisteswissenschaften. Göttingen: https://textgrid.de .

Viehhauser, Gabriel (2015): "Historische Stilometrie? Methodische Vorschläge für eine Annäherung textanalytischer Zugänge an die mediävistische Textualitätsdebatte", in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten der Digital Humanities* (Sonderband der Zeitschrift für digitale Geisteswissenschaften 1).

Mobile Anwendungen als multimodale Medien zur Vermittlung vormoderner Artefakte. Die ‚Historisches Paderborn'-App – ein interdisziplinäres Forschungs- und Lehrprojekt

Greulich, Markus
markus.greulich@uni-paderborn.de
Universität Paderborn, Deutschland

Oberthür, Simon
oberthuer@uni-paderborn.de
SICP – Software Innovation Campus Paderborn, Universität Paderborn, Deutschland

Karthaus, Nicola
karthaus@ieman.de
Universität Paderborn, Deutschland

Schmidt, Ariane
arianes@mail.uni-paderborn.de
Universität Paderborn, Deutschland

Wilk, Nicole M.
nicole.m.wilk@upb.de
Universität Paderborn, Deutschland

Stog, Kristina
kristina.stog@uni-paderborn.de
Universität Paderborn, Deutschland

Senft, Björn
bjoern.senft@uni-paderborn.de
SICP – Software Innovation Campus Paderborn, Universität Paderborn, Deutschland

Zusammenfassung der Sektion

Die Vermessung der Welt mittels digitaler Medien hat längst begonnen. Von der Durchdringung der Gesellschaft zeugen nicht nur Street View und weltweit verfügbare Satellitenaufnahmen, sondern auch Twitter und Facebook und nicht zuletzt die Auswirkungen auf die Wissenschaftskulturen, insbesondere auf die der Geisteswissenschaften. Hierfür hat sich der Begriff der Digital Humanities etabliert, der schillernd und komplex zugleich ist. Während zunächst historisches Material und Artefakte digitalisiert wurden, rückte in den letzten Jahren vor allem die Annotation von Digitalisaten und die Anlage und Aufbereitung von Datenbanken ins Zentrum des Interesses. Derzeit fächert sich das Spektrum der Digital Humanities weiter auf.

Anne Burdick, Johanna Drucker und andere (Burdick et al. 2012) weisen in ihrem intensiv rezipierten und vielfach zitierten Buch zum Konzept der Digital Humanities darauf hin, dass die Möglichkeiten und Chancen der Digital Humanities quasi einer Erweiterung der Geisteswissenschaften gleichkommen, die sowohl Werte, interpretative Praxis und Strategien der Bedeutung als auch die Ambiguitäten der menschlichen Existenz betreffen. Innerhalb der Sektion soll der Blick insbesondere auf zwei auch von Burdick und Drucker thematisierte Aspekte gelenkt werden: Zum einen ermöglicht die Erweiterung der Digital Humanities neue Wege der transmedialen Erforschung durch interdisziplinäre Kooperationen. Zum anderen darf nicht

nur die Anwendung von digitalen Werkzeugen und Datenbanken im Fokus stehen, sondern auch Konzeption, Entwicklung und Nutzung können und müssen neue Wege beschreiten.

Innerhalb dieser Sektion soll die kooperative, interdisziplinäre und interfakultäre Konzeption und Entwicklung einer mobilen Anwendung exemplarisch diese neuen Wege und Prinzipien als eine wichtige Option kulturwissenschaftlicher und informatischer Verschränkung vorgestellt werden.

An der Universität Paderborn hat sich vor zwei Jahren eine Forschergruppe innerhalb des akademischen Mittelbaus gebildet, die ein auf mehrere Semester angelegtes interdisziplinäres und interfakultäres Forschungs- und Lehrprojekt entwickelt hat. An diesem Projekt, das im September 2015 mit dem Forschungspreis der Universität Paderborn ausgezeichnet wurde, sind derzeit die Fächer germanistische Mediävistik und Linguistik, Geschichte, Informatik und Kunstgeschichte beteiligt. Die ‚Historisches Paderborn'-App (kurz HiP-App) zeigt die neuen Verknüpfungen, die durch das Konzept der Digital Humanities in den letzten Jahren proklamiert wurden, geradezu idealtypisch auf: Denn in der HiP-App ist die Informatik nicht lediglich Dienstleister für die Kulturwissenschaften und sind die Kulturwissenschaften nicht ausschließlich Content-Lieferanten für die Informatik. Vielmehr widmet sich das Projekt übergreifenden Forschungsfragen und ermöglicht darüber hinaus ebenso Individualforschung. In unserem Projekt fokussieren wir Fragestellungen, die sich aus der Konzeption, der Entwicklung und der Nutzung von digitalen Anwendungen (Apps) für mobile Endgeräte wie Smartphones oder Tablets ergeben.

Innerhalb dieser Sektion liegt der Fokus auf drei essenziellen Schwerpunkten des Projekts: (I.) auf dem Potenzial kooperativer Exploration und Konzeption mobiler Anwendungen für die Vermittlung wissenschaftlicher Inhalte in einem außeruniversitären Kontext, (II.) auf multimodaler Kommunikation und Raumwahrnehmung und (III.) auf der evolutiven Software-Entwicklung unter Berücksichtigung einer mensch-zentrierten Entwicklung, die in der interdisziplinären Kooperation zwischen Kulturwissenschaften und Informatik innerhalb unseres Projektes verwirklicht werden kann.

Ins Leben gerückt. Zum Potential mobiler Anwendungen für die Vermittlung vormoderner Artefakte

Markus Greulich, Nicola Karthaus und Ariane Schmidt (Universität Paderborn)

Insbesondere die historischen Geisteswissenschaften, im ganz besonderen Maße die mediävistischen Fächer, gelten manchem als längst überholte Wissenschaftsdisziplinen, aus denen weder ein Wert für die Gegenwart noch für die Zukunft zu erwarten ist. ‚Das' Mittelalter gilt als gut erforscht, die Einträge in online-Ressourcen vermitteln ein abgeschlossenes Bild: Wir wissen, wie unsere Vergangenheit war. Doch immer wieder gibt es Irritationen: Da geistern Pergamentfragmente durch die Tagesschau, widmen sich Regisseur_innen mit großem Erfolg den Lebensgeschichten von Kaiser_innen und Heiligen, zeigen internationale Serien, dass die Päpste des Mittelalters und der Renaissance ein nur begrenzt katholisches Leben pflegten und immer wieder geraten wertvolle Handschriften und Kunstobjekte des Mittelalters in den Fokus der Öffentlichkeit. So wie unsere Gegenwart nie als geschlossenes Bild vor uns stehen kann, so verändert sich auch ‚unser' Blick auf ‚das ' Mittelalter. Diesen Blick zu schärfen, vormoderne Artefakte lesbar zu machen und ihre eigene Geschichte als Teil ‚unserer' Geschichte, als Teil der Gegenwart erfahrbar zu machen – dies soll das interdisziplinäre Forschungs- und Lehrprojekt ‚Historisches Paderborn'-App, kurz *HiP*-App, leisten.

Die *HiP*-App ist eine Anwendung für mobile Endgeräte, die auf ansprechende Weise detaillierte und wissenschaftlich sinnvoll aufbereitete Materialien zur selbstständigen historischen Erkundung der Stadt Paderborn anbietet. Sie wird derzeit von Mittelbau-Vertreter_innen der Universität Paderborn aus den Bereichen Informatik, germanistische Mediävistik und Linguistik, Geschichte und Kunstgeschichte entwickelt. Die *HiP*-App ist zugleich Forschungsgegenstand und -infrastruktur, die durch ihren grundlegend evolutiven Charakter vielfältige Forschungsfragen erzeugt (Burdick et al. 2012). Existierende Denkmäler dienen als Ausgangspunkt, um germanistische, historische und kunsthistorische Inhalte zu erläutern. Das interaktive Front-End der *HiP*-App, d. h. die für den Benutzer sichtbare Oberfläche, wird u. a. durch aktuellste Präsentationsformen historischer Artefakte im Bereich der Augmented Reality, d. h. der virtuell erweiterten Realität, gestaltet. Hierbei können Kunstwerke und andere Objekte nicht nur schriftlich oder mündlich erläutert, sondern auch singuläre Details hervorgehoben oder verlorene Sinnzusammenhänge visualisiert werden. Sogar der ursprüngliche Kontext eines Werkes kann so rekonstruiert werden. Auch ist es möglich, kunsthistorische Vergleiche zu ziehen, verwandte Werke zu zeigen, zusätzliche Materialien anzubieten und eine kulturhistorische Kontextualisierung vorzunehmen. Die neuen technischen Möglichkeiten der medialen Aufbereitung von Kulturgeschichte sind verbunden mit neuen sozialen Praktiken des Wahrnehmens an der Schnittstelle zwischen physischem und digitalem Raum (Buschauer / Willis 2013; Schüttpelz 2006). Die so gestalteten historischen Aussagen werden im Hinblick auf eine für die Gegenwartsorientierung relevante (stadt)geschichtliche Sinnbildung (Rüsen 1996) motiviert. Somit soll die Neugier der Nutzer_innen geweckt und

durch eine veränderte Wahrnehmung des urbanen Umfelds eine weiterführende A useinandersetzung mit dem Paderborner Kulturraum gefördert werden.

Zentrales gemeinsames Forschungsanliegen ist es, am Beispiel der HiP-App Methoden, Prozesse und Analysen im Bereich der Digital Humanities zu entwickeln. Forschungsgegenstand und Grundlage hierfür bilden Genese, Entwicklung, Betrieb und Pflege der *HiP*-App, aber auch die kritische Reflexion der Kopplung physischer und digitaler Räume mithilfe mobiler Apps, die von der menschlichen Körperorientierung ausgehend den städtischen Raum diachron lesbar machen. So sollen durch Analyse der Dateneingabe etwa auch Aspekte der Software-Usability und der Entwicklung multimodaler Kommunikationsformate in den Blick genommen werden. Bezogen auf die Front-End-Entwicklung stehen verschiedene Formen des epistemologischen Präsentmachens durch Visualisierung, auditive Aufbereitung und weitere empirische Anknüpfung durch materielle Spurensuche auf dem Prüfstand (Kesselheim 2010). Die interdisziplinäre Kooperation lässt gleichzeitig Methoden der verschiedenen Disziplinen produktiv zusammenwirken. Entwicklung und Betrieb von Front- und Back-End finden aktuell in Form agiler Softwareentwicklung statt, in die die beteiligten Akteure aus den Kulturwissenschaften (insbesondere auch die Studierenden) fest eingebunden sind. Weitere Schwerpunkte liegen auf der leichten Nutzbarkeit von Technologien, beispielsweise einer einfachen Pflege der eingestellten Daten innerhalb eines Web-Back-Ends sowie einer flankierenden kulturwissenschaftlichen Reflexion der raumgenerierenden Potenziale neuer mobiler Systeme. Die Motivation des Projekts liegt somit auch darin, die aus der Informatik heraus entwickelten neuen Formen der Überblendung und Verkopplung physischer, kartographierter und medialer Räume in ihren produktiven Momenten der Verräumlichung (Habscheid / Reuther 2013) und der historischen Narrativierung kulturwissenschaftlich und experimentell (Back-End-Editieren, Nutzerverhalten) zu begleiten.

Im Rahmen der ersten Projektphase werden derzeit drei historische Stadtrundgänge zum Heiligen Liborius, zu Kaiser Karl dem Großen und zu Bischof Meinwerk von Paderborn entwickelt. Diese historischen Persönlichkeiten sind für die Genese und Entwicklung der Stadt Paderborn im Frühmittelalter von besonderer Bedeutung gewesen. Sie sind auch heute noch in vielfältiger Weise im Stadtbild präsent. Unter anderem wird dies in einem weiteren Rundgang thematisiert, der sich historischen Orts- und Straßennamen widmet, denn auch in der Namensgebung von Orten und Straßen artikuliert sich das kulturelle Gedächtnis. Ein zentrales Anliegen unseres Projekts ist es, den Stadtraum historisch erfahrbar zu machen. Ziel ist es, unter anderem die Ungleichzeitigkeit des Gleichzeitigen herauszuarbeiten: Die in der heutigen Präsenz als in sich geschlossene Einheit sichtbaren Objekte sollen als historisch gewachsen erfahren werden, ihre einzelnen Elemente als aus unterschiedlichen Epochen stammend.

Ein gutes Beispiel hierfür ist der Paderborner Dom (einführend: Quednau 2011), der sowohl von vielen Ortsansässigen als auch von Touristen mit seinen zahlreichen Artefakten und religiösen Objekten als gegebenes Bauwerk wahrgenommen wird. Dass dieses Bauwerk aber keinesfalls statisch, sondern vielmehr historisch gewachsen ist (u. a. Lobbedey 1986) und seit der Grundsteinlegung unter Karl dem Großen im Jahre 777 im Laufe der Jahrhunderte vielfältige bauliche Veränderungen, Erweiterungen und Überformungen erfahren hat, ist nur wenigen bewusst. Hier setzt die *HiP*-App an. Mit ihrer Hilfe ist es möglich, den Dom in seiner ganzen historischen Dimension und Vielschichtigkeit für den Betrachter sichtbar und erfahrbar zu machen. Als ein ganz konkretes Beispiel für die Ungleichzeitigkeit des Gleichzeitigkeit bietet sich das Paradiesportal des Paderborner Doms an:

Ursprünglich als romanische Vorhalle konzipiert, wurde das Paradies im Zuge eines Umbaus des Westquerhauses mit einem Figurenportal ausgestattet. Dieses wies zunächst die Struktur eines rein ornamentalen Portals auf, dem in der Konzeptionsphase Sandsteinfiguren hinzugefügt wurden. Die monumentalen gotischen Gewändefiguren entsprechen dem zeitgenössischen Geschmack des 13. Jahrhunderts und orientierten sich an Skulpturen der französischen Gotik, beispielsweise denen der Kathedralen von Paris und Reims (Sauerländer 1971). Auf diese Weise 'modernisierte' also die Paderborner Dombauhütte das Hauptportal des Sakralbaus bereits im Hochmittelalter, und zwar in Anlehnung an die damals aktuelle und hochwertige Bauskulptur des französischen Königreichs. Die beiden romanischen Holzskulpturen des Portals, die den hl. Kilian und den hl. Liborius darstellen, stammen hingegen aus dem 12. Jahrhundert, wurden jedoch erst deutlich später, nämlich 1815, an den Portaltüre n angebracht. Diese Veränderung und Ausgestaltung des Domportals eignet sich in geradezu idealer Weise, um den Nutzer_innen mit Hilfe von Augmented Reality innerhalb des Front-Ends die historische Dimension nicht nur des Portals, sondern exemplarisch auch des gesamten Doms vor Augen zu führen.

Für interessierte Nutzer_innen hält die *HiP*-App aber auch vertiefende Informationsangebote bereit: So z. B. erläuternde Texte zu einzelnen Skulpturen des Portals, die etwa typische, in der App farbig hervorgehobene Attribute der Heiligen erklären, oder aber kulturhistorische Einordnungen mit Blick auf die Nutzung (Tack 1958) oder die religiöse Praxis (Bawden 2014). Auch können Brücken und Verbindungen zwischen verschiedenen Fachdisziplinen geschlagen werden: So illustriert die zentrale Marienfigur des Trumeaus die hochmittelalterliche Marienverehrung anhand des Melker Marienlieds, eines wenig bekannten Textes aus dem 12. Jahrhundert. Die Nutzer_innen erfahren hier nicht nur Wissenswertes zum Text, sondern sehen mit Hilfe der App auch - vielleicht erstmals - eine mittelalterliche Handschrift. Dass Maria im Melker Marienlied als

Himmelspforte (porte des paradyses) bezeichnet wird, schlägt dabei eine Brücke zur Portalsymbolik und gibt - über die Einzelbetrac htung des Portals hinaus - einen Einblick in die Kulturgeschichte und spezieller noch in die Liturgie. War die Nutzung des Portals - u. a. als erzbischöflicher Zugang und als Gerichtsort - bereits zur Bauzeit vielschichtig, so erfuhr sie im Laufe der nachfolgenden Jahrhunderte noch weitere Veränderungen und Ergänzungen. So bildet das Domportal etwa noch heute den monumentalen Rahmen für die feierliche Liborius-Prozession: Sie erinnert alljährlich an die Translation der Gebeine des Heiligen aus dem damals westfränkischen, heute französischen Le Mans und an ihre Ankunft in Paderborn im Jahre 836. Darüber hinaus werden in der App regionale und überregionale Vergleichsobjekte (Sauerländer 1971; Lobbedey 1999) vorgestellt, die eine kunsthistorische Einordnung in die Portalskulptur geben. Damit macht die App eben gerade nicht nur die konkrete Stadt- und Baugeschichte Paderborns anschaulich erlebbar, sondern greift darüber hinaus. Es geht gerade auch darum, den Nutzer_innen der stadtgeschichtlichen App ein exemplarisches Wissen zu vermitteln, durch das Sehgewohnheiten verändert und historische Artefakte selbständig les- und erfahrbar werden.

Digitalisierung von geschichtlichem Wissen im Raum und raumgebundener Erinnerungskultur – am Beispiel von Straßennamen

Kristina Stog (Paderborn)/ Nicole M. Wilk (Paderborn)

Idee: Vervielfältigung der Lesarten durch neue Visualisierungsmethoden

Wissensarten sind an Darstellungsformate gebunden. Informationen über das Medienmaterial, seine Gestaltung, Beschaffenheit und Platzierung gehen bei der Verarbeitung von Daten im Zuge der Digitalisierung größtenteils verloren oder werden isoliert vom Textkorpus z. B. in Form von Metadaten gespeichert. Dieser Verlust wird in den interpretierenden Disziplinen oft in Kauf genommen, da quantitativ motivierte Fragestellungen bereits an diese Reduktionssituation angepasst sind. Doch es setzt sich gleichzeitig in linguistischen und (sozial)semiotischen Forschungskontexten die Erkenntnis durch, dass Wissen immer situiertes Wissen in materiellen und institutionellen Umgebungen ist (Fix 2008), und dass mit Blick auf multimodale Gebrauchsmuster die Wahl der semiotischen Ressource (O'Halloran 2004) und nicht zuletzt die Raumbasiertheit von Kommunikation semantische und Diskurs strukturierende Effekte haben können (Habscheid / Reuther 2013).

Kaum ein „Text“ ist so eng mit seinem Ort verknüpft wie ein aufgedruckter „Name“. Am Beispiel der Namen für Straßen, Gebäude und Plätze zeigen wir in unserem Beitrag Digitalisierungsmöglichkeiten auf, die die Verknüpftheit von stadtgeschichtlichem Wissen mit Orten auf (technisch) verschiedene Weise modellieren. Hierfür wird mit dem Smartphone ein mobiles Instrument gewählt, das über eine App Schnittstellen zwischen materiellem und digitalem Raum erzeugt (Weber 2012), um Stadtgeschichte in unterschiedlichen Deutungsrahmen (visuell, auditiv) verfügbar zu machen. Der Raum erweist sich dabei als interaktive Ressource (Hausendorf / Mondada / Schmitt 2012), auf die die kulturellen Sinnangebote ausgerichtet sind und die sie selbst als solche hervorbringen (reflexiver Raumbegriff).

Mit den neuen technologischen Verfahren erschöpft sich die interaktive Dimension nicht in der Aufmerksamkeitslenkung durch schriftbasierte oder bildliche Information, mithilfe standortbezogener Informationen durch Location-based Services kann Nutzer_innen vielmehr an konkreten Orten durch eine mittelalterliche Geräuschkulisse oder eine visuelle Anreicherung des Stadtbilds ein Einstieg in historische Szenarien geboten werden.

Hintergrund: Straßennamen als Kondensate kulturellen Wissens und der Inanspruchnahme von Geschichte

Namen von Straßen, Plätzen und anderen Örtlichkeiten (Toponyme) dienen der Orientierung. Sie erschließen den Raum und strukturieren ihn physisch und historisch, zugleich können Toponyme als Verweise auf die Geschichte sowie auf das Geschichtsbewusstsein einer Stadt gelesen werden. Anders als Denkmäler, die aufgrund ihrer erinnernden Funktion bewusst aufgesucht werden, stellen sie – in ihrer Sekundärfunktion – „Medien kultureller Erinnerung“ (Pöppinghege 2005: 10) dar, die von den Rezipient_innen im urbanen Raum täglich genutzt werden. Vor allem die frühen Orts- und Straßennamen, die aufgrund von gemeinsamen Gewohnheiten, Bedürfnissen oder Wahrnehmungen in der Interaktion (bereits im Mittelalter) gewachsen sind (Fuchshuber-Weiß 1996) geben Hinweise auf Vergangenes, im heutigen Stadtbild möglicherweise nicht mehr Sichtbares: Geographische bzw. topographische Merkmale oder Besonderheiten des Ortes, nennenswerte Gebäude in der Umgebung, eine bestimmte Nutzung des Bezugsbereiches (wie etwa dort angesiedeltes Gewerbe) oder soziokulturelle Tatbestände vor Ort (Fuchshuber-Weiß 1996).

Namen stellen dabei keinen schlichten Spiegel tatsächlicher Gegebenheiten dar, sondern geben Einblick in die kollektiven Sicht- und Vorstellungsweisen ihrer Nutzer und bilden „in ihrer auswählenden und akzentuierenden Thematisierung des Stadtraumes Dokumente einer Mentalitätsgeschichte des Sehens“ (Glasner 1999: 320). In dieser Hinsicht gibt auch die heutige Benennungspraxis Aufschluss über das Geschichtsbewusstsein einer Stadt: So spiegelt sich etwa in der Vergabe von Namen, die sich auf Historisches „vor Ort“ beziehen, auch das „kultur- und alltagsgeschichtliche Verständnis“ (Pöppinghege 2005: 10) einer Stadt. Lokalen Ereignissen, Personen, aber auch Intentionen oder Vorstellungen, mit denen sich eine Stadt identifiziert, werden in Form von Straßen(-namen) begehbare „Zeichen gesetzt“.

In der Erarbeitung toponymischen Wissens ergibt sich eine Herausforderung daraus, dass verschiedene Wissenssorten zusammenkommen: Legendenbildungen, Volksetymologien, Geschichtswissen der Historiker und ein zeitabhängiges Geschichtsbewusstsein, das teilweise mit einer intensiven Geschichte der Umbenennung einhergeht. Beobachtungen zur thematischen Verarbeitung von Namensgebung und Namensgeschichte in bestehenden Apps zur Stadtgeschichte belegen das allgemeine Bedürfnis nach „Lesbarkeit“ von urbanen Räumen: Namen werden als Spuren geschichtlicher Zusammenhänge aufgeschlossen. Doch wie werden durch sie Diskurse räumlich materialisiert? Wie können Wissenssorten auch durch mediale Operationen reflektiert werden? Das Beispiel der in Entwicklung befindlichen 'Historisches Paderborn'-App soll aufzeigen, wie die doppelte Situiertheit des stadtgeschichtlichen Wissens in einem metakommunikativen und in einem räumlichen Sinne durch Visualisierungstechniken darg estellt werden kann.

Ausgangslage: Typonyme in bisherigen Kommunikationsangeboten und Apps zur Stadtgeschichte

Um in mobilen Geräten Geschichtliches auf neue experimentelle Weise darzustellen, müssen zunächst die traditionellen Repräsentationen geschichtlichen Wissens im Stadtraum hinsichtlich ihrer raumstiftenden Qualitäten erschlossen werden. Straßen- und Gebäudenamen haben trotz des in ihnen sedimentierten impliziten Wissens über Ereignisse der Stadtgeschichte für viele Bewohner und Stadtbesucher einen primär pragmatischen Sinn und dienen der räumlichen Orientierung. Selbst Legenden und Volksetymologien, die sich um die Namen im Stadtraum ranken, drohen verloren zu gehen. Auf diese Situation reagieren ortsfeste und ortsgebundene digitale Kommunikationsangebote zur Stadtgeschichte, die den urbanen Raum als Medium des kollektiven Gedächtnisses und mit ihm eine völlig neue städtische Erzählkultur etablieren. In einer medienlinguistischen Studie zur Musterhaftigkeit ortsfester Kommunikationsangebote zur Stadtgeschichte konnten zwei wesentliche Tendenzen in der Entwicklung internetbasierter Formate und Stadtgeschichts-Apps festgestellt werden: der Ausbau einer dialogischen Sequenzierung stadtgeschichtlichen Wissens und die Narrativisierung urbaner Erzählsequenzen („Histörchen“). Ohne Angabe von Quellen und ohne Verweise auf die Deutungsvielfalt der historischen Dokumente werden Brauchtümer und Motive sprachlich so dargelegt als seien die Namen Repräsentationen einer zu allen Zeiten und eindeutig herauszulesenden historischen Faktizität (vgl. Wilk 2015). Der Einsatz multimodaler Darstellungsformen verspricht hier Möglichkeiten, Namen und Namenswandel exemplarisch in einem Spurkonzept zu modellieren (vgl. Müller 2012), das Typonyme im linguistischen Sinn weniger als objektive Zeugen eines geschichtlichen Geschehens aufschließt als vielmehr anhand der Namensgebung den Kampf um historische Lesarten und ihre Orientierung für die Zukunft verdeutlichen. Aufgabe der Medienlinguistik ist es dabei, anhand konkreter Textentwicklunge n zu beschreiben, wie unter Nutzung verschiedener Daten aus den Visualisierungen eines durch historische Szenarien erweiterten Stadtraums unterschiedliche historische Interpretationen hervorgehen.

Das Beispiel Paderborn – Motiviertheit und sozialer Sinn hinter den Spuren

Wie Namen als Spuren von Vergangenem im heutigen Stadtbild gelesen werden können, lässt sich anhand einiger Straßennamen in der Paderborner Innenstadt beispielhaft zeigen: Sie können Hinweise auf das historische Stadtbild geben, wie etwa der Name Grube, der als einer der ältesten Straßennamen in der Altstadt auf die heute nicht mehr sichtbare Grube, die nach Auffüllung eines Steinbruchs südlich der Domburg im 12. Jahrhundert zu sehen war (vgl. Liedtke 1999: 101), verweist. Namen wie Im Düstern oder Krummer Ellenbogen geben darüber hinaus Einblicke in die Wahrnehmung des städtischen Raums aus der Perspektive ihrer Nutzer_innen. Neben Anwohnergruppen (Weberberg) spiegeln sich in Namen bestimmte Nutzungsweisen von Straßen (wie etwa Kühe durch die Kuhgasse zur Tränke an die Pader zu treiben (vgl. Liedtke 1999)). Auch Spuren des Niederdeutschen, das als gesprochene Alltagssprache in Paderborn kaum noch existiert, finden sich in Namen wie Abtsbrede (Brede bezeichnet einen breiten Acker) oder Börnepader (börnen: tränken).

Nach Nübling (2012: 244) lassen sich aus diesen primären Straßennamen, die in engem Zusammenhang mit den Straßen, die sie bezeichnen, entstanden sind, „Topographie und Sozialgeschichte einer Stadt hervorragend rekonstruieren“. In einer App zum Historischen Paderborn bilden sie nicht nur einen Anknüpfungspunkt für die Auseinandersetzung mit dem historischen Raum, sondern auch mit den kollektiven Sicht- und Vorstellungsweisen ihrer Nutzer_innen. Dies gilt auch für die sekundären Straßennamen, die administrativ vergeben werden: Die Wahl der regionalen und überregionalen Personen und Ereignisse, nach denen eine Stadt ihre Straßen benennt, gibt Aufschluss über ihr (Stadt-)Geschichtsbewusstsein (vgl. Pöppinghege 2007). So deutet etwa die Benennung der Straßen eines Viertels in Paderborn, in dem neben der Karlsstraße und dem Karlsplatz auch Albin-, Gerold-, Einhard- und Widukindstraße auf historische Zusammenhänge und Personen im Umfeld Karls des Großen verweisen, auf die Bedeutung hin, die dieser im Selbstverständnis der Stadt einnimmt.

Politische und gesellschaftliche Umbrüche werden vor allem in der Auseinandersetzung mit den Umbenennungen von Straßen oder Plätzen sichtbar. In ihnen spiegeln sich die Vorstellungen, Ideen und Ideologien, zu deren Verbreitung Toponyme seit dem 18. Jahrhundert genutzt werden (vgl. Fuchshuber-Weiß 1994: 1472). Deutlich wird dies am Beispiel des Le-Mans-Walls in Paderborn, der bis 1938 als Wilhelmstraße, in der Zeit des Nationalsozialismus als Horst-Wessel-Wall und nach 1945 erneut als Wilhelmstraße bezeichnet wurde, bis 1967 mit der Städtepartnerschaft die Umbenennung nach der französischen Stadt Le Mans folgte (vgl. Liedtke 1999: 151). Mit der Benennung, die sich nun zugleich an mittelalterlichen Ereignissen orientierte, wurde der Straße – als Weg, über den 836 der Zug mit den Reliquien des Hl. Liborius von Le Mans in Richtung Dom geführt haben soll – somit auch eine größere Bedeutung innerhalb der Stadtgeschichte Paderborns zugesprochen.

In der *Hip*-App lässt sich die Geschichte dieser Umbenennung und der Vereinnahmung historischer Persönlichkeiten für die städtische Identität multiperspektivisch visualisieren, so dass anschaulich wird, wie zu verschiedenen Zeitpunkten Historisches in städtischen Strukturen repräsentiert (worden) ist. Diese Repräsentationen der historischen Traditionen schließen zudem die Konsequenzen für das (Geschichts-)Bild der gegenwärtigen Stadt und der Stadt der Zukunft auf.

Hierbei sollen epistemischen Effekte, d.h. insbesondere komplexitätsreduzierende Wissenseffekte verschiedener Darstellungsweisen (Karten, archäologische Modelle, Einblendungen) exemplarisch erfasst werden. Diese variieren mit der Auswahl der Kategorien, der Relationierung markanter Ereignisse und nicht zuletzt der Herstellung von Bezügen zum materiellen Raum. Die Differenzierung von Erzählzeit und erzählter Zeit reflektiert dabei die Variabilität historischer Sinnstiftung: So können historische Figuren (Könige, Ritter, Pilger etc.), die z. B. mit Pferdegetrappel an ausgewählten Stellen des Stadtrundgangs die Wege der Nutzer_innen kreuzen und damit historische Situationen auf dem Hellweg simulieren, eine mittelalterliche Vergangenheit einerseits behaupten. Andererseits lassen sie sich anschließen an eine stadttypische Rezeptionsgeschichte mittelalterlicher Quellen und Schriften. In der (zusätzlichen) Darstellung gewandelter städtischer Identitätsdiskurse z. B. über kartografierte Wissensbezüge (Bezug zum Mittelalter, zur niederdeutschen Varietät, zur Geografie etc.) lässt sich ein jeweils zeitabhängiges Geschichtsbewusstsein veranschaulichen.

Auf dem Weg zu einer experimentellen und evidenzbasierten Softwareentwicklung in den Digital Humanities

Björn Senft, Simon Oberthür – SICP – Software Innovation Campus Paderborn, Universität Paderborn

Das interdisziplinäre Projekt ‚Historisches Paderborn'-App, kurz *HiP*-App, ist ein gutes Beispiel für die sich verändernden Anforderungen an den Software-Entwicklungsprozess im DH-Kontext bzw. im Kontext des digitalen Wandels (Digital Transformation). Klassische Entwicklungsmodelle wie das Wasserfallmodell wurden für Situationen entworfen, in denen Funktionsumfang und Aufbau der Software zu Beginn der Entwicklung relativ genau festgelegt werden können.

Für die *HiP*-App ist dies jedoch aus mehreren Gründen nicht möglich. Aufgrund der Verwendung neuer Technologien in der App und der Raum- und Zeitgebundenheit der Inhalte ergeben sich vollkommen neue Wege, mit Wissen umzugehen. Das Dilemma ist nicht gerade selten: Informatiker verfügen über das Wissen um moderne Technologien, Kulturwissenschaftler verfügen über das Wissen der zu vermittelnden Inhalte. Ein sinnvoller und innovativer Einsatz neuer Technologien kann aber nur in enger Verzahnung mit konkreten Inhalten erfolgen. Oft kann auch eine Verzahnung vorab nicht hinreichend beurteilt werden, sondern muss beispielsweise experimentell bewertet werden. Die sinnvolle Anwendung neuer Technologien ist deshalb ein Forschungsdesiderat und muss durch geeignete Entwicklungsmethoden und -abläufe unterstützt werden.

Weil in unserem Kontext die Anforderungen vor der Implementierung nicht genau spezifiziert werden können, muss erforscht werden, welche Faktoren zielführend für die Lösung der zu bewältigenden Entwicklungsaufgabe sind. Da die Informatik mit dieser Situation relativ häufig konfrontiert ist, wurden hierfür

Methoden wie etwa die agile Entwicklung gebildet. Das ist allerdings nur ein möglicher Ansatz zur Lösung dieses Problems, da die Methoden nur einen lockeren Rahmen bieten und nicht genauer darauf eingehen, wie mit der Software experimentiert werden kann, geschweige denn, wie eine systematische Extraktion und Evaluation in diesem Kontext aussehen könnte. Sinnvoll wäre beispielsweise ein Softwareleitstand, der Experimente mit verschiedenen Nutzergruppen (Betatester, Endnutzer, Experten, etc.) bzw. Ebenen (Simulation, Menschen, etc.) ermöglicht. Erschwerend kommt dabei hinzu, dass Prozesse verschiedener Domänen (Kulturwissenschaften, kulturelle Institutionen, etc.) in den eigentlichen Softwareentwicklungsprozess integriert werden müssen.

Lösungsansatz

In unserem Projekt entwickeln wir ein mensch-zentriertes Prozessmodell (siehe Abbildung 1), dass das DevOps-Prinzip (Hüttermann 2012; Sharma / Coyne 2015) mit dem Führungskreislauf des St. Galler Management Modells (Ulrich / Krieg 1974) kombiniert, um so die vielschichtige Verzahnung von Technologien, Inhalten, Domänen und Akteuren (Informatiker, Kulturwissenschaftler, Usability-Experten, Nutzer, etc.) zu gewährleisten. Die Grundidee hinter diesem Modell ist die Aufteilung in generellere Phasen, um so Verzahnungspunkte für die unterschiedlichen Prozesse zu definieren. In der Ermittlungsphase werden mit Methoden der Informatik und der Kulturwissenschaft Daten aus der Realität (Verwendung des Prototyps, Interviews, Unit-Tests, ‚ausgelieferte' Software, etc.) ermittelt, um so ein Lagebild zu erstellen, das die Grundlage für die Weiterarbeit in der Strategie- und Analysephase bildet. In dieser kommen die unterschiedlichen Akteure zus ammen und analysieren gemeinsam die ermittelten Daten und beraten über die zukünftige Strategie, die in der Realisierungsphase (Implementierung, Entwickeln eines konkreten Interviews, etc.) umgesetzt und anschließend erneut in der Realität eingesetzt und evaluiert wird. Wichtig ist dabei - da es kein definiertes Ende gibt -, die einzelnen Phasen kontinuierlich und iterativ zu durchlaufen, da wir nach Drucker et al. (2012) davon ausgehen, dass sich die Software, aufgrund der hieraus entstehenden neuen Erkenntnisse, ständig weiterentwickeln wird:

„Digital Humanities work embraces the iterative, in which experiments are run over time and become objects open to constant revision. Critical design discourse is moving away from a strict problem-solving approach that seeks to find a final answer: Each new design opens up new problems and—productively—creates new questions." (Drucker et al. 2012: 22).

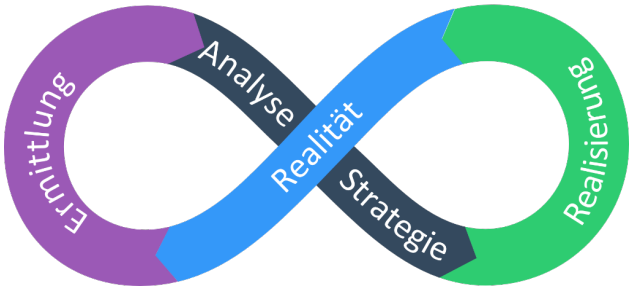


Abb. 1: Verwendetes Prozessmodell der *HiP*-App-Entwicklung im Digital Humanities Kontext

Erfahrungen und Erkenntnisse

Der bisher entwickelte Teil des Lösungsansatzes basiert auf den Erfahrungen, die bislang bei der Entwicklung der *HiP*-App gemacht wurden. Eine studentische Projektgruppe der Informatik entwickelte in stetiger Rückkopplung mit den Kulturwissenschaften das Backend zum Einpflegen der Daten. Die Studierenden entwickeln die Software nach Ansätzen der agilen Softwareentwicklung (Beck et al. 2001) und nach dem DevOps-Prinzip (Hüttermann 2012; Sharma / Coyne 2015), die für uns Schlüsselfaktoren sind, um einen kontinuierlichen mensch-zentrierten Softwareentwicklungsprozess (Mayhew / Follansbee 2012) mit explorativen Möglichkeiten zu erreichen.

Dass eine enge Kooperation von Informatik und Kulturwissenschaften im Sinne der Digital Humanities notwendig ist, hat sich bereits in den ersten Arbeitsphasen des Projekts bestätigt. Wie bereits erwähnt, können sinnvolle Anwendungen nur in enger Verzahnung von Inhalten und Technologien entstehen. Informatik und Kulturwissenschaften müssen deshalb interagierend Daten auswerten und die Strategie anpassen. Diese Erkenntnis ist eine Quintessenz aus unserer Projekterfahrung.

Den involvierten Kulturwissenschaftlern fehlten anfangs Bewusstsein und Wissen über die notwendige Spezifizität, über den Realisierungsaufwand und auch über die Nachhaltigkeit der Softwareentwicklung, die für eine zielgerichtete Entwicklung qualitativ hochwertiger Software jedoch essenziell sind. Die beteiligten Informatiker verloren sich dagegen allzu schnell in technologischen Herausforderungen anstatt die Nutzeranforderungen zu fokussieren. Es bedarf deshalb eines gemeinsamen Verständnisses und einer gemeinsamen Strategie, welche Merkmale einer Anwendung welche Priorität haben und wann diese implementiert werden sollen oder aber mit Hilfe anderer Frameworks zu realisieren sind. Um wichtige technologische Entscheidungen treffen zu können, müssen sich die Anforderungen herauskristallisieren, die sich im Detail aus den Prozessen und Inhalten ergeben. Daher erscheint uns ein Verhältnis Dienstleister (Informatik) und Content-Lieferant (Kulturwissenschaften) für die

Entwicklung von Software im Kontext von Digital Humanities wenig sinnvoll, ja sogar kontraproduktiv.

Festzuhalten ist, dass unser Prozessmodell möglichst kurze Durchläufe erlaubt, um so frühzeitig neue Erkenntnisse zu gewinnen, die dann zeitnah in die weitere Softwareentwicklung einfließen können. Die kurzen Wiederholungen im Prozessmodell helfen, nicht-verbalisierbares Nutzerwissen verfügbar zu machen. Solches Wissen ist beispielsweise grundlegend, um die Lösungen bestmöglich auf die Bedürfnisse der Nutzer auszurichten.

Unser Lösungsansatz ermöglicht ein Experimentieren, das nicht nur auf Softwareprototypen bezogen ist. Beim gemeinsamen „Design Thinking“ (Uebernickel et al. 2015) von Informatikern und Kulturwissenschaftlern haben wir die Erfahrung gemacht, dass erst und vor allem das häufige Evaluieren und Experimentieren dabei hilft, sich von technologischen und organisatorischen Restriktionen zu lösen und stattdessen sinnvolle Anwendungen zu identifizieren. Es müssen zudem technologische Konzepte entwickelt werden, die sowohl ein Experimentieren mit verschiedenen Varianten als auch eine Evolution von Software-Architektur und -Design und ein Reagieren auf fehlerhaften Code (Resilience) erlauben.

Abgrenzung vom aktuellen Stand der Forschung und Technik

Bisherige Ansätze in der Softwaretechnik sind vor allem mit Blick auf das Extrahieren und Experimentieren unzureichend. Ein klassisches Entwicklungsmodell in der Informatik ist das Wasserfallmodell (Royce 1970), in dem bestimmte Phasen wie Anforderungserhebung, Systementwurf und Implementierung lediglich einmal durchlaufen werden. Dieses Modell ist vor allem für solche Entwicklungen geeignet, die bereits existente Prozesse digitalisieren sollen. Wenn jedoch neue digitale Prozesse entwickelt werden sollen, wirkt sich bei diesem Modell nachteilig aus, dass Fehlentwicklungen erst am Ende des Prozesses sichtbar werden, also erst dann, wenn die Software als Ganzes bereits fertig ist. Um schneller auf sich ändernde Anforderungen reagieren zu können, wurden deshalb agile Methoden entwickelt, deren Rahmen mithilfe des agilen Manifest (Beck et al. 2001) definiert werden. Zur Grundidee der agilen Softwareentwicklung gehören kurze, feste Iterationen mit de m Ziel, möglichst frühzeitig lauffähige Produktinkremente auszuliefern. So können öfter Rückmeldungen vom Kunden eingeholt und Fehlentwicklungen frühzeitig erkannt werden. Vor allem aber ist dieser Prozess auch transparenter für den Kunden, da er regelmäßig Fortschritte sieht. Da dieser Ansatz davon ausgeht, dass man zwangsläufig ‚scheitern‘ wird, soll das Scheitern im Kleinen stattfinden, um so Potenzierungseffekte zu minimieren. Die agilen Methoden haben jedoch den Nachteil, dass sie lediglich einen Rahmen bilden und keine Spezifika bieten, wie z. B. konkret experimentiert werden kann oder soll. Als Anforderungen werden im agilen Ansatz Scrum User Stories verwendet, die aus Nutzersicht die gewünschten Funktionalitäten beschreiben. Um eine gute Produktqualität zu erreichen, muss Scrum (Sutherland / Schwaber 2007) mit klassischen Ansätzen kombiniert werden.

Im Gegensatz zu den hier erläuterten Ansätzen bietet das von uns vorgestellte Modell durch den ständig wiederkehrenden Dialog der domänenübergreifenden Akteure sowie das Experimentieren die Möglichkeit, die Entwicklung neuer Methoden und Werkzeuge systematisch zu unterstützen. Gestützt von Prozessen können so neue Technologien experimentell auf ihre Anwendbarkeit untersucht werden. Ein Aspekt, der sich aktuell in der Projektarbeit der *HiP*-App bereits bestätigt hat.

Übergeordnete offene Fragestellungen

In ihren Anfängen zeichneten sich die Digital Humanities hauptsächlich durch die Übertragung bewährter Konzepte der Informatik aus. Es ist aber zu fragen, ob die Informatik nicht stärker von den Kulturwissenschaften lernen kann? Wäre es für die Informatik beispielsweise nicht hilfreich, verstärkt auch soziologische Methoden (qualitative Methoden wie Experteninterviews, quantitative Verfahren, etc.) für den Prozess der Anforderungserhebung und des Experimentierens zu adaptieren? Wie aber könnte das in der Softwareentwicklung praktikabel und systematisch angewandt werden? Diese offenen Fragestellungen gilt es weiterhin im Auge zu behalten.

Bibliographie

Bawden, Tina (2014): *Die Schwelle im Mittelalter* (= Sensus 4). Köln / Weimar / Wien: Böhlau-Verlag.

Beck, Kent et al. (2001): *Manifesto for agile software development* http://agilemanifesto.org [letzter Zugriff 20. Januar 2016].

Burdick, Anne / Drucker, Johanna / Lunenfeld, Peter / Presner, Todd / Schnapp, Jeffrey (2012): *Digital_Humanities*. Cambridge: The MIT Press / Massachusetts Institute of Technology.

Buschauer, Regine / Willis, Katharine S. (2013): *Locative Media*. Medialität und Räumlichkeit. Multidisziplinäre Perspektiven zur Verortung der Medien. Bielefeld: Transcript.

Fix, Ulla (2008): "Nichtsprachliches als Textfaktor. Medialität, Materialität, Lokalität", in: *Zeitschrift für Germanistische Linguistik* 36, 3: 343–354.

Fuchshuber-Weiß, Elisabeth (1996): "Straßennamen: deutsch / Street Names: German / Noms de rues: domaine allemand", in: Eichler, Ernst / Hilty, Gerold / Löffler, Heinrich / Steger, Hugo / Zgusta, Ladislav (eds.): *Namenforschung*. Ein internationales Handbuch zur Onomastik / *Name Studies*. An International Handbook of Onomastics / *Les noms propres*. Manuel international d'onomastique (= Handbücher zur Sprach- und Kommunikationswissenschaft 11,2). Berlin: De Gruyter 1465-1468.

Glasner, Peter (1999): "Ein sprachhistorischer Beitrag zur Semiotik der Stadt: das Pilotprojekt 'Kölner Straßennamen'", in: *Muttersprache* 109: 316-330.

Habscheid, Stephan / Reuther, Nadine (2013): "Performatisierung und Verräumlichung von Diskursen. Zur soziomateriellen Herstellung von ‚Sicherheit' an öffentlichen Orten", in: Felder, Ekkehard (ed.): *Faktizitätsherstellung in Diskursen*. Die Macht des Deklarativen (= Sprache und Wissen 13). Berlin / New York: de Gruyter 127–145.

Hausendorf, Heiko / Mondada, Lorenza / Schmitt, Reinhold (eds.) (2012): *Raum als interaktive Ressource* (= Studien zur Deutschen Sprache 62). Tübingen: Narr.

Hüttermann, Michael (2012): *DevOps for Developers*. New York: Apress.

Kesselheim, Wolfgang (2010): "'Zeigen, erzählen und dazu gehen': Die Stadtführung als raumbasierte kommunikative Gattung", in: Costa, Marcella / Müller-Jacquier, Bernd (eds.): *Deutschland als fremde Kultur*. Vermittlungsverfahren in Touristenführungen. München: Iudicum 244–271.

Liedkte, Gerhard (1999): *Abbestraße bis Zwetschenweg*. Straßennamen in Paderborn. Paderborn: H&S Verlag.

Lobbedey, Uwe (1986): *Die Ausgrabungen im Dom zu Paderborn 1978/80 und 1983* (= Denkmalpflege und Forschung in Westfalen 11). Bonn. Habelt.

Lobbedey, Uwe (1999): *Romanik in Westfalen*. Würzburg: Zodiaque-Echter.

Mayhew, Deborah J. / Follansbee, Todd J. (2012): "User Experience Requirements Analysis within the Usability Engineering Lifecycle", in: Jacko, Julie A. (ed.): *The Human-Computer-Interaction-Handbook*. Boca Raton, FL, USA: CRC Press 945–953.

Müller, Marcus (2012): "Geschichte als Spur im Text", in: Jacko, Julie A. / Bär, Jochen A. (eds.): *Geschichte der Sprache – Sprache der Geschichte*. Probleme und Perspektiven der historischen Sprachwissenschaft des Deutschen. Berlin 159–179.

Nübling, Damaris (2012): *Namen*. Eine Einführung in die Onomastik. Tübingen: Narr.

O'Halloran, Kay L. (ed.) (2004): *Multi-modal Discourse Analysis*. Systemic Functional Perspectives. London / New York.

Pöppinghege, Rainer (2005): *Geschichte mit Füßen getreten: Straßennamen und Gedächtniskultur in Deutschland* (= Paderborner Universitätsreden 94). Paderborn: Universitätsverlag Paderborn.

Pöppinghege, Rainer (2007): *Wege des Erinnerns*. Was Straßennamen über das deutsche Geschichtsbewusstsein aussagen. Münster: Agenda.

Quednau, Ursula (2011): *Handbuch der deutschen Kunstdenkmäler*. Nordrhein-Westfalen II. Westfalen. Berlin / München: Deutscher Kunstverlag.

Royce, Winston W. (1970): "Managing the development of large software systems", in: *Proceedings of IEEE WESCON* 26, 8: 328-388.

Rüsen, Jörn (1996): "Historische Sinnbildung durch Erzählen. Eine Argumentationsskizze zum narrativistischen Paradigma der Geschichtswissenschaft und der Geschichtsdidaktik im Blick auf nicht-narrative Faktoren", in: *Internationale Schulbuchforschung* 18: 501–543.

Sauerländer Willibald (1971): "Die kunstgeschichtliche Stellung der Figurenportale des 13. Jahrhunderts in Westfalen", in: *Westfalen* 49: 1-76.

Schüttpelz, Erhard (2006): "Die medienanthropologische Kehre der Kulturtechniken", in: Engell, Lorenz / Siegert, Bernhard / Vogl, Joseph (eds.): *Kulturgeschichte als Mediengeschichte (oder vice versa?)*. Weimar: Universitätsverlag Weimar 87–110.

Sharma, Sanjeev / Coyne, Bernie (2015): *DevOps for Dummies*. 2nd IBM Limited Edition. Hoboken: John Wiley & Sons, Inc.

Sutherland, Jeff / Schwaber, Ken (2007): *The Scrum Papers*. Nuts, Bolts, and Origins of an Agile Method. Boston: Scrum, Inc.

Tack, Wilhelm (1958): "Die Paradies-Vorhalle des Paderborner Domes und die Wallfahrt nach Santiago de Compostela", in: *Alte und Neue Kunst im Erzbistum Paderborn* 8: 27-62.

Uebernickel, Falk / Brenner, Walter / Pukall, Britta / Naef, Therese / Schindlholzer, Bernhard (2015): *Design Thinking – Das Handbuch*. Frankfurt am Main: Frankfurter Allgemeine Buch.

Ulrich, Hans / Krieg, Walter (³1974): *St. Galler Management-Modell*. Bern: Haupt.

Weber, Heike (2012): "Urbanisierung und Umwelt: Ein Plädoyer für den Blick auf Materialitäten, Ressourcen und urbane ‚Metabolismen'", in: *IMS*. Informationen zur modernen Stadtgeschichte 2: 28-35.

Wilk, Nicole M. (2015): "'Gebäude erzählen Geschichte(n)'. Medienlinguistische und diskursgrammatische Untersuchung zur multimodalen Herstellung historischer Stadt-Räume durch Schilder, Pulte, Stelen, Mobile Tagging und Apps", in: *Networks*. Die Online-Schriftenreihe des Projekts mediensprache.net 72: http://www.mediensprache.net/networx/networx-72.pdf [letzter Zugriff 20. Januar 2016].

Arthistory's Next Topmodel? Der Trend zur Ontologie

Schelbert, Georg
georg.schelbert@hu-berlin.de
Humboldt-Universität zu Berlin

Hohmann, Georg
g.hohmann@deutsches-museum.de
Deutsches Museum, München

Kuroczyński, Piotr
piotr.kuroczynski@herder-institut.de
Herder-Institut für historische Ostmitteleuropaforschung, Marburg

Raspe, Martin
raspe@biblhertz.it
Bibliotheca Hertziana - Max-Planck-Institut für Kunstgeschichte, Rom

Sektionsbeschreibung

Die "Digitalisierung" von Objekten hat viele Gesichter. Sie kann in einer formalisierten Beschreibung bestehen, in Abbildungen, in einer Kombination aus beidem oder in der geometrischen Nachbildung, sei diese maschinell generiert (gescannt) oder gezeichnet und konstruiert. Um mit den entstehenden digitalen Daten nachhaltig arbeiten zu können, bedarf es der Strukturierung und der Vereinheitlichung. Anders etwa als die Gegenstände des Bibliothekswesens, die Bücher, haben Artefakte der Archäologie, Kunstgeschichte und anderer Disziplinen keine standardisierten formalen Eigenschaften, die direkt erfasst werden könnten. Die Merkmale der Beschreibung müssen von außen definiert werden. Langjährige Diskussionen um zugleich spezifische und universell anwendbare Metadatenschemata wie etwa CDWA oder MIDAS zeugen von den besonderen Mühen, diese Beschreibung zu formalisieren. Ein anderer Weg, sich dem Objekt zu nähern besteht darin, es abzubilden, zu vermessen und nachzuformen. Es tritt uns dann in verschiedenen bildlichen oder graphischen Formaten bis hin zur 3D-Visualisierung entgegen. Sieht man einmal davon ab, dass die ab- und nachbildenden Formate im Fall komplexerer Artefakte oft so heterogen ausfallen, dass sie nur schwer prozessierbar sind, dokumentieren sie vor allem nicht ohne Weiteres die Bedeutung und den Sinngehalt der Gegenstände, so dass zusätzliche beschreibende Informationen notwendig bleiben.

Hinzu kommt der Umstand, dass viele der vermeintlichen Eigenschaften der Objekte eigentlich als historische Ereignisse anzusehen sind, die den Kontext der Objekte bilden. Dies trifft etwa für die Fragen zu, wer einen Gegenstand in Auftrag gegeben, geschaffen, benutzt oder erworben hat.

Um diese komplexen Zusammenhänge angemessen darzustellen, zu vermitteln, nachnutzbar zu speichern, ist es praktisch unerlässlich, ein stringentes Datenmodell einzusetzen. Als Modell, das auch Ereignisse in den Blick nimmt, war das CIDOC Conceptual Reference Model (CRM) ein weitgehender Neuansatz auf der Bühne der Metadatenschemata und seit es 2006 zum ISO-Standard gekürt wurde, ist es in aller Munde. Es gilt als Referenzmodell für die Dokumentation und den Austausch von Daten im Bereich „cultural heritage". Doch wie sieht die Anwendung aus? Ist es praktikabel? Was repräsentiert eigentlich die Ontologie – eine Abstraktion der historischen Wirklichkeit? Eine gedankliche Struktur, die das kulturhistorische Wissen innerlich zusammenhält? Oder lediglich ein Datenformat zum Austausch zwischen verschiedenen Projekten und Beständen? Wozu sollen wir überhaupt "modellieren" - reicht nicht einfaches "information retrival" oder "resource discovery"?

Die Sektion will anhand von drei verschiedenen Seiten aus Licht auf diese Fragen werfen.

Zunächst werden die ursprünglichen Planungen, der aktuelle Status quo und die zukünftigen Perspektiven der Praxis im Kulturerbe und Museumsbereich, für dessen Bedürfnisse das Conceptual Reference Model zuerst entwickelt wurde, durch einen ausgewiesenen Kenner des Gebietes, Georg Hohmann (Deutsches Museum, München), dargestellt.

Aus der spezifischen Perspektive der Forschung beleuchtet Martin Raspe (Bibliotheca Hertziana, Max-Planck-Institut für Kunstgeschichte) die Rolle von Datenmodell und Ontologie nicht nur für die Beschreibung der Artefakte, sondern auch für Referenzierbarkeit im ständig fortschreitenden Wissenschaftsdiskurs.

Einen dritten Perspektivpunkt bieten digitale Architekturmodelle. Piotr Kuroczynski (Herder-Institut Marburg und Mitbegründer der DHd-Arbeitsgruppe Digitale Rekonstruktion) wird die aktuellen Bestrebungen darstellen, die geometrischen Daten durch zusätzliche, semantisch organisierte Angaben zu kontextualisieren. Damit nähern sich die Konzepte des Strukturmodells (Datenmodell, Ontologie) und des repräsentierenden Modells (vereinfachtes Abbild, 3D-Modell) einander an und eröffnen weitere Fragen zur Datenmodellierung im Allgemeinen.

Ontologien und das kulturelle Erbe

Georg Hohmann, Deutsches Museum München

In den letzten Jahren haben sich Museen und museumsnahe Projekte zunehmend dem Thema Linked Open Data (LOD) zugewandt. Die prominentesten Vertreter sind in dieser Richtung sind derzeit das Rijksmuseum Amsterdam und das British Museum London, die beide nahezu ihren gesamten Datenbestand als LOD unter einer freien Lizenz zur Verfügung stellen. Beide Museen sind daher auch als einzige Vertreter in der aktuellen Linked Data Cloud vertreten (s. unter http://lod-cloud.net/). Zur Modellierung der Datenbestände werden dabei unterschiedliche Ansätze verfolgt.

Das Rijksmuseum veröffentlicht Daten zu fast 550.000 Sammlungsobjekten und bedient sich des Europeana Data Model (EDM), welches von der Europeana entwickelt wurde und dort als universelles Modell für die Datenaggregation eingesetzt wird (Dijkshoorn et al. 2014). Das EDM ist ein sehr leichtgewichtiges Modell, das darauf optimiert ist, sehr heterogene Datenbestände möglichst einfach zu integrieren.

Während hier die Abstraktion und die Einfachheit im Vordergrund stehen, ist in einem anderen Ansatz die Semantik der Daten von größerer Bedeutung. In Anlehnung an der ursprünglichen Idee von Tim Berners-Lee wird hier Linked Data als niederschwelliger Einstieg in das Semantic Web interpretiert. Grundlegend ist hier die Verwendung einer Ontologie zur Datenmodellierung, die im Gegensatz zu einfachen Modellen wie dem EDM die Entitäten der jeweiligen Domäne, also ihrer Diskursumgebung, sowie die Beziehungen zwischen diesen Entitäten eindeutig und explizit definiert.

Eine erste umfassende und nicht als Versuchsfeld konzipierte Anwendung einer solchen Ontologie im Bereich des Kulturerbes fand im archäologisch fokussierten CLAROS-Verbund statt. Die gesamten Daten stehen heute im Format des CIDOC Conceptual Reference Model (CRM) zur Verfügung. Das CRM ist eine formale Ontologie für das kulturelle Erbe, die von einer Arbeitsgruppe des Komitees für Dokumentation im Internationalen Museumsrat ICOM entwickelt wird. In Hinsicht auf ihren Umfang und ihrem Detailgrad ist diese Ontologie singulär im Bereich der historischen Wissenschaften. Die Nutzung des CRM wird seitdem vor allem von der Andrew W. Mellon Foundation durch Förderung des Projekts ResearchSpace vorangetrieben, dass sich zum Ziel gesetzt hat, das CRM als lingua franca für das kulturelle Erbe zu etablieren. Als bislang bedeutendstes Ergebnis dieser Bestrebung ist zu bewerten, dass das British Museum seine Bestände über eine CRM-kompatible Schnittstelle zur Verfügung stellt. Der frei verfügbare Datenbestand des British Museum enthält Daten von über 2 Millionen Objekten und ist damit einer der größten CRM-basierten Datenbestände weltweit.

Im Gegensatz zu Modellen, die wie das EDM durch ihre Einfachheit und ihren Abstraktionsgrad einfache Datenintegration gewährleisten sollen, verspricht das CRM eine semantische Vernetzung gleichzeitiger Integration der Daten. Doch während ersteres bereits erfolgreich eingesetzt wird, fristet letzteres bislang ein Nischendasein.

Es muss also die Frage nach den Gründen gestellt werden. Exemplarisch wird hier die LOD-Datenrepräsentation des British Museum auf Basis des CRM unter Verwendung der Web Ontology Language (OWL) herangezogen. Hier wurden viele Design-Entscheidungen getroffen, die Fehlstellungen der zugrundenliegende Ontologie offenbaren: Wie sind die im CRM definierten Eigenschaften von Eigenschaften, die in OWL untersagt sind, umzusetzen? Welche Auswirkung hat die Open World Assumption von OWL? Wann sollten sog. Shortcuts anstatt der voll ausgeführten Pfade verwendet werden? Sind CRM-Shortcuts als Property Chains in OWL umzusetzen? Wann sollten die bestehende Klassen erweitert werden, wann soll auf das Typen-System des CRM zurückgegriffen werden? Wie ist damit umzugehen, dass im CRM mehrere Möglichkeiten zu Repräsentation eines Sachverhalts geboten werden? Wie sind Identifikatoren externer Vokabulare und Datenbestände einzubinden?

Anhand dieser Fragen wird deutlich, dass das CRM viele interpretative Freiräume und noch einige Lücken offen lässt, die je nach Implementation anderes geschlossen werden, was trotz der Verwendung eines vermeintlich einheitlichen Modells die Datenintegration erschwert. Eine Lösungsmöglichkeit stellt die Einrichtung von sog. Profiles vor, wie sie auch bei der Umsetzung des bibliothekarischen Standards METS (Metadata Encoding & Transmission Standard) zu Einsatz kommen. Ein Profil für die Anwendung des CRM für Kerndaten von Museumsdokumentationssystemen könnte dazu eingesetzt werden, Modellierungsalternativen verbindlich zu regeln, die Nutzung spezifischer Vokabulare zu empfehlen oder Mapping-Aufgaben zu erleichtern.

Wer hats gesagt? Metadata und überprüfbares Wissen im kunsthistorischen Datenmodell

Martin Raspe, Bibliotheca Hertziana - Max-Planck-Institut für Kunstgeschichte, Rom

Seit sich der Mensch intellektuell mit Kunst- und Bauwerken aus der Vergangenheit beschäftigt, möchte er mehr darüber wissen: Er verlangt nach zusätzlichen “Daten”, nach Angaben über den Verfertiger, über die historische Epoche, über Material, Maße und Herstellungsverfahren sowie über die Bedeutung - wer hat’s gemacht, was stellt es dar? Seit er sich wissenschaftlich damit befasst, möchte er noch mehr wissen: Wie kommt unser Wissen zustande, wie gesichert ist es, auf welcher Grundlage beruht es - wer hat’s gesagt? Um diese beiden Aspekte soll es in dem Beitrag gehen. Die Datenmodelle, die derzeit in der Kunstgeschichte en vogue sind, berücksichtigen fast ausschließlich das erste

Wissensbedürfnis. Das zweite jedoch, welches eigentlich erst mit vollem Recht Wissenschaft genannt werden kann, wird in der digitalen Welt weiterhin auf traditionelle, vergleichsweise umständliche Weise erfüllt.

Während das "Allgemeinwissen" auf immer durchdachtere Weise abgebildet wird, beispielsweise mit Hilfe hochdifferenzierter, fachspezifischer Ontologien, wird das reflektierte, in Aussagen, Begründungen und Belegen bestehende Spezialwissen der Disziplinen auch heute noch in Fußnoten und Literaturverzeichnissen verwaltet. Nur ganz selten kann man ein Quellenzitat per Mausklick direkt zu seiner Herkunft verfolgen, und wenn, dann landet man meist bei einem Eintrag im Bibliothekskatalog, nicht an der zitierten Stelle. Die meisten online erhältlichen Textausgaben lassen sich nicht punktgenau von außen referenzieren, und wenn doch, dann jedenfalls nicht mit Hilfe eines standardisierten, persistenten Verfahrens. Die Adresse der Belegstelle, auf die man heute verweist, kann schon morgen geändert sein - schon führt der Link ins Nichts, wie das Gnomon der Sonnenuhr an einem bedeckten Tag.

Noch viel seltener besteht die Möglichkeit, vom Beleg aus die Links, die auf ihn verweisen, automatisch rückwärts zu verfolgen, also vom Zitierten zum Zitierenden zu gelangen. Das wäre ein großer Fortschritt, denn ein Hauptteil der wissenschaftlichen Arbeit besteht ja darin, zu einem Primärwerk die "fortuna critica" zusammenzutragen. Ohne Bezug auf den aktuellen Stand der Forschung bliebe die eigene Untersuchung ohne Fundament. Würden solche Verweise in standardisierter Form gemacht und von Suchmaschinen indexiert, könnte die Datenverarbeitung den Wissenschaftler erheblich entlasten. Dass das prinzipiell möglich ist, zeigen die "citation indexes" der Naturwissenschaften, die verzeichnen, welche Artikel wo und wie oft zitiert werden. Anstelle eines fragwürdigen Bewertungsinstruments könnte die Informationstechnologie dem Forscher ein neues, äußerst nutzbringendes Instrument an die Hand geben.

Dass dem so ist, liegt gewiss daran, dass in der Digitalen Kunstgeschichte anfangs nicht neue Anforderungen an ein wissenschaftliches Verweissystem im Vordergrund standen, sondern vor allem die Möglichkeiten der Digitalisierung des Primärmaterials gesehen und genutzt wurden. Kunst- und Quellenwerke standen dadurch in besserer Abbildungsqualität als je zuvor zur Verfügung, und zwar direkt auf dem eigenen Arbeitsplatz. Aufgrund der kontinuierlich wachsenden Zahl von Digitalisaten, die auf Servern in aller Welt verstreut lagen, wurde die Aufgabe wichtig, diese zu finden. Mit dem Aufkommen mächtiger Suchmaschinen wie Google, die die Indexierung dieser Primärdaten übernahmen, gewannen die sogenannten Metadaten an Bedeutung. Ziel der Anstrengungen ist seitdem, diese Metadaten möglichst zu vereinheitlichen, um weltweit mit den gleichen Suchbegriffen arbeiten zu können. Die üblichen Bezeichnungen der Artefakte, die Namen der Künstler und Auftraggeber und viele weitere Angaben werden zur Identifizierung und Katalogisierung herangezogen. Große Meta-Datenbanken wie die "Europeana" entstehen, in denen die Bestände einzelner lokaler Datenquellen gemeinsam, zentral und nach einheitlichen Prinzipien erschlossen werden. Digitale Editionen von Büchern und Texten finden Eingang in traditionelle Bibliothekskataloge und können nun auch auf konventionellem Wege gefunden werden.

Zentraler Gegenstand dieser Phase der elektronischen Geisteswissenschaften ist die digitale Ressource, und "resource discovery" ist das Schlagwort, das für die Kernaufgabe der Wissenschaft verwendet wird. Alle anderen Tätigkeiten sind dem Fischen im weltweiten Datenozean nachgeordnet. Die Devise lautet, zunächst möglichst viel in dieses Meer hineinzuwerfen und die Mühe, die Gegenstände bei Bedarf wieder herauszuangeln, der Zukunft bzw. einer anonymen community zu überlassen. "Suchen" heißt folgerichtig das erste (und längste) Kapitel in Hubertus Kohles (2013) Handbuch "Digitale Bildwissenschaft". Auf diese Anforderung antworten auch die Datenmodelle, die im Bereich der Bildwissenschaften entwickelt werden.

Die vorrangige Aufgabe ist zunächst die eindeutige Identifikation der Ressource. Sie muss präzise im Netz angesprochen werden, denn anders ist sie nicht zu finden. Paradoxerweise existiert für viele digitale Bucheditionen heutzutage nur noch eine einzelne Adresse im Netz. Weiß man diese nicht, kann man das Buch nicht auffinden. In analogen Zeiten musste man lediglich Titel, Autor und Jahr kennen, dann konnte man das Buch in zahlreichen Bibliotheken anhand seiner lokalen Signatur auffinden. Daher brauchte auf eine präzise Formalisierung nicht so streng geachtet zu werden. Heutzutage reicht ein Fehler bei der Eingabe der Metadaten, und die Ressource wird nicht gefunden. Normdaten, die möglichst international abzugleichen sind, und persistent identifiers sind das Mittel, mit dem man dem Problem beikommen möchte.

Andererseits hat die präzise Ansprache mit einem eindeutigen Identifikator auch den Vorteil, dass sie Verwechslungen ausschließt, die in der analogen Welt nicht selten waren, insbesondere bei weniger deutlich profilierten Museumsbeständen, aber selbst bei Kupferstichen und Zeichnungen vorkamen.

Die zweite Aufgabe des Metadaten-Modells, das zur Erschließung einer Ressource dient, ist die Klassifikation. Erst wenn man unterscheidet, ob "Paris Gütersloh" zur Kategorie "Ortsbezeichnung" gehört oder vielmehr eine "Person" repräsentiert, kann man fehlerhafte Suchergebnisse vermeiden.

Der dritte Aspekt eines kulturwissenschaftlichen Datenmodells sind die Beziehungen zwischen der Ressource und den Metadaten. Auch hier geht es in erster Linie darum, Verwechslungen zu vermeiden und die Treffermenge bei der Suche einzugrenzen oder überschaubar zu halten, nicht primär um die adäquate Repräsentation des Wissens über den Gegenstand. Es ist wichtig zu unterscheiden, ob das Kunstwerk, nach dem wir mit Hilfe eines Personennamens suchen, von

dieser Person geschaffen wurde, ob es von ihr gekauft oder von ihr beschrieben wurde. Demzufolge muss ein Datenmodell in der Kunstgeschichte für diese verschiedenen Fälle Platz bieten, am besten in möglichst umfassender Form, so dass alle in Frage kommenden Relationen berücksichtigt werden.

Der Gesamtbestand von möglichen Klassifikationen und Relationen kann formal beschrieben und in maschinenlesbarer Form gespeichert werden. Ein solcherart verzweigtes Datenmodell bezeichnet man heutzutage im informatischen Kontext als ”Ontologie”. Das bekannteste Beispiel in der Kunst- und Kulturwissenschaft ist das CIDOC Conceptual Reference Model (CRM), das durch seinen Status als ISO-Norm den Anspruch erhebt, ein gangbares Muster für den bereich cultural heritage abzugeben.

Das Ziel dieses Beitrags ist es nicht, Kritik an diesem sehr verdienstvollen Unterfangen zu üben, sondern seine Stellung im Bereich einer digitalen Geisteswissenschaft zu verorten. Die digital humanities sind längst nicht am Ende ihrer Möglichkeiten angekommen. Wie zu Beginn angedeutet, kann bei einer Metadaten-Erschließung noch nicht von einer Repräsentation begründeten und reflektierten Wissens die Rede sein, auch wenn im Laufe der Zeit immer deutlicher wird, dass die Interessen der Forschung sich nicht darauf beschränken, der digitalen Ressourcen besser habhaft zu werden. Das, was heute als Metadata bezeichnet wird, dient ja nicht nur der Erschließung digitaler Ressourcen, sondern repräsentiert eigene Forschungsgegenstände. Wenn ein Kunsthistoriker über einen prominenten Mäzen oder einen Kritiker forscht, dann steht dieser nicht als Digitalisat zur Verfügung. Das gleiche gilt für Gattungen, Kunstlandschaften, ikonographische Themen oder künstlerische Techniken. Datenmodelle, die derartige konzeptuelle Fragestellungen repräsentieren und das verfügbare Wissen in digitaler Form zu formen und zu speichern vermögen, gibt es noch nicht. Auch ICONCLASS ist letztlich lediglich der Versuch einer Normierung von Konzepten (im Sinne von Metadata). Die Codes unterstützen die Forschung über ikonographische Themen nur indirekt, indem man mit ihrer Hilfe entsprechend verschlagwortete digitale Ressourcen finden kann.

Noch ist das Datenmodell nichts als eine Daten-Model, ähnlich einer Guss- oder Backform, die bestimmt, in welchem Format und in welchem Zusammenhang die einzelnen Inhalte anzuordnen sind, damit sie nicht nur von menschlichen, sondern auch von Elektronengehirnen verarbeitet werden können. Ein wirkliches Modell des Wissens stellen sie noch nicht dar. Wissen ist immer gekoppelt an Autorität und Glaubwürdigkeit, und diese Aspekte werden bisher wenig berücksichtigt. Bei kaum einer inhaltlichen Angabe im digitalen Datenmodell ist klar ersichtlich, wer eine Erkenntnis formuliert hat, auf welcher Grundlage sie beruht, wann sie zuerst gemacht wurde und wer sie unterstützt. All das sind wesentlich Bestandteile des Wissens im Sinne einer wissenschaftlichen Nachprüfbarkeit. Diese muss derzeit immer noch auf konventionellem, sprich analogem Wege bestätigt werden.

Der Beitrag ist ein Plädoyer dafür, in Zukunft nicht nur der resource discovery Aufmerksamkeit zu schenken, sondern auch dem Auffinden von wissenschaftlicher Aussagen, Verweise und Evidenz. Manche dieser Aspekte werden derzeit unter dem Schlagwort “Annotationen” zusammengefasst, aber auch darin offenbart sich die Ressourcen-Zentriertheit der heutigen Debatte: Wissenschaftliche Aussagen und Hinweise, Kerngebiete des Wissens, werden reduziert auf Post-Its, die den Gegenständen angeheftet werden. Wissen, das dauerhaft zugänglich und nutzbar sein soll, würde man diesem Medium wohl nicht anvertrauen.

Die semantische Leiter hoch hinauf – Potenziale und Herausforderungen einer Applikationsontologie für digitale 3D Rekonstruktionen

Piotr Kuroczyński, Herder-Institut für historische Ostmitteleuropaforschung, Institut der Leibniz-Gemeinschaft, Marburg

Seit den 1990er Jahren beobachten wir den Einsatz von Computern bei der Rekonstruktion und Vermittlung von verlorenen und/oder nie existierenden Kunst- und Bauwerken. Bislang dienen die Ergebnisse der „Virtuellen Rekonstruktion“ meist dem Wissenstransfer in Form eines Bildes oder einer Filmanimation. Die digitalen 3D-Modelle werden selbst an sich weniger als ein Forschungsgegenstand bzw. als ein Informationsmodell betrachtet.

Die rapide Entwicklung der Informationstechnologien führt in letzter Zeit zu weit sichtbaren Veränderungen analog basierter Forschungsmethoden in den Geistes- und Sozialwissenschaften. Zugleich etabliert die neu bildende Querschnittsdisziplin „Digital Humanities“ innovative Methoden und Standards innerhalb der Geisteswissenschaften.

Das wirkmächtigste Konzept des derzeitigen Transformationsprozesses ist sicherlich das Semantic Web, vor allem in der Form von Linked Open Data (LOD). Hierbei bilden kontrollierte Vokabulare, Thesauri und Ontologien, indem sie die Typen, Eigenschaften und Beziehungen zwischen Entitäten in einem Diskursraum definieren, das formelle Rahmenwerk zur Wissensorganisation und Wissensrepräsentation. Auf dem Gebiet des kulturellen Kulturerbes (Cultural Heritage) stellt CIDOC-CRM den nunmehr seit längerer Zeit kontinuierlich vorangetriebenen Versuch dar, den „semantischen Leim“ zwischen den Informationseinheiten fachübergreifend zu schaffen.

Im Fall der „Virtuellen Archäologie” und „Digitalen Kunstgeschichte“ fokussiert sich die semantische

Erschließung derzeit noch vor allem auf digitale Bildarchive und Objektkataloge. Stellvertretend sind hier vor allem die digitale Bibliothek EUROPEANA mit ihrem Datenmodell und die damit verbundene Entwicklung von Metadaten Schemata zu nennen.

Da die zunehmende Digitalisierung von Kulturgut auch zu - in Zukunft vermutlich exponentiell schnell - wachsenden Beständen von 3D-Daten führt, gibt es bereits mehrere EU-Projekte, die an gemeinsamen Beschreibungsformaten (Metadaten Schemata), z. B. 3D-COFORM, CARARE und 3D-ICONS, arbeiten. Aktuell kommt den Bedürfnissen der Forschung, die an einer möglichst umfassenden Dokumentation und semantischen Erschließung der 3D-Datensätze interessiert ist, das Metadaten Schema CARARE 2.0 entgegen. Dieses Schema entspricht vor allem den Anforderungen hinsichtlich der Beschreibung und semantischen Anreicherung von 3D-Digitalisaten, die beispielsweise mit moderner Fotogrammetrie, Streifenlichtscanning, Laserscanning oder der „Structure from Motion“ Technologie von existierenden räumlichen Objekte erstellt werden. Die CARARE 2.0-Basiskategorien – wie Heritage Asset, Activities, Digital Resources und Collections – berücksichtigen jedoch nicht die besonderen Anforderungen einer digitalen hypothetischen 3D-Rekonstruktion von nicht (mehr) existierenden Kunst- und/oder Bauwerken. Somit eignet sich dieses Schema vorzugsweise für die Beschreibung von 3D-Digitalisaten von existierenden Objekten, weniger aber für die Beschreibung von hypothetischen, hoch interpretativen, 3D-Rekonstruktionen auf der Basis primärerer und sekundärer Quellen. Dies ist aber gerade das Arbeitsfeld der (kunst-)historischen Wissenschaften, die nicht nur Vorhandenes in einem anderen Medium abbilden, sondern wissensbasierte Annahmen über vergangene oder vielleicht auch nur geplante Sachverhalte in ihrer ganzen Komplexität erforschen und dokumentieren möchten.

Aus der Sicht der digitalen 3D-Rekonstruktion wird ein Metadaten Schema mit dem Namen „Cultural Heritage Markup Language“ (CHML) entwickelt, das den gesamten Arbeitsablauf einer hypothetischen 3D-Rekonstruktion von der Datenerhebung, über die Datenverarbeitung bis zum händischen 3D-Modellieren am Rechner ausdrückt.

Der Vortrag erläutert zunächst die Idee vom Semantic Web sowie die existierenden Standards und Technologien. Anschließend werden die Vorteile und Herausforderungen der Entwicklung einer auf CIDOC-CRM referenzierten Applikationsontologie für die digitale 3D-Rekonstruktion von verlorenen Kunst- und Bauwerken, kritisch beleuchtet.

Ein konkreter Vergleich zwischen CARARE 2.0 und CHML soll die Unterschiede der Metadatenstandards und die Stärken von CHML für die Dokumentation von hypothetischen 3D-Rekonstruktionen aufzeigen. An konkreten Beispiel wird die Implementierung von CHML in eine Applikationsontologie (OWL DL, Erlangen-CRM) gezeigt. Dabei werden insbesondere die Herausforderungen der fachspezifischen Abbildung von Sachverhalten, wie beispielsweise von „nicht existierender Objekten“, in CIDOC-CRM behandelt.

Abschließend möchte die Präsentation eine mittelfristige Perspektive aufzeigen, welchen Einfluß die Semantic Web Technologien auf die digitale 3D-Rekonstruktion im Wissenschaftsbereich haben können, indem sie fundierte, informationsgesättigte semantische 3D-Modelle anstelle der zurzeit vorwiegend auf reine Visualisierung bezogenen 3D-Modelle etablieren. Im Ausblick werden die Debatten um „Serious 3D”, die semantische Anreicherung, eine äquivalente Dokumentation und die Probleme der Publikation von 3D-Daten vor dem Hintergrund des aufkommenden Web 3.0 (Semantic Web) aufgerufen

Bibliographie

Dijkshoorn, Chris / Ter Weele, Wesley / Jongma, Lizzy / Aroyo, Lora (2014): "The Rijksmuseum Collection as Linked Data", in: *Semantic Web Journal* http://www.semantic-web-journal.net/content/rijksmuseum-collection-linked-data . [letzter Zugriff 05. Februar 2016].

Kohle, Hubertus (2013): *Digitale Bildwissenschaft*. Glückstadt: Verlag Werner Hülsbusch.

Transbiblionome Daten in der Literaturwissenschaft. Texttechnologische Erschließung und digitale Visualisierung intertextueller Beziehungen digitaler Korpora

Wagner, Benno
wagner@lit-wiss.uni-siegen.de
Beijing Institute of Technology, China, Volksrepublik

Mehler, Alexander
mehler@em.uni-frankfurt.de
Universität Frankfurt

Biber, Hanno
Hanno.Biber@oeaw.ac.at
Österreichische Akademie der Wissenschaften

Introduction

> "digital humanities" boils down to using computers to do exactly the same silo-ed and intellectually buttoned down work that people did before. But it's always easier to get money for equipment (i.e. computers to make a million concordances) than it is to re-envision a field. People in this kind of digital humanities are very concerned with "preservation" in every sense of the word — preservation of the status quo, of themselves and their jobs, and of the methods and fields of the past.
>
> (Lisa Nakamura)

> Letztlich reicht es nicht aus, auf Seiten der Objektbasis unablässig neu digitales Material zu akkumulieren. Parallel dazu müsste auf Seiten der Forschung die Bereitschaft zum aktiven Einsatz technologisch und methodisch innovativer Verfahren gefördert werden. Die Digitalisierung allein ohne eine begleitende Theoriedebatte und ohne ein verfeinertes methodisches Rüstzeug betreiben zu wollen dürfte zu verkürzten Ergebnissen führen.
>
> (Embach / Andrea 2008)

Die Sektion und ihr Leitkonzept, d. h. transbiblionome Daten, reagiert auf die zitierten Postulate, indem sie die Konferenz-Leitfrage: „Was sind Daten in den Geisteswissenschaften?" an einem spezifischen Aspekt literarischer Texte und ihrer Nutzung aufgreift, nämlich ihrer Intertextualität. Seit den Medienverbünden der literarischen Moderne funktioniert Literatur essentiell als Intertext, als Relais zwischen anderen Texten. Lektüre wird damit zu einer Suche „von Buch zu Buch" nach „etwas, was zwischen den Inhalten aller einzelnen Bücher schwebt, was diese Inhalte in eins zu verknüpfen vermöchte" (von Hofmannsthal 2000: 111). Im Unterschied zu *paratextuellen* Daten wie Textvarianten, Konkordanzen, Erläuterungen, Quellen etc. sind *intertextuelle* Daten relational, d. h. sie bezeichnen die (z. B. semantischen) *Beziehungen* eines literarischen Textes (Matrix-Text) zum Archiv aller vorgängigen und zeitgleichen Texte. Sie sind Daten (gegeben) in dem eingeschränkten Sinn, dass sie jeweils aus der Differenz zwischen zwei Texten emergieren. Diese *transbiblionomen*, die Ordnung des Buches und seiner Einheiten überschreitenden Daten bewirken zudem eine laufende Modifizierung der Daten des Matrix-Textes, indem sie dessen (lexikalischen, syntaktischen und semantischen) Einheiten je nach Kontextualisierung unterschiedliche Signifikanz bzw. Verknüpfungsdichte zuweisen. Hieraus ergibt sich für den textbezogenen Sektor der *Digital Humanities* die Herausforderung, das transbiblionome Medium des Computers nicht länger für rein biblionom formulierte Aufgaben zu verwenden, sondern dessen Potential für die Erschließung, Aufbereitung und kollaborative Nutzbarmachung transbiblionomer (intertextueller) Daten zu erproben.

In den Beiträgen dieser Sektion (1) wird eine literatur- und medientheoretische Definition transbiblionomer Daten gegeben, ihre Problematisierungsgeschichte knapp resümiert und ihr Stellenwert für Literaturforschung im digitalen Zeitalter illustriert; (2) ein erfolgreich etabliertes digitales Korpus (fackel-online des Austrian Academy Corpus) vorgestellt, dessen Textbestand aufgrund seiner programmatischen (zitierenden, anspielenden, parodierenden) Referenz auf ein umfangreiches Korpus anderer Texte (die zeitgenössische Presseberichterstattung) besonders nachdrücklich die Öffnung auf transbiblionome Daten (die extrem große Zahl der zitierten Pressetexte) nahelegt; (3) eine Texttechnologie namens *Wikidition* zur automatischen Generierung lexikalischer, syntaktischer, semantischer und textueller Links vorgeführt, die es Literaturforschern erstmals ermöglichen wird, intertextuelle Bezüge zwischen Matrix-Texten und großen Kontext-Korpora (etwa Jahresbänden einer Tagezeitung) umfassend zu erschließen, zu evaluieren, und zu visualisieren. Die Sektion reagiert damit theoretisch und praktisch auf die Leitfragen nach dem theoretischen und technologischen Status von Daten in den *Digital Humanities* sowie nach den Verfahren ihrer Modellierung und Visualisierung.

Transbiblionome Daten: Problemgeschichtlicher Abriss und aktuelle Herausforderungen

Benno Wagner; Beijing Institute of Technology

Moderne Literatur funktioniert essentiell intertextuell und intermedial. Mit dem Heraufkommen der neuen Konkurrenz-Medien (Photographie und Film, Telegraf und Telefon) und mit der Ablösung der Referenzinstitution Bibliothek durch globale Informationssysteme (Rayward 2008) sowie das alle Lebensbereiche durchdringende Wissen moderner Verwaltungen existiert Schrift nurmehr im Spannungsverhältnis zwischen Buchgebundenheit und einem zunehmend *transbiblionom* organisierten kulturellen Kontext. Der literarische Text gerät auf diese Weise einerseits zum „geometrischen Ort eines hors-texte", zu einem „Kreuzungspunkt von Schichten, die Myriaden von Horizonten entspringen" (Topia 1984: 103). Anderseits wird Literatur unter diesen Bedingungen zu einer besonderen Instanz des kulturellen Gedächtnisses. Sie lässt sich als komplexer „Spurenkörper" (Pêcheux 1983: 55) beschreiben, oder – jedenfalls in ihren raffiniertesten und reflektiertesten Schreibweisen – auch als „hypermnemische Maschine" (Derrida 1984: 147), als dynamischer Erinnerungsapparat, dessen virtuelles Verweispotential auf andere Texte die Ordnungsraster realer Wissensspeicher (Enzyklopädien, Bibliotheken,

Archive) durchkreuzt. Bilden die letzteren Datensätze aus Relationen erster Ordnung, so sind die genuin intertextuellen Daten stets Relationen zweiter Ordnung (Relationen von Relationen). M.a.W.: sie lassen im Paradigma eines Textsyntagmas nicht nur lexikalische Einheiten erscheinen, sondern ganze Textfragmente, also wiederum Syntagmen.

So schrieb Derrida in der *Grammatologie* zunächst:

> „Es geht […] nicht darum, der Buchhülle noch nie dagewesene Schriften einzuverleiben, sondern endlich das zu lesen, was in den vorhandenen Bänden schon immer zwischen den Zeilen geschrieben stand. Mit dem Beginn einer zeilenlosen Schrift wird man auch die vergangene Schrift unter einem veränderten Organisationsprinzip lesen. […] Was es heute zu denken gilt, kann in Form der Zeile oder des Buches nicht niedergeschrieben werden; ein derartiges Unterfangen käme dem Versuch gleich, die moderne Mathematik mit Hilfe einer Rechenschiebermaschine zu bewältigen.“ Stattdessen avisiert er, diesmal mit Leroi-Gourhan, „eine andere[n], bereits vorstellbare[n] Art der Speicherung […], deren rasche Verfügbarkeit der des Buches überlegen sein wird: die große ‚Magnetothek‘ mit elektronischer Auswahl wird in naher Zukunft vorselektierte und sofort verfügbare Informationen liefern“ (Derrida 1967: 154-155).

Knapp zwei Jahrzehnte später hingegen, in einem Aufsatz aus dem Jahre 1984, konfrontiert uns Derrida mit einem ganz anderen und scheinbar diametral entgegengesetzten Szenario. Im Bezug auf den *Ulysses* von Joyce heißt es nun:

> „for there be no simple confusion between him [Joyce] and a sadistic demiurge, setting up a hypermnesiac machine, there in advance, decades in advance, to compute you, control you, forbid you the slightest inaugural syllable because you can say nothing that is not programmed on this 1000th generation computer […] beside which the current technology of our computers and micro-computerfied archives and translating machines remains a bricolage of a prehistoric child's toys” (Derrida 1984: 147).

Hier hat sich offenbar das Komplexitätsgefälle zwischen Druckschrift und elektronischem Speicher verkehrt. Die lineare Schrift ist nicht länger Komplexitäts-Engpass, sondern sie fungiert selbst als Quelle einer überbordenden Komplexität. *Sie* ist nun der "Computer der 1000. Generation", im Vergleich zu dem die elektronischen Speichermedien als Problem erscheinen, als eine dem Gegenstand der Druckschrift unangemessene, prähistorische Spielerei.

Betrachtet man nun den Einsatz von Computern zu Zwecken der Literaturforschung seit den 1990er Jahren, so drängt sich der Eindruck auf, als habe jedes der beiden Zitate eine Arbeitsperspektive eröffnet, die von der jeweils anderen nichts zu wissen scheint. So haben Autoren wie George Landow und Jay Bolter, in einer eigentümlichen Einebnung des Unterschieds zwischen der syntagmatischen und der paradigmatischen Text-Dimension, den elektronischen Hypertext kurzerhand zu jenem Medium deklariert, mit dessen Hilfe sich das intertextuelle Verweispotential eines literarischen Textes restlos implementieren, der vieldimensionale literarische Text sich aus dem Zwang der Zeile befreien ließe. Hierzu konstatiert Moritz Baßler:

> „Landows Parallelisierung von Hypertext mit jenem poststrukturalistischen Textbegriff, den Barthes in S/Z entwickelt, setzt sich über den elementaren Unterschied von syntagmatischer und paradigmatischer Textdimension großzügig hinweg. […] Dabei handelt es sich jedoch um zwei vollkommen verschiedene Dinge, denn ein Hypertext mag so nonlinear sein wie er will – das betrifft doch immer nur die Sequenz, in seiner paradigmatischen Dimension dagegen unterscheidet er sich nicht vom normalen Text“ (Baßler 2005: 307-308).

Sehr viel erfolgreicher gestaltet sich computergestützte Literaturforschung immer dann, wenn sie die transbiblionome Dimension der Intertextualität von vornherein aus ihrem Gegenstandsbereich ausschließt. Dies geschieht zumeist stillschweigend, bisweilen aber auch mit programmatischem Nachdruck, wenn sich etwa die in Deutschland etablierte Computerphilologie explizit auf die Befassung mit „traditionellen philologischen Gegenständen“ und damit das Potential des Computers auf die Optimierung biblionomer Funktionen sowie auf die Herstellung dezentraler Forschungskollaborationen beschränkt (Meister 2005: 326). Als ein auf dieser biblionomen Basis erfolgreich etabliertes Projekt stellen wir die digitale Edition der *Fackel* des Austrian Academy Corpus vor. Das digitale *Fackel*-Korpus kann zugleich als Testfall für die hier postulierte transbiblionome Erweiterung dienen, weil sein Textbestand aufgrund der programmatischen (zitierenden, anspielenden, parodierenden) satirischen und polemischen Referenz auf umfangreiche Korpora anderer Texte (die zeitgenössische Presseberichterstattung, die zeitgenössische Literatur) besonders nachdrücklich die Öffnung auf intertextuelle Daten (die extrem große Zahl der zitierten Pressetexte) nahelegt.

Eine auf intertextuelle Verlinkung orientierte *transbiblionome Literaturforschung* wird also auf biblionome Digitalisierungen von Datensätzen erster Ordnung funktional aufsetzen. Ihre methodische und technische Entwicklung hätte sich vor dem skizzierten Erfahrungshintergrund an drei Leitlinien zu orientieren:

(1) *Zielsetzung*: Transbiblionome Digital Humanities zielt auf die computergestützte Erschließung und Darstellung der intertextuellen und intermedialen Dimension literarischer Texte jenseits einer Beschränkung auf einen biblionomen Forschungshorizont und der Fixierung auf das Visualisierungspotential von Hypertext. Als technische Grundlage hierfür hätte eine auf die Generierung und Verwaltung intertextueller Daten

zielende digitale Arbeitsumgebung zu dienen, die zugleich die dezentrale Kollaboration von Experten(gruppen) und eine nutzerspezifisch differenzierte Aufbereitung der Forschungsergebnisse ermöglicht. In dieser Sektion stellen wir mit *Wikidition* eine Software vor, welche die ersten beiden Aspekte dieses Anforderungsprofils texttechnologisch umsetzt.

(2) *Theorie*: Unter den genannten Bedingungen kann und muss sich das zugrundeliegende Intertextualitäts-Modell der methodischen Alternative entziehen, mittels derer die printorientierte Methodendiskussion (insbesondere der 70er und 80er Jahre) sich um eine „Zähmung“ (Lachmann 1984: 137) des transbiblionomen Potentials ‚moderner‘ Intertextualität bemüht hatte: Hier eine ‚geschlossene‘, durch unterstellte Verknüpfungsabsichten des Autors oder faktische Verknüpfungsoperationen des Lesers begrenzte, dort eine unmittelbar auf das System der langue bezogene ‚offene‘ und daher, zumal unter Bedingungen einer printfixierten Forschung, forschungspraktisch niemals einholbare Intertextualität. Baßlers Entwurf eines ‚archivimmanenten Strukturalismus‘ (Baßler 2005), der den intertextuellen Raum (das ‚Paragrammaire‘ nach J. Kristeva) eines Bezugstextes auf eine historische Positivität von Kontext-Dokumenten bezieht, die er ‚Archiv‘ nennt, kann hier als fundierte und konstruktive Alternative dienen, deren Begrifflichkeit sich unmittelbar auf die Konzepte und Leistungen einer digitalen Texttechnologie beziehen lässt. Präzisierungen und Erweiterungen der Theorie werden dort anzustreben sein, wo das 'Archiv' nicht nur den biblionomen Raum, sondern zugleich den der Textualität überschreitet, indem es sich multimedial konstituiert.

(3 *) Methode*: Intertextuelle Computerphilologie dieser transbiblionomen Art zielt nicht auf die *Implementierung* literarischer Intertextualität: ihre ‚Befreiung‘ aus dem ‚Gefängnis‘ der Druckzeile und ‚vollständige Entfaltung‘ im barrierefreien digitalen Schreibraum, sondern auf ihre *Supplementierung*: auf die forschungstechnische Unterstützung der selbstverständlich stets selektiven und (projekt- und methodenspezifisch) perspektivischen Erschließung des intertextuellen Potentials eines je gegebenen literarischen Texts. Bei der Entwicklung einer zweckmäßigen Arbeitsumgebung hätte die Kooperation zwischen Philologie, Medienwissenschaft, Texttechnologie und Informatik einer Logik *pragmatischer Schnittstellenbildung* zu folgen. Statt entweder digitale Lösungsmöglichkeiten mit philologischen Problemstellungen zu überfordern, oder umgekehrt von vornherein die philologischen Problemstellungen an das Leistungsvermögen digitaler tools anzupassen, wären für jede Teilaufgabe die Schnittstellen zwischen humaner Intelligenz und künstlicher Intelligenz präzise zu definieren, um die Leistungsvermögen von Menschen und Rechnern möglichst effizient miteinander zu verschalten.

Ausgehend von diesen Überlegungen stellen wir, mit den Bezugshorizonten Franz Kafka und Karl Kraus, eine texttechnologische Anwendung (*Wikidition*) für die Entwicklung eines *Literary Memory Information System* (LiMeS) vor. Dh. einer literaturwissenschaftlichen Forschungsumgebung, die *literarische Texte* nicht einfach als Gegenstände, sondern *als Medien des kulturellen Gedächtnisses* behandelt, indem sie ihr intertextuelles Verweispotential erschließbar, darstellbar und für unterschiedliche Verwertungszusammenhänge nutzbar macht. Die texttechnologischen Entwicklungen der letzten Jahre bieten u. E. eine tragfähige Basis für die Konzeption und Implementierung einer digitalen Arbeitsumgebung für eine solche intertextuell und transbiblionom orientierte Literaturforschung.

"Die grellsten Erfindungen sind Zitate". Corpusbasierte Erkennung und Analyse von Zitaten und intertextuellen Referenzen in literarischen Texten

Hanno Biber; Österreichische Akademie der Wissenschaften

„Die unwahrscheinlichsten Taten, die hier gemeldet werden, sind wirklich geschehen; ich habe gemalt, was sie nur taten. Die unwahrscheinlichsten Gespräche, die hier geführt werden, sind wörtlich gesprochen worden; die grellsten Erfindungen sind Zitate.“ (Kraus 1926: VII) Im Vorwort der Buchausgabe des aus den die Worte und Taten, die Presseberichte und Propaganda, die Phrasen und Tonfälle des Ersten Weltkrieges aus- und aufrufenden, satirisch in Szene gesetzten Dokumenten bestehenden monumentalen Dramas "Die letzten Tage der Menschheit" steht diese Aussage von Karl Kraus zu seiner satirischen Methode des Zitats. Sie hebt hervor, wie der Autor das gestaltet hat, was er für seine Leser, die noch hören und lesen konnten, was und wie es gesagt wurde, wie auch für zukünftige Leser hörbar und sichtbar gemacht hat, in den aus Zitaten bestehenden Szenen seines Dramas ebenso wie in den Texten seiner Zeitschrift *Die Fackel*.

Die *AAC-Fackel* , die digitale Ausgabe der von Karl Kraus vom 1. April 1899 bis Februar 1936 in Wien herausgegebenen Zeitschrift *Die Fackel* wurde unter Anwendung computerphilologischer Methoden im Rahmen des *AAC-Austrian Academy Corpus* an der Österreichischen Akademie der Wissenschaften erstellt. Das literarische Werk des wohl bedeutendsten Satirikers deutscher Sprache und darin die 1899 bis 1936 in Wien herausgegebene Zeitschrift *Die Fackel* des 1874 in Jičín in Böhmen geborenen Karl Kraus, der in Wien gelebt, in vielen Städten Europas seine Texte vorgetragen hat und weit darüber hinaus rezipiert wurde und wird, ist als überaus bedeutender Beitrag der deutschsprachigen Literatur zur Weltliteratur zu betrachten. Die Überlieferung und Verfügbarkeit der satirischen und polemischen Texte

im digitalen Medium der *AAC-Fackel*, der Texte der *Fackel* mit ihrer thematischen Vielfalt, sprachlichen Komplexität und historischen Relevanz, muss für eine von computergestützten Verfahren der digitalen Literaturwissenschaft bestimmten wissenschaftlichen Beschäftigung, die sich der zeitgemäßen Erforschung des literarischen Werkes nach neuesten wissenschaftlichen Standards verpflichtet, als eine überaus wichtige Aufgabe betrachtet werden. In der digitalen Edition der *AAC-Fackel* wird der gesamte Text dieser Zeitschrift nicht nur einer an Literatur, Sprache und Geschichte interessierten Leserschaft zugänglich gemacht, sondern auch die einzigartige sprachliche und literarische Qualität dieser Texte unter Nutzung computerphilologischer Methoden und texttechnologischer Instrumente durch verschiedene Suchmöglichkeiten und Register in einer neuen Weise erschlossen.

Die Methode der Satire von Karl Kraus ist das Zitat. Seine Texte folgen mithin einer transbiblionomen Logik des laufenden Verweises auf andere Texte. Aus diesem Grunde kann unter Nutzung texttechnologischer Möglichkeiten, wie sie in der im anschließenden Beitrag beschriebenen *Wikidition* implementiert werden, das digital verfügbare Textcorpus der *Fackel* zur Analyse der in ihm zitierten Texte genutzt werden. Karl Kraus findet zumeist in den zeitgenössischen Texten der Presse und der Publizistik die Prätexte für seine Texte, indem er das, was er verarbeitet, einfach zitiert, oder einschöpft in seine Texte und satirisch kommentiert. In seinen satirischen Texten muss er nichts mehr erfinden, sondern vielfach einfach nur zitieren, in Anführungszeichen setzten, den Text einbetten, ihn graphisch so anordnen, dass seine satirische Qualität sichtbar werden kann, dem Leser sichtbar gemacht werden kann. Die in den tausenden Zitaten wiedergegebenen, satirisch bearbeiteten Sprachstücke der *Fackel* sind vom Autor montiert und typographisch aufbereitet, was in vielen Fällen, besonders in den Glossen der 'Fackel' durch einfache Spationierung von bestimmten Textpassagen erfolgen kann. Die Textsorte der Glosse ist beispielhaft für das satirische Verfahren von Karl Kraus, wo es auf die besondere Behandlung eines zitierten, aus der Presse stammenden Prätextes ankommt. Kurt Krolop schreibt dazu in einem Aufsatz über Karel Čapek und Karl Kraus über die wichtigste Funktionsweise der Satire von Karl Kraus: „Das auf solche Weise schöpferische Vernichtungsarbeit leistende Zitat, in dem die Spationierung den Gestus sichtbar macht, reduziert wolkigen Bombast auf den zugrunde liegenden reinen Unsinn, vorgetäuschte Fülle auf die tatsächliche Leere, falsches Pathos auf das echte Blech, die hohle Phrase auf die dicke Lüge. Der gestischen Entlarvung dient auch die für die Fackel so charakteristische Verschränkung von klassischem Sprachgut und Jargon, welche etwa die wahre Haltung eines liberalen Journalisten zur Anschauung bringen soll, wenn dieser Posas berühmte Aufforderung zitiert und das so bewirkt „Sire, bittsie, was liegt Ihnen schon dran, geben Sie Gedankenfreiheit!“ (890,276)“ (Krolop 1992: 305-306.) Der Text „Blendwerk der Hölle“ (F 366,30)[1], in dem der Satiriker auch am Ende seines Textes über die besondere Funktionsweise seiner Satire reflektiert, ist beispielhaft für dieses dann in der satirischen Tragödie "Die letzten Tage der Menschheit" dramatisch gestalteten Verfahrens. Diese Glosse ist wie viele andere auch Teil einer Reihe von oft thematisch in Zusammenhang stehenden Glossen in den Heften der Zeitschrift, einer für diese besondere Form der Satire von Karl Kraus typischen Textsorte und thematisiert in besonderer Weise die Arbeit des Satirikers mit der „übervollen Realität“ (F 366,32), wie sie sich ihm sprachlich deutlich und in seinen eigenen Texten durch Zitierung in seiner Zeitschrift deutlich gemacht in der Presse seiner Zeit darbietet. Karl Kraus demonstriert und reflektiert seine Methode der Satire, die nicht Satiren verfassen muss, sondern Texte aus der Zeitung bloß noch anordnen muss, damit sie zu Satiren werden. Diese besondere Methode der Satire ist die Methode des Zitats. Der Schriftgrößenwechsel zur Kenntlichmachung des Zitats und die Hervorhebung einzelner sprachlicher Äußerungen durch Sperrung sind der zeitgenössischen typographischen Konvention entliehene Zeichen, die hier zum kritischen Zweck der Hervorhebung im satirischen Text eingesetzt werden.

Im Zentrum des satirischen Verfahrens steht das Zitat. Die zitierten Prätexte der satirischen Texte, die in den Glossen, im Drama und in den anderen Texten von Karl Kraus aufgerufen und zitiert werden, können mit Hilfe von texttechnologischen und corpusbasierten Methoden der digitalen Literaturwissenschaft ebenso wie andere intertextuelle Verbindungen zwischen den Texten aufgefunden und systematisch erforscht werden. Wie erste Versuche eines solchen Verfahrens aussehen könnten, wäre ein wichtiger Beitrag zur Erforschung der komplexen Textreferenzen der *Fackel*. Als ein wichtiges Instrument dazu kann neben anderen zur Verfügung stehenden Ressourcen die in der Österreichischen Akademie der Wissenschaften erstellte Textsammlung des *AAC-Austrian Academy Corpus* dienen. Das *AAC* ist ein umfangreiches und komplex strukturiertes Textcorpus zur deutschen Sprache und Literatur im Untersuchungszeitraum von 1848 bis 1989 und Basis für die Entwicklung texttechnologischer Anwendungen im Bereich der Corpusforschung sowie philologischer und textwissenschaftliche Forschungen, die am Institut für Corpuslinguistik und Texttechnologie der Österreichischen Akademie der Wissenschaften mit diesen Textressourcen neben anderen Textcorpora in den Bereichen der Corpusforschung im Kontext digitaler Literaturwissenschaft und philologischer Grundlagenforschung durchgeführt werden. Die *AAC-Fackel* kann neben der digitalen Edition der literarischen Zeitschrift *Der Brenner* in diesem Kontext als exemplarischer Anwendungsfall einer aus der Corpusforschung und computergestützten Textwissenschaft entstandenen digitalen Musteredition eines literaturgeschichtlich bedeutenden Textes

betrachtet werden. Ihr Zustandekommen und die dafür notwendigen Bedingungen resultieren aus einer sich mit Sprache und Fragen des besonderen Sprachgebrauchs auf empirischer Textbasis in einem bestimmten, mit besonderen Eigenschaften ausgestatteten Textcorpus befassenden Forschungsrichtung, wie sie im *ICLTT* und seinen Vorgängerunternehmungen seit vielen Jahren betrieben wird. Die computergestützte Auffindung von Zitaten in Texten und der Erforschung des Gebrauchs von Zitaten in Texten, die der Satiriker Karl Kraus in seinen Texten literarisch nutzt, wird durch große Textcorpora erst systematisch möglich und soll als exemplarischer Anwendungsfall einer derart texttechnologisch ausgeübten Literaturwissenschaft verstanden und gezeigt werden.

Automatische intra- und intertextuelle Vernetzung literarischer Texte mit Hilfe von Wikidition

Alexander Mehler; Universität Frankfurt

Der Beitrag führt *Wikidition* als neuartiges *Literary Memory Information System* ein, das die in den beiden vorigen Beiträgen theoretisch und im Hinblick auf ein existierenden digitales Korpus formulierten transbiblionomen Postulate einlösen soll. *Wikidition* zielt auf die automatische Generierung von online Editionen von Korpora natürlichsprachlicher Texte. Dazu zählen insbesondere literarische Texte, welche als Matrixtexte im Kontext ihrer Echotexte untersucht werden (siehe Abbildung 1). *Wikidition* kombiniert eine Vielzahl von Text-Mining-Verfahren für die automatische Vernetzung (*Linkification*) von Wort-, Satz- und Texteinheiten unter den Bedingungen ihrer syntagmatischen *Verwendung* einerseits und ihres paradigmatischen *Gebrauchs* andererseits. Darüber hinaus beinhaltet *Wikidition* eine Lexikonisierungskomponente zur Ermittlung jenes Teillexikons einer Sprache, welches der Erzeugung der jeweiligen Inputtexte zugrunde liegt. Auf diese Weise lassen sich die Texte des Eingabekorpus nicht nur mehrschichtig (auf Wort-, Satz- oder Textebene) hypertextuell traversieren. Vielmehr können die Leser einer *Wikidition* das jeweilige Autorenvokabular bis hinab zur Ebene einzelner syntaktischer Wörter analysieren (siehe Abbildung 2). Während die Linkifizierungskomponente an Prinzipien von *Wikipedia* bzw. *WikiSource* anknüpft, ist es das *Wiktionary*-Projekt, an dem sich die Lexikonisierungskomponente orientiert. *Wikidition* verwendet zahlreiche computerlinguistische Methoden zur Gewährleistung von Interoperabilität, Offenheit und Erweiterbarkeit der resultierenden Editionen nach dem Wiki-Prinzip. Auf diese Weise wird die Dissemination computerbasierter Methoden über die *Digital Humanities* hinaus unterstützt, indem es Geisteswissenschaftlern ermöglicht wird, eigene explorative Analysen durchzuführen, und zwar ohne Implementierungsaufwand.

	$\{\text{Echo Text}_i \mid i = 1\}$	$\{\text{Echo Text}_i \mid i > 1\}$	$\{\text{Echo Text}_i \mid i \gg 1\}$
$\{\text{Matrix Text}_i \mid i = 1\}$	1	2	3
$\{\text{Matrix Text}_i \mid i > 1\}$	4	5	6
$\{\text{Matrix Text}_i \mid i \gg 1\}$	7	8	9

Abb. 1: Transbiblionome Szenarien, auf welche *Wikidition* zielt. Beispiele: (1) Kafkas Bericht für eine Akademie (als Matrixtext) versus Hauffs Der Affe als Mensch (als Echotext); (2) Kafkas Bericht für eine Akademie versus alle „Affentexte“ (R. Borgards) seit Ende des 18. Jahrhunderts (Hauff, Hoffmann, Flaubert, …); (3) Kafkas Beim Bau der Chinesischen Mauer versus Prager Tagblatt vom 08.1914–03.1917; (4) Kafkas Werk versus ein Einzeltext aus dem Werk Nietzsches; (5) eine Werkauswahl Kafkas versus eine Werkauswahl Nietzsches; (6) Kafkas Werk versus ein Zeitungskorpus (z. B. basierend auf dem Prager Tagblatt); (7) sämtliche Werke einer Reihe von Schriftstellern versus ein Einzeltext (z. B. Goethes Faust); (8) sämtliche Werke einer Reihe von Schriftstellern versus ein Korpus von Faust-Texten; (9) sämtliche Werke einer Reihe deutschsprachiger Schriftsteller versus sämtliche Werke einer Reihe englischsprachiger Schriftsteller. Anstelle von Matrix- und Echotexten (Wagner 2015) kann alternativ von Hyper- und Hypotexten (Genette 1997) gesprochen werden. In ersterem Fall liegt eine rezeptionsorientierte Sichtweise zugrunde.

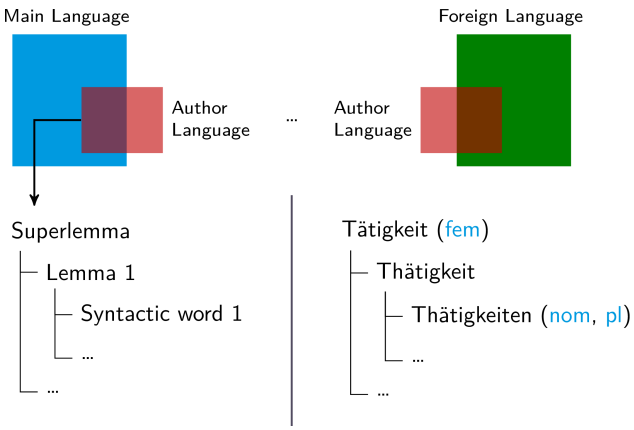


Abb. 2: Linke Bildseite: schematischer Ausschnitt (rot) des autorenspezifischen Vokabulars im Schnittmengenbereich der zugrundeliegenden Matrixsprache (blau). Beispielhafte Fragestellung: *Welches Deutsch verwendet Franz Kafka in seinem Werk und worin weicht sein Deutsch von der jeweiligen Referenzsprache ab?* Da Autoren mehrsprachige Texte verfassen können, ist diese Betrachtungsweise auf mehrere Bezugssprachen auszudehnen (rechte Bildseite). Hierzu unterscheidet *Wikidition* drei Ebenen der

lexikalischen Resolution: die Ebene der Superlemmata, der Lemmata und der syntaktischen Wörter (Wortformen plus grammatische Merkmale) (untere Bildhälfte).

Die Basis für die Vernetzung von Texteinheiten bildet das Konzept der Intertextualität einerseits bzw. der intratextuellen Kohärenz andererseits. Hierzu werden je Sprachebene (*Wort*, *Satz*, *Paragraph*, *Text*) zwei Arten von Relationen unterschieden: syntagmatische Relationen, welche auf Kontiguitätsrelationen von Texteinheiten beruhen, sowie paradigmatische Relationen, welche auf Relationen der Similarität bzw. Substituierbarkeit von Texteinheiten basieren. Dieses Modell orientiert sich an Hjelmslevs (1969) Konzept der sprachlichen Relation, um eine linguistische Basis für die automatisch zu ermittelnden Vernetzungsrelationen zu gewinnen. So werden beispielsweise alle Sätze des Inputkorpus danach paarweise analysiert ob sie (1) syntaktisch ähnlich konstruiert sind (syntagmatische Ähnlichkeit[2]) oder (2) lexikalisch ähnlich aufgebaut sind (paradigmatische Ähnlichkeit[3]). Ausgehend von einem zweidimensionalen Koordinatensystem, in dem syntagmatische und paradigmatische Ähnlichkeit die Dimensionen aufspannen, kann auf dieser Grundlage für alle texttechnologischen Satzähnlichkeitsmaße theoretisch bestimmt werden, wo sie innerhalb dieses Bezugssystems angesiedelt sind. Mit Hilfe von *Wikidition* soll auf diese Weise zugleich eine theoretische und praktische Basis der Text- bzw. Satzähnlichkeitsmessung gelegt werden, die es Usern erlaubt, den Ergebnissen der von *Wikidition* vorgelegten Ähnlichkeitsmessungen explorativ nachzuspüren. Gleichzeitig wird so eine Basis für die Analyse von Satzähnlichkeitsphänomenen im Hinblick auf Zitate gelegt, wie sie insbesondere für das Werk von Karl Kraus maßgeblich sind.

Analog zu diesem Ansatz wird auf Wortebene zwischen der (syntagmatischen) Tendenz von Wörtern zur Kookkurrenz einerseits und der (paradigmatischen) Tendenz zum Vorkommen in ähnlichen Kontexten andererseits unterschieden. Mittels texttechnologischer Erweiterungen der MediaWiki-Software werden zudem sämtliche dieser Verweisbeziehungen interaktiv visualisiert, so dass Nutzer beispielsweise in Echotexten jenen Satzkandidaten nachspüren können, welche dem Autor des jeweiligen Matrixtexts als Vorlage gedient haben könnten (siehe Abbildung 3).

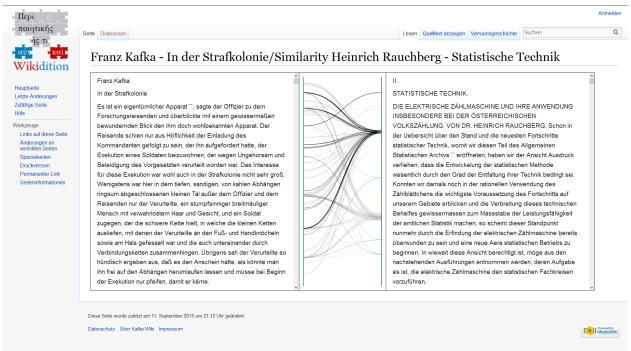


Abb. 3: Franz Kafkas *In der Strafkolonie* (als Matrixtext) im Verhältnis zu Heinrich Rauchbergs *Statistische Technik* (als Echotext). In dem bipartiten Graphen (Bildmitte) verbinden Kanten solche Sätze, die als *Text-Reuse*-Kandidaten ermittelt wurden. Nutzer können per *Fish-Eye*-View den bipartiten Graphen traversieren und Verweisbeziehungen einzelner Sätze, welche in dem links- bzw. rechtsseitig dargestellten Text entsprechend hervorgehoben werden, analysieren.

Der Beitrag exemplifiziert *Wikidition* anhand ausgewählter Texte von Franz Kafka einerseits und von Karl Kraus andererseits. Er ergänzt die beiden literaturwissenschaftlichen Vorträge der Sektion, indem er das texttechnologische Instrumentarium einer transbiblionomen Literaturwissenschaft entwirft und experimentell erprobt. Insbesondere die Verweisstruktur der resultierenden *Wikiditionen* wird auf der Grundlage einer größeren Zahl von Netzwerkcharakteristika quantitativ evaluiert. Diese Evaluation zeigt, dass *Wikiditionen* jeweils auf der Basis einer sehr kleinen Zahl von Verweisen, die theoretisch möglich sind, hochgradig clusternde Graphen erzeugen, die es ihren Lesern erlauben, dasselbe Textkorpus in jedwede Richtung und ausgehend von jedwedem (Wort-, Satz- oder Text-)Kontext zu traversieren. Auf diese Weise wird eine Brücke zwischen Literaturwissenschaft und Informatik gespannt, und zwar am Beispiel der Analyse und Synthese intertextueller Beziehungen in digitalen Korpora literarischer Texte.

Notes

1. Kurznotation für Die Fackel Nr. 366, S. 30
2. Hier sprechen wir auch von syntagmatischer Kontiguität der betroffenen Sätze in dem Sinne, dass ihre Konstituenten füreinander austauschbar sind, ohne paradigmatisch oder inhaltlich verwandt sein zu müssen
3. Hier sprechen wir von paradigmatischer Similarität der betroffenen Sätze in dem Sinne, dass ihre Konstituenten füreinander austauschbar sind, ohne dass die Sätze syntaktisch ähnlich zu sein brauchen. Im Extremfall zweier identischer Sätze gehen syntagmatische und paradigmatische Ähnlichkeit zusammen.

Bibliographie

AAC – Austrian Academy Corpus: *AAC-Fackel*. Online Version: 'Die Fackel. Herausgeber: Karl Kraus, Wien 1899–1936.' AAC Digital Edition No. 1. http://www.aac.ac.at/fackel [letzter Zugriff 15. Oktober 2015].

AAC-Austrian Academy Corpus und Brenner-Archiv: *Brenner Online*. Online Version: 'Der Brenner. Herausgeber: Ludwig Ficker, Innsbruck 1910–1954.' AAC Digital Edition No.2. http://www.aac.ac.at/brenner [letzter Zugriff 15. Oktober 2015].

Baßler, Moritz (2005): *Die kulturpoetische Funktion und das Archiv.* Eine literaturwissenschaftliche Text-Kontext-Theorie. Tübingen: Francke.

Borgards, Roland (2012): „Der Affe als Mensch und der Europäer als Ureinwohner“, in: Wellbery, David E. (ed.): *Kultur-Schreiben als romantisches Projekt.* Romantische Ethnographie im Spannungsfeld zwischen Imagination und Wissenschaft. Tübingen: Königshausen & Neumann 17-42.

Derrida, Jacques (1967): *Grammatologie*. Frankfurt a.M.: Suhrkamp.

Derrida, Jacques (1984): “Two words for Joyce”, in: Attridge, Derek / Ferrer, Daniel (eds.): *Poststructuralist Joyce.* Essays from the French. Cambridge: CUP 145-159.

Embach, Michael / Rapp, Andrea (05.08.2008): „Die intelligentere Expansion der Gutenberg-Galaxis“, in: *Frankfurter Allgemeine Zeitung*.

Genette, Gérard (1993): *Palimpseste.* Die Literatur auf zweiter Stufe. Frankfurt am Main: Suhrkamp.

Hjelmslev, Louis (1969): *Prolegomena to a Theory of Language*. Madison: University of Wisconsin Press.

Kraus, Karl (1926): *Die letzten Tage der Menschheit*. Wien: Verlag Die Fackel.

Krolop, Kurt (1992): "'Solche Erfolche erreichen nur deutsche Molche'", in: Krolop, Kurt: *Sprachsatire als Zeitsatire bei Karl Kraus*. Neun Studien. Berlin: Akademiverlag 305-306.

Meister, Jan Christoph (2005): „Projekt *Computerphilologie*. Über Geschichte, Verfahren und Theorie rechnergestützter Literaturwissenschaft“, in: Segeberg, Harro / Winko, Simone (eds.): *Literarität und Digitalität.* Zur Zukunft der Literatur. München: Fink 315-341.

Pêcheux, Michel (1983): „Über die Rolle des Gedächtnisses als interdiskursives Material“, in: Geier, Manfred / Woetzel, Harold (eds.): *Das Subjekt des Diskurses.* Beiträge zur sprachlichen Bildung von Subjektivität. Berlin: Argument-Verlag 50-58.

Rayward, Warden Boyd (2008): *European modernism and the information society.* Informing the present, understanding the past. Aldershot: Ashgate.

Topia, André (1984): "The Matrix and the Echo: Intertextuality in Ulysses", in: Attridge, Derek / Ferrer, Daniel (eds.): *Post-Structuralist Joyce.* Essays From the French. Cambridge: Cambridge UP 103-125.

von Hofmannsthal, Hugo (2000): *Der Brief des Lord Chandos*. Schriften zur Literatur, Kultur und Geschichte, hg. v. Mathias Mayer. Stuttgart: Reclam.

Vorträge

ePoetics – Korpuserschließung und Visualisierung deutschsprachiger Poetiken (1770-1960) für den ‚Algorithmic Criticism

Alscher, Stefan
stefan.alscher@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Bender, Michael
mbender@linglit.tu-darmstadt.de
Technische Universität Darmstadt

John, Markus
Markus.John@vis.uni-stuttgart.de
Universität Stuttgart, Deutschland

Müller, Andreas
A.Mueller3@gmx.net
Universität Stuttgart, Deutschland

Richter, Sandra
sandra.richter@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Rapp, Andrea
rapp@linglit.tu-darmstadt.de
Technische Universität Darmstadt

Ertl, Thomas
thomas.ertl@vis.uni-stuttgart.de
Universität Stuttgart, Deutschland

Koch, Steffen
Steffen.Koch@vis.uni-stuttgart.de
Universität Stuttgart, Deutschland

Kuhn, Jonas
jonas.kuhn@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

ePoetics ist ein Forschungskooperationsprojekt der Universität Stuttgart und der Technischen Universität Darmstadt. Gefördert vom Bundesministerium für Bildung und Forschung zielt es gleichermaßen auf einen Erkenntnisgewinn für die Informatik sowie die Sprach- und Literaturwissenschaft dank einer wechselseitigen Anregung und Ergänzung im Sinne des ‚Algorithmic Criticism' nach Stephen Ramsay (Ramsay 2007). Dieser Ansatz ist explizit nicht darauf ausgerichtet, lediglich hermeneutische Hypothesen mit algorithmischen Verfahren zu überprüfen. Vielmehr zielt er darauf, durch den iterativen Einsatz analoger und digitaler Methoden verschiedene Perspektiven auf Texte einnehmen und abgleichen zu können. Darüber hinaus ist ein zentraler Aspekt dieses Forschungsparadigmas, Erschließungsentscheidungen und -verfahren sowie Analyseschritte transparent bzw. nachvollziehbar und nachnutzbar zu machen. Das Projekt ePoetics ist der Digitalisierung, Annotation, Analyse und Visualisierung eines für die Geisteswissenschaften zentralen Textkorpus' gewidmet: Poetiken und Ästhetiken von 1770 bis 1960. Diese Texte dokumentieren das Denken und Schreiben über Literatur und andere Künste in der zentralen Periode nach der Abkehr von der Normen- und Regelpoetik (vor 1770) und vor dem Übergang zur Literaturtheorie und damit dem Ende der Poetik als literaturwissenschaftlicher Textgattung (nach 1960). Sie enthalten dabei grundlegendes Wissen über Sprache und Literatur (-wissenschaft), etwa die Erläuterungen zentraler Begriffe und deren Zusammenhänge. ePoetics betreibt die Entwicklung und Untersuchung eines Testkorpus' von zwanzig Poetiken, ausgewählt aus einem Gesamtkorpus von 1240 Texten (inkl. aller Auflagen), die Sandra Richter in ihrer Studie ‚A history of Poetics' (Richter 2010) als zur Gattung ‚Deutschsprachiger Poetik' zählbarer Werke bibliographiert hat. Die Auswahl des Testkorpus' enthält – historisch und systematisch betrachtet – die repräsentativsten Texte des Gesamtkorpus', d. h. die, die am häufigsten zitiert und in den meisten Auflagen herausgegeben wurden, und stellt dennoch auf den ersten Blick ein sehr heterogenes Korpus dar. Aus sprach- und literaturwissenschaftlicher Sich zeigen wir auf, wie sich diese Heterogenität im Einzelnen darstellt, aber auch, welche tiefergehenden Gemeinsamkeiten und Abhängigkeiten die Texte auf den zweiten Blick aufweisen und auf welche Ursprünge sich diese zurückführen lassen. Für ausführlichere Informationen zum ausgewählten Textkorpus und zum Projekt insgesamt besuchen Sie unsere Homepage (vgl.).

Im Zentrum unseres Interesses steht aktuell beispielsweise der Begriff der Metapher als ein zentrales sprach- und literaturwissenschaftliches Konzept, das in unserem Textkorpus verhandelt wird. Die mit diesem zusammenhängenden Fragen lauten: Wie wird der Begriff in einzelnen Poetiken verstanden und erklärt? Wie ändert sich dieses Verständnis innerhalb unseres Testkorpus'? Welche literarischen oder theoretischen Werke werden im Zusammenhang damit genannt oder zitiert? Wie verändert sich der ‚Kanon' dieser Werke? Verändern sich die Zusammenhänge, in denen die Werke zitiert werden? Und

schließlich: Wie verändert sich insgesamt der Umgang mit Zitaten und deren Nachweisen?

Problemstellungen für die digitale Annotation mit dem Ziel der computergestützten Auswertbarkeit liegen bei solchen Texten und Anforderungen auf mehreren Ebenen vor: Das jeweilige Metaphernverständnis muss differenziert erschlossen und die Komponenten der Begriffsbestimmung müssen trennscharf kategorisiert werden können. Beispiele aus der Primärliteratur müssen eindeutig erkannt und den jeweiligen theoretischen Aspekten, für die sie stehen, zugeordnet werden. Und schließlich müssen die Textebenen und Referenzstrukturen der Poetik explizit gemacht werden – also wo der Autor selbst theoretisiert, wo zitiert oder paraphrasiert wird, inwiefern dies kenntlich gemacht wird oder nicht und sogar, wo bei Zitaten vom ursprünglichen Text abgewichen wird. Dies wird durch die Annotation nach einem komplexen Schema umgesetzt. Die Annotationen werden einerseits in TEI-konformen XML-Dateien publiziert, andererseits aber auch als Grundlage von computergestützten Analysen und Visualisierungen genutzt. Abbildung 1 veranschaulicht das Vorgehen im Projekt ePoetics im Sinne eines ‚Algorithmic Criticism' nach Stephen Ramsay (2007).

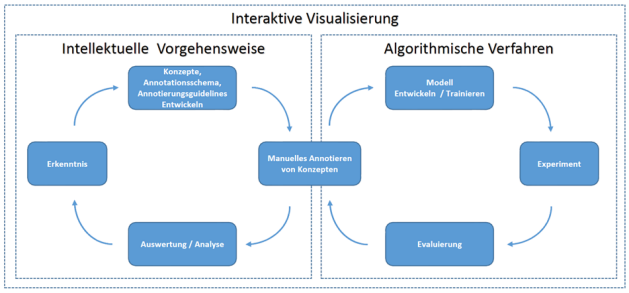


Abb. 1: Intellektuelle Vorgehensweise und die algorithmischen Verfahren in ePoetics, in Anlehnung an Kuhn und Reiter (2015).

Intellektuelle Vorgehensweise (Sprach- und Literaturwissenschaft)

Die Texte des Testkorpus' stehen als Image-Digitalisate und als nach dem ‚Double Keying'-Verfahren transkribierte und aufbereitete digitale Volltexte zur Verfügung. Die strukturellen (und auch die semantischen) Annotationen des Korpus' erfolgen nach den Konventionen der Text Encoding Initiative (TEI). Das Korpus wird in virtuelle Forschungs-Infrastrukturen wie TextGrid und das Deutsche Textarchiv (DTA) integriert und dort mit den vorhandenen Referenztexten verlinkt.

Nach der Identifikation relevanter und interessanter Begriffe und Konzepte wurden zu einzelnen ausgesuchten Begriffen wie der Metapher mithilfe des UAM CorpusTool Annotationsschemata für manuelle Annotationen erstellt. Diese wurden unter ausführlicher Dokumentation von Annotationsguidelines durch mehrere Annotatoren getestet, kontinuierlich verbessert, ausgebaut und schließlich in den Poetiken durchgeführt.

Abbildung 2 zeigt eine vereinfache Version des daraus hervorgegangenen Annotationsschemas, das sich in zwei Teilbereiche gliedern lässt, die teils direkt und teils mit leichten Veränderungen auch auf andere Begriffe übertragen werden können. Das Schema resultiert aus den oben genannten sprach- und literaturwissenschaftlichen Fragen, die sich in die Aspekte der Repräsentation und des Verständnisses bzw. der Anwendung des Metaphernbegriffs in den Poetiken aufteilen lassen.

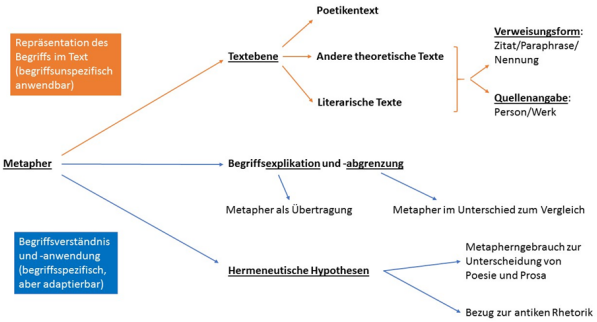


Abb. 2: Vereinfachte Darstellung des Annotationsschemas zur Metapher: Erkennbar ist der begriffsunspezifische (oben, orange) und -spezifische Bereich (unten, blau). Während sich der Teil des Schemas, der die Repräsentation des Begriffs im Text abbildet, sofort auf andere Begriffe anwenden lässt, ist der Teil, der sich dem Begriffsverständnis und der -anwendung widmet, begriffsspezifisch. D. h. die hier zu annotierenden Kategorien lassen sich nicht für andere Begriffe verwenden, aber leicht durch passende für den jeweiligen Begriff ersetzen.

Das Annotationsschema stellt eine Systematisierung des Begriffs, d. h. seines Vorkommens und Verständnisses in den Poetiken dar. Die für den Begriff relevanten Textstellen werden zunächst dahingehend klassifiziert, ob es sich um Poetikentext handelt (also Text vom Autor der Poetik selbst), oder ob andere theoretische oder literarische Texte zitiert, paraphrasiert oder genannt werden. Neben den Verweisungsformen annotieren wir hierbei auch die Quellenangaben – beides im Übrigen nicht nur, wenn es explizit angegeben ist. So berücksichtigen wir auch die Möglichkeit von „versteckten" Zitaten oder solchen, bei denen die Quelle nicht oder unvollständig benannt ist. Das Auffinden bestimmter Muster sowie zum Beispiel Titel und Personennamen oder Zitate wird dabei unterstützt durch computerlinguistische Methoden und Verfahren der interaktiven Visualisierung. Darüber hinaus systematisieren wir das vorliegende Begriffsverständnis, d. h. ob die Metapher z. B. als Übertragung erklärt wird, und grenzen sie von anderen Begriffen ab, z. B. im Unterschied zum Vergleich. Zusätzlich lassen sich auch Beobachtungen zu konkreten hermeneutischen Hypothesen annotieren, z. B. ob anhand

des Metapherngebrauchs zwischen poetischer und prosaischer Sprache unterschieden wird.

Schon durch die Annotation von implizitem Wissen entsteht somit bereits bei den manuellen Annotationen eine Metaebene an Informationen, mit der der digitalisierte Poetikentext angereichert wird. Die Systematisierung erfordert eine andere Herangehensweise an den Gegenstand, als es bei einer rein hermeneutischen Analyse der Fall wäre. Ebenso führt diese zwangsläufig zur Problematisierung der Systematisierungs(un)möglichkeit eines per se komplexen, weil heterogenen Untersuchungsgegenstandes. Das Ziel der algorithmischen Weiterverarbeitung wird zum Paradigma für die systematisch-kategorisierende Ausdifferenzierung von theoretischen Begriffen, wobei diesbzgl. neue Erkenntnisse, aber auch Grenzen aufgezeigt werden können. Die Operationalisierung der Daten führt so bereits zu Erkenntnissen, bevor computertechnologische Auswertungen durchgeführt werden, womit sie sich über den Status bloßer Vorverarbeitung erheben und einen Eigenwert besitzen.

Algorithmische Verfahren (Computerlinguistik)

Mit algorithmischen Verfahren können aus kleinen Mengen annotierter Daten (aus der manuellen und damit zeitaufwendigen Annotation) große Mengen gemacht werden, indem die annotierten Arten von Informationen automatisch auf größere Datenmengen übertragen werden.

Im Folgenden wird anhand eines Beispiels in Anlehnung an den rechten Teil von Abbildung 1 beschrieben, wie die manuelle Annotation, das Training von Klassifikationsmodellen und die Analyse der Klassifikationsergebnisse ineinander greifen. Zur Klassifizierung von Text zwischen Anführungszeichen als eine der drei Klassen ‚Hervorhebung' (Wörter deren Bedeutung hervorgehoben wird), ‚Titel' (Werktitel) und ‚Zitat' (Zitate aus anderen Werken) wurde manuell ein Korpus annotiert, in dem jeder Text zwischen Anführungszeichen einer dieser drei Klassen zugewiesen wurde (Manuelles Annotieren von Konzepten). Auf der Basis dieses Korpus wurden Klassifikationsmodelle zur automatischen Erkennung dieser drei Klassen trainiert (Modell trainieren / entwickeln). Die automatischen Modelle wiederum wurden benutzt, um in anderen Poetiken Text in Anführungszeichen automatisch in diese drei Klassen einzuteilen. Es wurde durch Stichproben und formale Evaluation auf einem für diesen Zweck annotierten separaten Korpus erkannt, dass die Klassifikation gut funktioniert (Evaluierung). Da so unter anderem direkte Zitate und Werktitel automatisch erkannt werden, ermöglicht dieser Schritt wiederum die automatische Verlinkung von Werktiteln und Zitaten mit ihren Einträgen (sofern vorhanden) im TextGridRepository-Korpus (Evaluierung). Durch diese Information kann vom Analysten manuell die Verteilung von Werken und Zitaten in den Poetiken untersucht und bedeutende Werke / Zitate erkannt werden. Diese Erkenntnisse können dann wiederum als Metadaten im Dokument annotiert werden (Manuelles Annotieren von Konzepten und Metadaten).

Interaktive Visualisierung

Interaktive Visualisierung spielt eine wesentliche Rolle in der Vorgehensweise von ePoetics, siehe Abbildung 1, da sie eine zusätzliche Interaktion zwischen Forschern und den Untersuchungsgegenständen ermöglicht. Zum einen können interaktive Systeme die hermeneutischen Vorgehensweise unterstützen, indem sie den Geisteswissenschaftlern die Möglichkeiten bieten, Annotationsschemata und -guidelines zu entwerfen, Konzepte und Metadaten in Texten manuell zu annotieren sowie diese Ergebnisse zu analysieren und darzustellen. Zum anderen kann die computerlinguistische Vorgehensweise unterstützt werden, so dass Forscher Einfluss auf komplexe Prozesse nehmen können wie beispielsweise dem Trainieren maschineller Lernmethoden durch visuelle Veränderungsparameter. Durch diese Art der Interaktion kann unterstützt werden, dass Modelle mit Hilfe des Experten entwickelt, angepasst, trainiert sowie die Ergebnisse evaluiert werden können. Um diese Herausforderungen umzusetzen, wurden zwei interaktive visuelle Analysewerkzeuge konzipiert und entwickelt. Der VarifocalReader (Ertl et al. 2014), der auf einem hierarchischen Navigationskonzept basiert (Wörner / Ertl 2013), ermöglicht den Anwendern einen direkten Zugang zu Details und Dokumentquellen, während sie auf unterschiedlichen Abstraktionsebenen mit Zusammenfassungen vorhandener Annotationen interagieren können. Des Weiteren bietet das System die Möglichkeit, computerlinguistische Modelle anzupassen bzw. zu trainieren sowie Metadaten zu analysieren, zu annotieren und zu korrigieren. Eine beispielhafte Analyse ist in Abbildung 3 dargestellt, in der der Forscher einen schnellen Überblick und Zugang zur ausgewählten Annotation „Wallenstein" (in der 3. Word Cloud sichtbar) erhält.

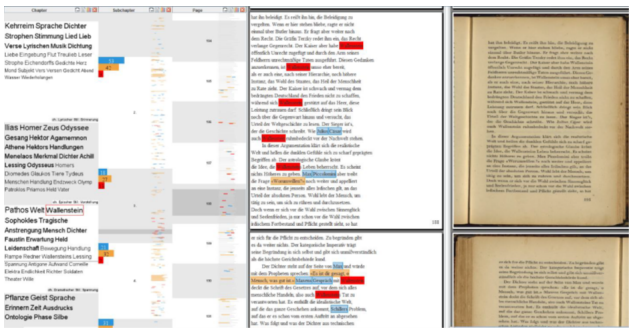


Abb. 3: Emil Staigers "Grundbegriffe der Poetik" unterteilt (von links nach rechts) in unterschiedliche Ebenen. Kapitel (mit Word Clouds), Unterkapitel (mit Balkendiagrammen und Piktogrammen), Seiten (mit Piktogrammen), Textzeilen und gescannte Digitalisate der aktuellen Seite.

Der zweite Ansatz (Heimerl et al. 2014) wurde konzipiert, um eine textvergleichende Analyse zu

ermöglichen (siehe Abbildung 4). Die Visualisierung bietet einen Vergleich von mehreren Dokumenten auf einer abstrakten Ebene in Bezug auf die Verteilung der Annotationen, während die Textfelder eine flexible Navigation durch die einzelnen Texte ermöglichen. Zusätzlich unterstützt dieser focus+context Ansatz einen reibungslosen Übergang zwischen close und distant reading.

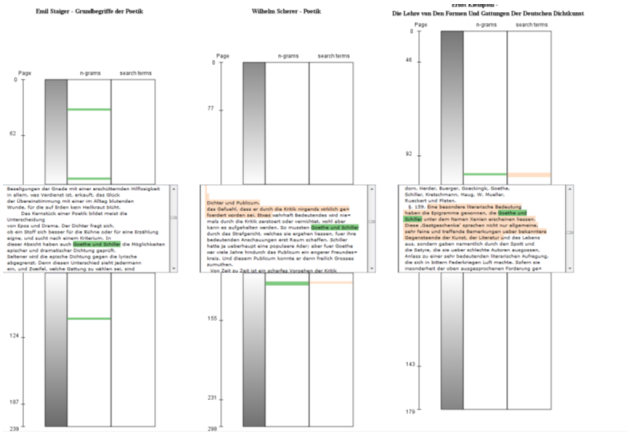


Abb. 4: In dieser Abbildung sind drei ausgewählte Texte nebeneinander dargestellt. Jedes dieser Dokumente verfügt über Seitenangabenskala (linke Seite) und jeweils zwei Bänder, die zum einen Annotation darstellen (grüne Balken) und zu anderen Suchergebnisse (orangene Balken). Der Analyst kann durch die einzelnen Dokumente navigieren (Textboxscrollleiste) und per Mausklick zwischen den einzelnen Annotationen (Balken) springen.

Conclusion

Ergebnis des Projekts ePoetics ist ein digitalisiertes und annotiertes Korpus poetologischer Texte (TEI-konform und nachnutzbar), in denen zentrale Konzepte der Sprach- und Literaturtheorie durch XML-Auszeichnung explizit gemacht und systematisiert werden. Durch Korpus-übergreifende Analysen dieser Auszeichnungen können Gemeinsamkeiten und Unterschiede sowie diachrone Entwicklungen gezeigt werden. Darüber hinaus werden die Referenz- und Diskursstrukturen erschlossen (auch implizite, „versteckte“ Verweisungen), die auf verschiedenen Ebenen der Texte bestehen – einerseits Verweisungen auf andere Poetiken sowie die Identifikation bestimmter Denkschulen bzw. Theorielinien, die bis auf Ansätze aus der Antike zurückgehen (z. B. Aristoteles, Quintilian), andererseits die Diskussion von literarischen Beispielen, die Rückschlüsse auf die Entwicklungen des Literaturkanons erlauben. Die manuellen Annotationen werden iterativ gestützt durch automatisierte Methoden und Verfahren der interaktiven Visualisierung. Die dabei entwickelten computerlinguistischen Anwendungen und Visualisierungssysteme (siehe Abbildungen 3 und 4) stellen ebenfalls Ergebnisse des Projekts dar.

Bibliographie

Ertl, Thomas / Wörner, Michael (2013): “Smoothscroll. A multi-scale, multi-layer slider”, in: *Computer Vision, Imaging and Computer Graphics - Theory and Applications* 274: 142–154.

Ertl, Thomas / John, Markus / Koch, Steffen / Wörner, Michael (2014): “VarifocalReader – In-Depth Visual Analysis of Large Text Documents”, in: *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 20, 12: 1723-1732.

Ertl, Thomas / Kuhn, Jonas / Richter, Sandra / Alscher, Stefan / Rapp, Andrea (2013-2016): *ePoetics*. Universität Stuttgart [letzter Zugriff 03. Februar 2016].

Heimerl, Florian / John, Markus / Koch, Steffen / Müller, Andreas (2014): “A Visual Focus+Context Approach for Text Comparison Tasks”, in: *VisLR Workshop, LREC 2014*.

Kuhn, Jonas / Reiter, Nils (2015): “A plea for a method-driven agenda in the Digital Humanities”, in: *Proceedings of the Digital Humanities Conference, Sydney, Australia 2015*.

Ramsay, Stephen (2007): “Algorithmic Criticism”, in: Schreibman, Susan / Siemens, Ray (eds.): *A Companion to Digital Literary Studies*. Malden, MA: Blackwell 477-491.

Richter, Sandra (2010): *A History of Poetics*. German Scholarly Aesthetics and Poetics in International Context, 1770-1960. With Bibliographies by Anja Zenk, Jasmin Azazmah, Eva Jost and Sandra Richter. Berlin / New York: de Gruyter.

Editionsphilologie zwischen Bibliothek, Archiv und Fachwissenschaft: Der standardisierte Open-Source-Workflow der digitalen Edition der Korrespondenz August Wilhelm Schlegels

Bamberg, Claudia
claudia.bamberg@staff.uni-marburg.de
Philipps-Universität Marburg, Deutschland

Jochen, Strobel
jochen.strobel@staff.uni-marburg.de
Philipps-Universität Marburg, Deutschland

Im prädigitalen Zeitalter war der Herausgeber einer wissenschaftlichen (Buch-)Edition nicht nur verantwortlich für die Textkonstitution und für die Kommentierung des edierten Textes, sondern er war es auch für die Materialrecherche, -beschaffung und für die Autopsie der Handschriften. Zugleich entwickelt jede Buchedition im Hinblick auf die Aufarbeitung und Präsentation der Materialien ihre je eigenen Kriterien. Auch wenn sich in den letzten Jahrzehnten in der Editionsphilologie im Hinblick auf die Qualität von Editionen ein wissenschaftlicher Standard etabliert hat, weicht die Präsentation der edierten Quellen in Buchform zumeist stark voneinander ab.

Mit der Möglichkeit, Quellen digital aufarbeiten und präsentieren zu können, hat sich der sehr aufwändige Arbeitsprozess des Erschließens und Edierens bekanntlich grundlegend gewandelt (Sahle 2013). So sieht etwa in einer digitalen Edition nicht nur der Bearbeiter das Material ein, da es nun in digitalisierter Form mitgeliefert wird und vom Nutzer problemlos eingesehen werden kann. Vor allem trägt der Wunsch nach Standardisierungen bereits Früchte: Die meisten Netz-Editionen beziehen sich inzwischen auf die Gemeinsame Normdatei der Deutschen Nationalbibliothek (GND) sowie weitere einschlägige Datenbanken und Dateiformate (VIAF, Beacon etc.) und arbeiten mit der Standardauszeichnungssprache XML/TEI (Hochstrasser 2014).

Was allerdings noch immer fehlt, ist ein allgemeinverbindlicher, d. h. standardisierter und nachnutzbarer Editionsworkflow, der für die einzelnen Schritte beim Erarbeiten einer digitalen Edition von der Aufnahme der Handschriftendigitalisate und deren normierten Metadaten über die Transkription und Präsentation der Quellen bis hin zur Langzeitarchivierung genutzt wird. Noch immer liefern viele Archive ihre Digitalisate auf unterschiedlichen Medienträgern mit teils stark voneinander abweichenden Qualitätskriterien bzw. senden sie per E-Mail an den Editor, der sie in Auftrag gegeben hat. Das ist für beide Seiten umständlich, zumal so zunächst keine Möglichkeit der Nachnutzbarkeit oder Weiterverarbeitung existiert – sowohl was die Digitalisate als auch was die Metadaten angeht. Auch die vom Philologen korrigierten Metadaten fließen i.d.R. nicht in die Archive und die zentralen Datenbanken wie *Kalliope* zurück – ein eklatanter Mangel, der für alle Beteiligten, auch für den Nutzer, der nach aktuellen und verlässlichen Informationen sucht, von großem Nachteil ist.

Das in Dresden, Marburg und Trier angesiedelte DFG-Projekt „Digitale Edition der Korrespondenz August Wilhelm Schlegels“ () hat in den letzten Jahren ein beispielhaftes Modell der Datenverarbeitung entwickelt, das diese Lücke schließen will, indem es eine standardisierte Infrastruktur für digitale Editionen bereitstellt. In der zweiten Projektphase, die nach der Bewilligung der DFG am 1. Oktober 2015 beginnt und Ende 2018 endet, soll diese Infrastruktur vollständig eingerichtet werden. Anders gesagt: Der Open-Source-Editionsworkflow des Schlegel-Projekts, der in weiten Teilen bereits funktioniert und im Vortrag im Einzelnen vorgestellt werden soll, möchte beispielgebend sein. Er will zukünftigen Projekten als vollständig freie, quelloffene und einfach nachnutzbare technisch-organisatorische Lösung zur Nachnutzung zur Verfügung stehen. Das bedeutet auch, dass die erhobenen standardisierten und, wo nötig, korrigierten Briefdaten sofort frei zugänglich sind und nicht nur in der eigenen Editionssoftware verzeichnet werden, sondern auch an die zentralen Datenbanken und an die Archive zurückgespielt werden. Der Nutzer kann somit sofort den neuesten editorischen Stand abrufen.

Dafür ist eine enge und kontinuierliche Zusammenarbeit zwischen der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB), dem Trier Center for Digital Humanities (TCDH), der Universität Marburg und den Partnerinstitutionen – den zahlreichen, über hundert Archiven – erforderlich. Die Werkzeuge, die in diesem Workflow zum Einsatz kommen, weiterentwickelt werden und ineinander spielen, sind die Digitalisierungssoftware *Goobi* (Bonte) sowie die in Trier entwickelte virtuelle Forschungsumgebung und Editionsplattform *Forschungsnetzwerk und Datenbanksystem* FuD (FuD 2014 ; Bamberg / Burch 2014) *Goobi* wird inzwischen von rund 50 Anwendern genutzt, mit *FuD* arbeiten zahlreiche vom TCDH betreute Editionsprojekte. Über Standardschnittstellen können die Metadaten mitsamt Image-Digitalisaten der Handschriften automatisch abgerufen und übertragen werden.

Im Schlegel-Projekt hat zunächst der Projektpartner in Dresden, die SLUB, unter der Leitung von Prof. Dr. Thomas Bürger seine Schlegeliana (rund 3.800 Briefe) von der zentralen Autographendatenbank *Kalliope* nach *Goobi*, wo die Image-Digitalisate der Autographen hinzugefügt wurden, und von dort nach *FuD* importiert, wo sie editorisch bearbeitet werden können. Zugleich stehen die Autographen in den eigenen digitalen Sammlungen frei zur Verfügung. Auch die Universitäts- und Landesbibliothek Bonn hat als erster Pilotpartner des Schlegel-Projekts ihre Autographensammlung zu A. W. Schlegel (rund 350 Schreiben) über *Goobi* nach *FuD* transferiert.

Grundlegend für eine gelingende und nachhaltige Infrastrukturbildung ist die normierte Verzeichnung der Metadaten, so dass die einzelnen Systeme diese richtig erkennen und zuordnen und sie außerdem weiterverarbeitet werden können und die Edition mit anderen Projekten vernetzbar ist. Für die Autographen, die nach internationalen Standards (METS / EAD bzw. METS / MODS) in den Katalogen und Datenbanken verzeichnet sind, wird eine Schnittstelle in *FuD* implementiert, so dass eine konsistente Datenerhaltung zwischen den Katalogen und *FuD* – mithin in beide Richtungen – gewährleistet ist. Für jene Briefdaten, die noch nicht auf diese Weise verfügbar sind, wird, nach der Aufnahme und Bearbeitung in *FuD*, ein Exportfilter

entwickelt, so dass sie aus *FuD* nach *Kalliope* und von hier aus in die lokalen Kataloge exportiert werden können (Bamberg / Burch 2014: 293).

Der geplante Vortrag möchte diesen im Schlegel-Projekt entwickelten Open-Source-Workflow anhand einiger Beispiele aus der digitalen Briefedition darstellen und dabei zeigen, dass eine solche Infrastrukturbildung, die Quellen aus den unterschiedlichsten Archiven in einer Plattform einheitlich strukturiert zusammenführt, um sie weiter verfüg- und vernetzbar zu machen, auch eine neue Erschließungstiefe eines umfangreichen Briefkorpus ermöglicht.

Bibliographie

Bamberg, Claudia / Burch, Thomas (2014): „Inventarisieren, Analysieren und Archivieren vernetzt. Digitalisierung und Edition größerer Briefkorpora mit der virtuellen Editionsplattform 'Forschungsnetzwerk und Datenbanksystem' (FuD)“, in: Delf von Wolzogen, Hanna / Falk, Rainer (eds.): *Fontanes Briefe ediert*. Würzburg: Königshausen und Neumann 265–282.

Bonte, Achim (ed.): *Goobi*. SLUB: Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden http://www.goobi.org/ [letzter Zugriff 08. Januar 2016].

FuD (2014): *FuD*. Die virtuelle Forschungsumgebung für die Geisteswissenschaften. Trier Center for Digital Humanities (TCDH) & Forschungszentrum Europa (FZE). Universität Trier http://fud.uni-trier.de/de/ [letzter Zugriff 08. Januar 2016].

Hochstrasser, Daniel (2014): „Anforderungen an digitale Briefeditionen“, in: Delf von Wolzogen, Hanna / Falk, Rainer (eds.): *Fontanes Briefe ediert*. Würzburg: Königshausen und Neumann 266–277.

Sahle, Patrick (2013): *Digitale Editionsformen*. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. 3 Bde. Norderstedt: Books on Demand.

Stadler, Peter (2014): „Interoperabilität von digitalen Briefeditionen“, in: Delf von Wolzogen, Hanna / Falk, Rainer (ed.): *Fontanes Briefe ediert*. Würzburg: Königshausen und Neumann 278–287.

Technical and social Infrastructures for the Humanities: The Example of the Dagaare-English-Cantonese Dictionary

Bodomo, Adams
adams.bodomo@univie.ac.at
Universität Wien, Institut für Afrikawissenschaften; AT

Wandl-Vogt, Eveline
eveline.wandl-vogt@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities; AT

Mörth, Karlheinz
karlheinz.moerth@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities; AT

Introduction

This paper introduces into the transformation process of the Dagaare – Cantonese – English dictionary into an open, online research infrastructure in the framework of European research infrastructures and – in doing so – open those for Non-European researchers, research data as well as topics.

The trilingual dictionary is designed for use in lexicographical and linguistic field methods training. It serves as a database to illustrate many linguistic principles and phenomena in phonology, morphology, syntax and semantics. First and foremost it is intended as a reference source for Chinese and English speaking students. Dagaare is a language spoken in Ghana and Burkina Faso by about two million people. It belongs to the Gur branch of the Niger-Congo family. In spite of the fact that Dagaare is genetically unrelated to Chinese, there are some interesting typological features under which the two languages can be compared. To illustrate, both Dagaare and Chinese are tone languages, unfortunately lacking audio files in the printed dictionary version. But while Chinese has a complex system of four to nine tonemes, Dagaare - like most West African languages - has a two-tone system. The first part of the dictionary includes information of the orthography and sound system of Dagaare followed by an explanation of the verbal and nominal morphology of this language. Part two is the proper dictionary which comprises more than thousand head words and a total of 3,000 to 4,000 words. Subsequent to the lexicon are represented sample field work projects that are intended to aid both the field trainer and trainee. They cover the areas of phonology, morphosyntax, lexical semantics ans sociolinguistics.

The valuable lexicographical data mentioned before were meant to be made sustainable available on the internet. To this end, they had to be transferred into an existing infrastructure, the research infrastructure for lexicography available at the Austrian Centre of Digital Humanities (ACDH) of the Austrian Academy of Sciences. The Academy has a long-standing tradition in eLexicography to which several departments contributed

over more than hundreds of years. Most recently, the ACDH hosts a research group on eLexicography, the lexicography laboratory (1.1.2015-), to support, coordinate and methodically explore experimental scholarship in the fields of lexicography.
The emerging infrastructure is made up of several components: (1) an editor, (2) a formalised encoding framework, (3) a depositing back end and (4) a publishing system, all of which have been integrated into one system. An important keyword in this endeavour has been modularisation, the system not being one single piece of software but a number of complementary components that interlock neatly through clearly defined interfaces.

Workflow

The work of integrating the lexicographical data into the infrastructure was performed by the eLexciography working group of the ACDH. The workflow is a five step procedure:
(1) analysing and discussing the research focus and data structure
(2) converting the data from a simple table into a standards-based XML format (TEI P5)
(3) importing it into the database
(4) manual post processing
(5) and publication on the internet.
At the end of the process the data will be available in a persistent manner.

Editor

The Viennese Lexicographical Editor (VLE) is a fairly new piece of software that first came into existence as a by-product of an entirely different development activity: the creation of an interactive online learning system for university students. Thus, it was first used in a collaborative glossary editing project carried out as part of university language courses at the University of Vienna. As the tool proved to be flexible and adaptable enough, it was also used and further developed in a number of other projects collecting lexical data.

The interface is built around an XML editor that allows to process standard-based lexicographic and terminological data. Basically any XML-based formats such as LMF, TBX, RDF or TEI can be handled. The program provides a number of useful functions to automate editing procedures. It can check the structural integrity (well-formedness) of input on the fly. Technologically, it draws not only on the XML core specification but also on several cognate technologies. XSLT and XPath play an important role both for visualising and modifying existing datasets. Lexicographers can insert elements on the basis of predefined XML schemas. Most of the functions can be applied both to single and multiple lemmas. One of the most recent improvements is a versioning system and an improved working mode that allows lexicographers to work on the XML data without actually seeing the tags. Furthermore, the editor also has a configurable interface enabling lexicographers to access external corpora and to integrate example sentences from them into dictionary entries. The communication between the dictionary client and the server has been implemented as a RESTful web service.

The tool forms part of Austria's contribution to the pan-European CLARIN Research Infrastructure Consortium and is freely available from the ACDH Website .

Formalised encoding framework

While the list of formats used in the lexicographic community is unfortunately very long, there exists a de-facto standard which has been used widely in many digital humanities projects, in numerous lexicographic projects and most of the ACDH's lexicographic endeavours: the Guidelines of the Text Encoding Initiative (TEI). The application of digital (de-facto) standards in building digital language resources is of particular concern when we think about interoperability and re-usability of resources. The buzz-word of open life-cycles for research data will remain meaningless unless researchers succeed in achieving a certain degree of harmonisation in structuring their data and meta data. The ACDH has been working on specialised schemata based on the TEI (P5) dictionary module for quite some time. In all these efforts, they have also aimed at a high degree of interoperability with the ISO standard LMF (Lexical Markup Framework). In order to realise mechanisms for cross-dictionary access, they have also been working with semantic technologies such as RDF and SKOS.

The basic reduced TEI schemata have been documented in form of guidelines which give detailed accounts of how dictionaries in the ACDH collection were encoded. These guidelines document and discuss the schema and furnish a number of examples taken from actual dictionaries. The target group for this guide are both the lexicographers working on ACDH projects as well as others who might want to work along similar lines. These particular Guidelines were themselves produced making use of the TEI framework.

Depositing infrastructure

The dictionary editor is a web-based application that allows lexicographers to work in groups. The data is stored on a server of CLARIN Centre Vienna. Being part of an official infrastructure, long-term availability will be vouchsafed. In addition, the owners of the lexicographical data can draw copies of their dictionaries at any time of the compilation process.

Publishing framework

The publishing infrastructure builds on corpus_shell, a service-oriented architecture and a distributed and heterogeneous virtual landscape. The core functionality of this modular framework is to expose well-defined interfaces based on acknowledged standards. The principle idea behind the architecture is to decouple the modules serving data from the user-interface components. One of the nice features of the system is that you can build new interfaces via XSLT styles almost on the fly.

Status and outlook

The conversion and import of the data has already been undertaken. Dagaare – English – Cantonese Dictionary is already available online. However, the working group is still improving the web-interface, a stable URL will be assigned by the end of 2015.

The Dagaare – English – Cantonese Dictionary is about to be improved from a content as well as collaboration / social infrastructure point of view:

(1) Audio files are to be added to support – mainly – the representation of the tone languages Dagaare and Cantonese. Doing so, we enlarge the network of people participating into the project for both,

(a) free and open Wikimedia audio tools as well as

(b) high performance audio tools e.g. supported by Forschungszentrum Telekommunikation Wien (FTW http://www.ftw.at/) or Phonogrammarchiv at the AAS http://www.phonogrammarchiv.at/ ,

(c) speakers with Cantonese mother tongue. (Bodomo Adams himselves represents Dagaare mother tongue).

(2) The dictionary will be fully embedded into the lexicographical research infrastructure of ACDH as well as the research of the Institut für Afrikawissenschaften at the University of Vienna. This implies both,

(a) experimental development of the dictionary content applying methods of other disciplines e.g. Natural Language Processing and Semantic Technologies for interlinking with other dictionaries, semi-automatic translation into other languages starting with German, connecting with cultural content etc., e.g. interlinking with cultural resources like songs; (b) embedding it into a research framework for African Diaspora studies.
In doing so, the representatives of both institutes open towards collaboration of global communities of several disciplines that are until now not in touch.

Bibliographie

Bodomo, Adams (2004): *Dagaare – Cantonese – English Dictionary for Lexicographical Field Research Training* (= Afrikawissenschaftliche Lehrbücher 14). Köln: Köppe.

Bodomo, Adams / Mora, Manolete (2007): "Documenting Spoken and Sung Texts of the Dagaaba of West Afrika", in: *Empirical Musicology Review* 2, 3: 81-102.

Budin, Gerhard / Majewski, Stefan / Moerth, Karlheinz (2012): "Creating Lexical Resources in TEI P5", in: *Journal of the Text Encoding Initiative* 3 https://jtei.revues.org/522 [letzter Zugriff 08. Februar 2016].

Budin, Gerhard / Moerth, Karlheinz (2011): "Hooking up to the corpus: the Viennese Lexicographic Editor's corpus interface",in: Kosem, Iztok / Kosem, Karmen (eds.): *Electronic lexicography in the 21st century*. New applications for new users. Proceedings of eLex 2011 conference. Bled, Slovenia: Trojina, Institute for Applied Slovene Studies 52-59.

Budin, Gerhard / Moerth, Karlheinz / Durco, Matej (2013): "European Lexicography Infrastructure Components", in: Kosem, Iztok / Kallas, Jelena / Gantar, Polona / Krek, Simon / Langemets, Margit / Tuulik, Maria (eds.): *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference, 17-19 October 2013. Tallin, Estonia: Trojina, Institute for Applied Slovene Studies / Eesti Keele Instituut 76-92.

Declerck, Thierry / Lendvai, Pirsoka / Moerth, Karlheinz (2013): "Collaborative Tools: From Wiktionary to LMF, for Synchronic and Diachronic Language Data", in: Francopoulo, Gil (ed.): *LMF*. Lexical Markup Framework. London / Hoboken: John Wiley & Sons 175-186.

Declerck, Thierry / Moerth, Karlheinz / Wandl-Vogt, Eveline (2014): "A SKOS-based Schema for TEI encoded Dictionaries at ICLTT", in: *LREC 2014, Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland: European Language Resources Association 414-417.

Über den Mehrwert der Vernetzung von OCR-Verfahren zur Erfassung von Texten des 17. Jahrhunderts

Boenig, Matthias
boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften - Berlin, Deutschland

Würzner, Kay-Michael
wuerzner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften - Berlin, Deutschland

Binder, Arne
binder@informatik.hu-berlin.de
Berlin-Brandenburgische Akademie der Wissenschaften - Berlin, Deutschland

Springmann, Uwe
springmann@cis.uni-muenchen.de
Centrum für Informations- und Sprachverarbeitung - Ludwig-Maximilians-Universität München, Deutschland

Einleitung

Dieser Beitrag stellt eine neuartige Methode zur optischen Zeichenerkennung (*Optical Character Recognition,* OCR) speziell für Textvorlagen des 17. Jahrhunderts vor. Anstatt ein neues OCR-Verfahren zu entwickeln, werden zwei etablierte Open-Source-Lösungen genutzt. Die Ausgaben der Programme werden computergestützt kombiniert, um so eine möglichst genaues Textergebnis zu erhalten. Die Besonderheiten und die Güte der Methode wird anhand der Texterfassung von Gelegenheitsgedichten von Simon Dach illustriert.

OCR

OCR bezeichnet die Gesamtheit von Verfahren, die in der Lage sind, aus Rastergrafiken Schriftzeichen zu erkennen. Der Begriff wird sowohl für die eigentliche Mustererkennung als auch für den gesamten Prozess der Bildverarbeitung verwendet. Letzterer gliedert sich normalerweise in drei Schritte: **1. Bildoptimierung**: Diese besteht aus der Bitonalisierung der Digitalisate, ihrer Begradigung (sog. *Deskewing*) und aus der Entfernung von Artefakten (sog. *Despeckling*). Außerdem können beim Scannen entstandene Wellen in einzelnen Zeilen automatisch begradigt werden (sog. *Dewarping*). **2. Strukturerkennung** (*Optical Layout Recognition*, OLR): Die einzelnen Seiten werden u. a. in Spalten, Absätze und Zeilen gegliedert. **3. Mustererkennung** (OCR): Für diese Aufgabe gibt es verschiedene Lösungsvorschläge sowohl im kommerziellen wie auch im Open-Source-Bereich. Besonders verbreitet sind die Software *FineReader* der Firma ABBYY sowie *BITAlpha* aus dem Hause Tomasi, die u. a. von Bibliotheken eingesetzt werden. Die bekanntesten Open-Source-Lösungen sind das ursprünglich von Hewlett-Packard entwickelte und heute von Google betreute *Tesseract* (GitHub 2016a) und das ursprünglich am DFKI Kaiserslautern entwickelte *OCRopus* (GitHub 2016b).

Grundsätzlich lassen sich bei OCR zwei unterschiedliche Erkennungsansätze unterscheiden: zeichenorientierte Verfahren wie Tesseract vergleichen das Bild eines Zeichens Pixel für Pixel mit einer Datenbasis (dem sog. Modell) und geben das ähnlichste Zeichen zurück. Sequenzorientierte (segmentierungsfreie) Verfahren wie OCRopus legen ein Raster fester Größe über eine Zeile und bestimmen anhand der Folgen der einzelnen Spalten, repräsentiert als Bitvektoren (0 entspricht weiß, 1 schwarz) die wahrscheinlichste Zeichensequenz.

Gelegenheitsgedichte

Unsere Studie beschäftigt sich mit OCR am Beispiel von Gelegenheitsgedichten des 17. Jahrhunderts, denen durch die von Segebrecht (1977) initiierte literaturwissenschaftliche Neubewertung eine zunehmende kulturgeschichtliche Bedeutung zukommt (vgl. Klöker 2010: 39). Der Zugriff auf diese Drucke wurde durch das VD17 (HAB 2007-2016)[1] und durch das *Handbuch des personalen Gelegenheitsschrifttums in europäischen Bibliotheken und Archiven* (Garber 2001-2013) erleichtert. Dennoch kann ein digitales Korpus für diese Textsorte heute nur als Desiderat wahrgenommen werden. Für Werke von Simon Dach ist die Ausgangslage scheinbar besser: Mit der digitalisierten vierbändigen Ausgabe von Ziesemer (Ziesemer 1936-1938) steht ein großer Teil der heute bekannten Gedichte zur Verfügung (vgl. auch Dach o. J.; TextGrid 2015). [2] Jedoch trübt sich dieser Eindruck beim textkritischen Blick. [3]

111 Funeralschriften Simon Dachs wurden im Verlauf des DFG-Pilotprojektes zum *OCR-Einsatz bei der Digitalisierung der Funeralschriften der Staatsbibliothek zu Berlin* (2009-2011) (Federbusch / Polzin 2013) digitalisiert und per OCR erfasst. Die in der vorliegenden Studie genutzten Drucke zeichnen sich dahingehend aus, dass eine einheitliche Schrifttype sowie ein einfaches Layout vorliegen. Im Unterschied zu Texten des 18. und 19. Jahrhunderts war für diese Drucke noch ein relativ hoher manueller Aufwand erforderlich. Die Schrifttypen weisen daher eine vergleichsweise hohe Varianz bzgl. ihrer Form auf. Die 111 Trauergedichte weisen eine Textgenauigkeit von bis zu 95% auf. Der Schwerpunkt der folgenden Studie liegt auf der Entwicklung und Prüfung von Methoden, die perspektivisch eine korrektere Übertragung der Textquellen aus dem 17. Jahrhundert liefern soll.

Arbeitsablauf

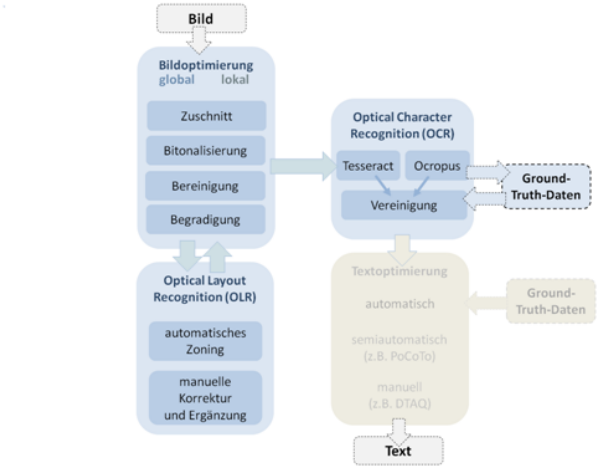


Abb. 1: Modell eines vollständigen Erfassungsworkflows (diese Studie betrifft die eingefärbten Stationen).

Abbildung 1 gibt einen Überblick über den Arbeitsablauf der hier vorgestellten Methode. Im Unterschied zu existierenden Workflows unterteilt unser Vorschlag die Bildoptimierung in zwei Phasen: 1. *global*: Das komplette Digitalisat wird beschnitten, binarisiert, begradigt und von Artefakten befreit. Danach findet die Optische Layouterkennung (OLR) statt. 2. *lokal*: Die identifizierten Textzonen werden aus dem Bild der Seite ausgeschnitten und nochmals begradigt. Dadurch wird die häufig zu beobachtende Trapezform der Digitalisate, die durch Scannen von Büchern ohne Auftrennen des Buchrückens entsteht, behandelt. Die Bilder für die einzelnen Zonen werden anschließend in Zeilen zerschnitten und den OCR-Engines übergeben.

Unser Vorgehen bei der OCR orientiert sich an der manuellen Texterfassung per *Double Keying*: Dabei werden Texte von zwei unabhängigen Erfassern transkribiert. Im Vergleich der beiden Textversionen werden die Unterschiede ermittelt und die korrekte Version ausgewählt. Um den Genauigkeitsgewinn durch die Mehrfacherfassung zu erhöhen, wurden zwei paradigmatisch verschiedene OCR-Verfahren, Tesseract und OCRopus, mit unterschiedlichen Stärken und Schwächen eingesetzt. Beide Open-Source-Programme erlauben ein Training auf die vorwendeten Typen und die Anwendung spezifischer OCR-Modelle. Dies ist wie Springmann et al. (2015) zeigen ein wesentlicher Vorteil gegenüber den meisten Closed-Source-Lösungen, da die mitgelieferten OCR-Modelle insbesondere für frühe Druckerzeugnisse bzw. gebrochene Schriften sehr schlechte Ergebnisse bzgl. der Textgenauigkeit liefern. Die automatische Vereinigung der beiden Textversionen findet im Wesentlichen auf Basis einer Textdifferenzberechnung mit Hilfe von *diff* (Hunt / McIlroy 1976) statt, wobei im Falle von Unterschieden verschiedene Bewertungsheuristiken zur Bestimmung der *korrekten* Textversion eingesetzt werden. Das skizzierte Vorgehen erlaubt auch die Kombination von mehr als zwei Textversionen sowie den anschließenden Einsatz von OCR-Nachkorrekturverfahren (vgl. z. B. Vobl et al. 2014).

Evaluation

Die Güte der hier vorgestellten Methode wird anhand der Volltexterfassung von Funeralschriften Simon Dachs (vgl. 1.2) evaluiert. Dabei konzentriert sich die Evaluation auf drei Punkte:

- Welchen Einfluss hat die Wahl der Binarisierungsmethode auf die Textgenauigkeit?
- Wie groß ist der Unterschied zwischen einem Standardmodell und einem speziell für die zu erfassenden Texte trainierten Modell bzgl. der Textgenauigkeit?
- Kann die Vereinigung zweier durch OCR erzeugter Texte die Textgenauigkeit erhöhen?

Ein typisches Beispiel für die Untersuchungsgrundlage sowie die entsprechenden OCR-Ausgaben gibt Abbildung 2.

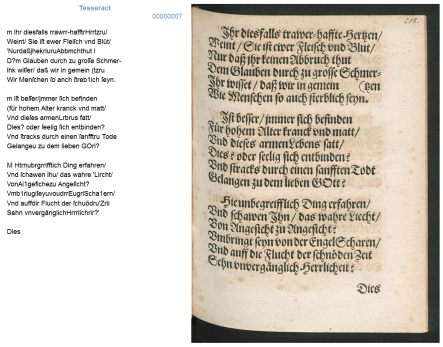


Abb. 2: Vergleich der OCR-Ergebnisse.

Material

Ground Truth

Voraussetzung für die Evaluation und das Modelltraining ist fehlerfreier Volltext (*Ground Truth*). Um für die Studie entsprechende Daten zu gewinnen, wurde eine manuelle Korrektur aller 111 Texte vorgenommen. Die Korrektur schloss nicht nur die Text-, sondern auch die datenstrukturelle Ebene ein. Der Aufwand belief sich auf 150 Stunden. Im Ergebnis liegen alle Texte im DTA-Basisformat vor und sind über die Qualitätssicherungsplattform DTAQ zugänglich.

Materialauswahl

Für das Training der spezifischen OCR-Modelle wurden 30 Seiten Ground-Truth zufällig ausgewählt. Für die Evaluation der Modelle wurden 25 andere zufällig ausgewählte Seiten verwendet.

Referenzlexikon

Zur Vereinigung beider OCR-Versionen wurde ein Referenzlexikon gültiger historischer Schreibungen des 17. Jahrhunderts herangezogen. Dazu wurden Wortformen (*n*=217067) aus DTA-Texten dieses Zeitraums extrahiert.

Durchführung

Vorverarbeitung

Für Beschneidung und Begradigung wurde das Programm *Scantailor* (GitHub 2016 a) eingesetzt. Für die Binarisierung, Artefaktbereinigung und Zeilenglättung wurde sowohl Scantailor als auch das in OCRopus enthaltene Werkzeug *nlbin* verwendet.

OLR

Die einzelnen Textzonen (Abschnitte und Kustoden) wurden mit Hilfe von *Leptonica* (Bloomberg 2001-2015) lokalisiert und manuell nachkorrigiert. Für die Untergliederung der Zonen in Zeilen wurde ebenfalls Leptonica eingesetzt.

OCR

Die Zeichenerkennung erfolgte sowohl mit OCRopus als auch mit Tesseract. Die erste Versuchsreihe basierte auf mitgelieferten Modellen. Für die zweite Versuchsreihe wurden die OCR-Programme mit Ground-Truth-Daten trainiert. Für das Training der OCRopus-Modelle wurde OCRopus eingesetzt. Dabei wurde für das Training aus Gründen der Modellvergleichbarkeit eine feste Anzahl von Iterationsschritten (*n*=30000) festgelegt. Die Tesseract-Modelle wurden mit Hilfe von *VietOCR* erstellt.

Textvereinigung

Die Textvereinigung wurde in *Python* mit Hilfe des Moduls *difflib* implementiert. Neben dem Referenzlexikon standen zur Konfliktauflösung auch die von den OCR-Programmen zurückgelieferten Konfidenzen auf Zeichenebene zur Verfügung. Waren sich die beiden Engines bzgl. eines Wortes bzw. einer Textsequenz uneins, wurde zunächst dem Wort Vorrang gegeben, dass sich im Referenzlexikon befindet. Konnte dort keine der beiden Versionen gefunden werden, wurde die Entscheidung auf Basis der Konfidenzwerte getroffen.

Qualitätsmessung

Die Bestimmung der Textqualität erfolgte durch Messung des Anteils falsch erkannter Zeichen (Fehlerrate in Prozent) im Vergleich zum fehlerfreien Volltext.

Ergebnisse und Diskussion

Tabelle 1 gibt einen Überblick über die Ergebnisse der Evaluation bzgl. der Fehlerrate auf Zeichenebene unter Berücksichtigung der Vorverarbeitung des Trainings- und Testmaterials, der Modellklasse (standard vs. spezifisch) und der eingesetzten OCR-Software (OCRopus, Tesseract). Das beste (grün) und das schlechteste Ergebnis (rot) sind hervorgehoben. Da wir keinen Einfluss auf die Vorverarbeitung der Trainingsmaterialien der mitgelieferten Modelle haben, ist die Matrix in dieser Hinsicht unvollständig.

				OCRopus		Tesseract			
Vorverarbeitung Training		Vorverarbeitung Test		standard		spezifisch		standard	spezifisch
nlbin		nlbin		25,41 %		6,04 %		-	53,10 %
		Scantailor		21,05 %		3,89 %		-	40,91 %
Scantailor		nlbin		-		6,95 %		37,37 %*	29,81 %
		Scantailor		-		4,21 %		27,15 %*	16,48 %

Tab. 1: Darstellung der Ergebnisse auf Einzel-OCR-Ebene im Bezug auf Vorverarbeitungsmethode für Trainings- und Testmaterial, Modelltyp und verwendete OCR-Software.

Die geringste erreichte Fehlerrate (3,89 %) liegt etwa im Bereich der Textgenauigkeit der 111 Gedichte aus der Pilotstudie von Federbusch (Federbusch / Polzin 2013). Die Fehlerrate von Tesseract ist jeweils höher als die von OCRopus. Der sequenzorientierte Ansatz hat klare Vorteile bei der Erkennung von Schriftzeichen, die die typischen Charakteristika früher Drucke aufweisen. [5]

Desweiteren zeigt sich, dass die Vorverarbeitung mit nlbin für Tesseract sowohl auf Trainings- als auch auf Testebene jeweils schlechtere Ergebnisse bringt. Für OCRopus sind die Ergebnisse bzgl. der Vorverarbeitung differenzierter: Die beste Kombination liefert eine Vorverarbeitung des Trainingsmaterials mit nlbin bei einer nachfolgenden Vorverarbeitung des Testmaterials mit Scantailor. Unterschiede im Ergebnis der Vorverarbeitung beider Programme illustriert Abbildung 3.

Abb. 3: Bild einer Textzeile nach der Vorverarbeitung mit nlbin (oben) und Scantailor (unten).

Die von Scantailor durchgeführte Bildvorverarbeitung ist deutlich normativer und für einen zeichenorientierten Ansatz wie Tesseract besser geeignet. Das Training sequenzorientierter Ansätze leidet unter dieser Vergröberung.

Es zeigt sich erneut, dass spezifisch trainierte Modelle eine massive Textgenauigkeitsverbesserung mit sich bringen können (vgl. auch Springmann et al. 2015).

Textvereinigung

Betrachtet man die Beispielausgaben in Abbildung 2, so wird der Qualitätsunterschied zwischen beiden OCR-Programmen ersichtlich. An einzelnen Stellen jedoch (z. B. Großbuchstaben am Anfang der Zeile im letzten Abschnitt) hat Tesseract Erkennungsvorteile.

Ausgehend von diesem Befund wurde der jeweils genaueste Text von OCRopus und Tesseract miteinander vereinigt. Es hat sich gezeigt, dass die Konfidenzen, die die Programme für jedes Zeichen zurückliefern, kein verlässliches Kriterium sind, um Konflikte aufzulösen. Die Fehlerrate nimmt zu. Die Strategie, Wörter bzw. Sequenzen zu bevorzugen, die sich im Referenzlexikon befinden, hat dagegen eine messbare Verbesserung mit sich gebracht. Die Anzahl der falsch erkannten Zeichen konnte um 14 % reduziert werden (Fehlerrate 3,34 %). Es ist zu vermuten, dass der Effekt größer wäre, wenn zwei OCR-Ergebnisse mit vergleichbarer Qualität vorlägen. Dies bleibt jedoch zum jetzigen Zeitpunkt für Drucke des 17. Jahrhunderts ein Desiderat.

Notes

1. Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 17. Jahrhunderts.
2. Vgl auch Dach (o. J.) in http://www.zeno.org/Literatur/M/Dach,+Simon/Gedichte sowie TextGrid (2015).
3. „Ziesemers Dach-Ausgabe ist textlich zu wenig genau, um auch für die dort abgedruckten, fast ausnahmslos deutschsprachigen, Gedichte den Rückgriff auf die kasualen Einzeldrucke und andere zeitgenössische Ausgaben entbehren zu können. Jede Stichprobe erweist für jedes einzelne Gedicht Transkriptionsfehler und unerklärte Texteingriffe.“ (Walter 2008: 466)
5. Für Frakturdrucke des 19. Jahrhunderts ist ein solch starker Unterschied zwischen den Tesseract und OCRopus nicht nachgewiesen.

Bibliographie

Bloomberg, Dan (2001-2015): Leptonica http://www.leptonica.com/ [letzter Zugriff: 15. Oktober 2015].

Dach, Simon (o. J.): *Gedichte* http://www.zeno.org/Literatur/M/Dach,+Simon/Gedichte [letzter Zugriff 15. Oktober 2015].

Federbusch, Maria / Polzin, Christian (2013): *Volltext via OCR - Möglichkeiten und Grenzen*. Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz. Berlin Staatsbibliothek zu Berlin http://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/historische_drucke/pdf/SBB_OCR_STUDIE_WEBVERSION_Final.pdf [letzter Zugriff 15. Oktober 2015].

Garber, Klaus (2001-2013): *Handbuch des personalen Gelegenheitsschrifttums in europäischen Bibliotheken und Archiven*. 13 Bände. Hildesheim / Zürich / New York: Olms / Weidmann.

GitHub Inc. (2016a): *ScanTailor* http://scantailor.org/ [letzter Zugriff 15. Oktober 2015].

GitHub Inc. (2016b): *OCRopus* https://github.com/tmbdev/ocropy [letzter Zugriff 15. Oktober 2015].

GitHub Inc. (2016c): *Tesseract* https://github.com/tesseract-ocr [letzter Zugriff 15. Oktober 2015].

HAB = Herzog August Bibliothek Wolfenbüttel (2007-2016): *VD17*. Das Verzeichnis der im deutschen Sprachraum erschienenen Druck des 17. Jahrhunderts http://www.vd17.de/index.php?category_id=1&article_id=1&clang=0 .

Hunt, James W. / McIlroy, M. Douglas (1976): "An Algorithm for Differential File Comparison" in: *Computing Science Technical Report* (Bell Laboratories) 41 http://www.cs.dartmouth.edu/~doug/diff.pdf

Klöker, Martin (2010): "Das Testfeld der Poesie. Empirische Betrachtungen aus dem Osnabrücker Projekt zur 'Erfassung und Erschließung von personalen Gelegenheitsgedichten'", in: Keller, Andreas / Lösel, Elke / Wels, Ulrike / Wels, Volkhard (eds.): *Theorie und Praxis der Kasualdichtung in der Frühen Neuzeit* (= Chloe. Beihefte zu Daphne 43). Amsterdam / New York: Rodopi 39-84.

Python Software Fundation (1990-2016): *difflib - Helpers for Computing Deltas* https://docs.python.org/2/library/difflib.html [letzter Zugriff 15. Oktober 2015].

Segebrecht, Wulf (1977): *Das Gelegenheitsgedicht*. Ein Beitrag zur Geschichte und Poetik der deutschen Lyrik. Suttgart: Metzler.

Springmann, Uwe / Lüdeling, Anke / Schremmer, Felix (2015): "Zur OCR frühneuzeitlicher Drucke am Beispiel des RIDGES-Korpus von Kräutertexten (Poster)", in: *Tagung der DHd (Digitale Geisteswissenschaften im deutschsprachigen Raum)*, Graz https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/pdf/

SpringmannLuedelingSchremmer2015.pdf [letzter Zugriff 15. Oktober 2015].

TextGrid (2015): *Die digitale Bibliothek bei TextGrid* https://textgrid.de/digitale-bibliothek [letzter Zugriff 15. Oktober 2015]

VietOCR http://vietocr.sourceforge.net/ [letzter Zugriff: 15. Oktober 2015].

Vobl, Thorsten / Gotscharek, Annette / Reffle, Uli / Ringlstetter, Christoph / Schulz, Klaus U. (2014): "PoCoTo - an open source system for efficient interactive postcorrection of OCRed historical texts" in: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)*: 57-61 http://dl.acm.org/citation.cfm?id=2595197 [letzter Zugriff 15. Oktober 2015].

Walter, Axel E.(2008): "Dach digital? Vorschläge zu einer Bibliographie und Edition des Gesamtwerks von Simon Dach nebst einigen erläuterten Beispielen vernachlässigter bzw. unbekannter Gedichte", in: Walter, Axel E. (ed.) in: *Simon Dach (1605–1659)*. Werk und Nachwirken. Tübingen: Niemeyer: 465-522.

Ziesemer, Walter (ed.) (1936-1938): *Simon Dach: Gedichte*. Vier Bände. Halle an der Saale: Niemeyer.

Pattern Mining in Keilschriftzeichnungen

Bogacz, Bartosz

bg.bartek@gmail.com
Universität Heidelberg, Deutschland

Mara, Hubert

hubert.mara@iwr.uni-heidelberg.de
Universität Heidelberg, Deutschland

Keilschrifttafeln gehören zu den ältesten Textzeugen, die im Umfang mit den Texten in lateinischer und alt-griechischer Sprache vergleichbar sind. Da diese Tafeln aus dem gesamten Alten Orient über beinahe viertausend Jahre in Verwendung waren (Soden 1994), lassen sich damit viele interessante Fragestellungen zur Entwicklung von Religion, Politik, Wissenschaft, Handel bis hin zu Klimaveränderungen (Kaniewski et al. 2013) beantworten. Die aus Ton geformten Tafeln, bei denen Zeichen (Borger 2010) als keilförmige Abdrücke mit einem eckigen Stylus eingedrückt wurden, erfordern neue informationstechnische Methoden zu der Dokumentation und Analyse als die in Archiven üblichen Flachwaren.

Keilschrifttafeln werden mit Hilfe verschiedenster Methoden digitalisiert und in verschiedene, untereinander nicht kompatible Formate übertragen. Sie werden photographisch mit wechselnden Lichtverhältnissen aufgezeichnet, handschriftlich oder digital abgezeichnet oder mit Hilfe eines 3D-Scanners aufgenommen (Mara et al. 2010; Mara / Krömker 2013). Jede dieser Repräsentationen erfordert ein eigenes Tool-Set zur Analyse und die textuelle Analyse ist auf die jeweilige Repräsentation beschränkt.

Die Initiative für eine digitale Keilschriftdatenbank (Cuneiform Digital Library Initiative - CDLI) stellt mehr als 300.000 Keilschrifttafeln je nach Verfügbarkeit in Form von handgefertigten Abschriften, Photographien oder Umschriften zur Verfügung. Diese Datenbank besitzt keine Möglichkeit Keilschrifttafeln nach den Keilsymbolen zu durchsuchen.

In unserer bisherigen Arbeit (Bogacz / Massa et al. 2015) stellten wir Verfahren und einen Ablauf zur Homogenisierung von den drei gängigsten Datenquellen vor. Keilschriftabdrücke wurden handschriftlichen Zeichnungen, digital abgezeichneten und 3D-gescannten Tafeln entnommen. Die Datenquellen wurden zuerst, falls nötig, in das SVG Format (Scalable Vector Graphics) vektorisiert. SVG Dateien sind ein offener Standard zur Beschreibung von Vektorgrafiken, der sich den XML Standard zu nutze macht.

Die Nutzung dieses Dateiformates ermöglicht uns Wörter in den digitalen Abzeichnungen mit ihrer Übersetzung zu Annotieren und als XML-Tags zu den Grafikpfaden, die den Wörtern entsprechen, in den SVG Dateien selbst abzuspeichern. Wir nutzten diese Annotationen, um die Genauigkeit unserer Worterkennung zu überprüfen (Bogacz / Gertz et al. 2015).

Auf Grundlage der homogenisierten Datenbasis führten wir eine minimale und einheitliche Beschreibung von Keilabdrücken mit Hilfe von Merkmalsvektoren ein. Die Abdrücke einer Keilschrifttafel in dem jeweiligen Datenformat werden erkannt und extrahiert. Bei der Extraktion werden die einzelnen Keile durch mehrere verschiedene, sich ausschließende, Merkmalsvektoren modelliert. Die abschließend gewählte Untermenge von Keilmodellen für die gegebenen Keile einer Tafel ist eine global optimale Zuordnung von Keilmodellen zu den jeweiligen Keilabdrücken. Dieser Ansatz wurde gewählt, da die Abdrücke oft beschädigt oder nicht eindeutig identifizierbar sind.

Die reduzierte Darstellung als Merkmalsvektoren ermöglicht eine Analyse der Daten mit gängigen Methoden aus dem Bereich des maschinellen Lernens, wie der Priniciple Component Analysis (PCA) Dimensionsreduktion, dem k-Means Algorithmus oder auch einem Entscheidungsbaum (Mohri et al. 2012), und das Abspeichern der Keilabdrücke und der Keilschrifttafeln in austauschbaren XML Dateien zur weiteren Analyse oder in einer effizienten Suchstruktur als Grundlage für einen Suchalgorithmus.

In dieser Arbeit stellen wir ein Verfahren zur vollständig automatisierten Suche von Keilschriftsymbolen vor. Wir übernehmen die Idee von "Query Words" und adaptieren sie für geometrische Symbole. Anstatt ausschließlich Übersetzungen von Keilschrifttafeln zu durchsuchen und nicht übersetzte Tafeln auszulassen, können wir alle homogenisierten

Tafeln nach Keilkonfigurationen durchsuchen. Eine beliebige geometrische Anordnung von Keilen im Merkmalsvektor Repräsentation wird als Query (Abfrage) genutzt, nach welcher Tafeln abgesucht werden können.

Unser Verfahren baut eine Suchstruktur auf, die danach mit Keilkonfigurationen abgesucht werden kann. Zuerst wird durch eine Radial Basis Function (RBF) Kernel-PCA Dimensionsreduktion (Schölkopf 1997) der Merkmalsraum der Merkmalsvektoren reduziert. Es gibt nur wenige Keiltypen und diese werden durch die hochdimensionalen (12 Merkmale pro Keil) Merkmalsvektoren überspezifisch beschrieben. Danach wird ein k-Means Clustering (Kanungo et al. 2002) durchgeführt, um die einzelnen Keiltypen automatisiert zu erkennen. Die gefunden Gruppierungen bilden die Basis für ein Wörterbuch an bekannten Keilkonfigurationen. Dieses Wörterbuch wird nun erweitert indem ein spatiales Frequent Pattern Mining (Han et al. 2007) der Tafeln durchgeführt wird. Häufig vorkommende und dicht zusammen liegende Keiltypen werden zu neuen Einträgen zusammengefasst. Keilschrifttafeln werden somit anhand der Positionen von im Wörterbuch vorhandenen Keilkonfigurationen beschrieben.

Ein Keilschriftszeichen wird gesucht, indem es in im Wörterbuch bekannte Keilkonfiguration unterteilt wird. Dazu werden die Merkmalsvektoren des Zeichens mit gelernten PCA reduziert und dem gelernten k-Means klassifiziert. Danach werden bekannte Konfigurationen im gesuchten Zeichen durch erneutes spatiales Frequent Pattern Mining identifiziert. Nun wird eine Schnittmenge von bekannten Konfigurationen im gesuchten Zeichen mit der Menge an bekannten Konfigurationen auf der Tafel gebildet. Übereinstimmungen werden durch ein genaueres Verfahren verglichen (Bogacz / Gertz et al. 2015).

Unser Verfahren Pattern Mining a Dictionary of Complex Structures (PDCS) macht sich die geringe Anzahl von Keiltypen (Winkelhaken, stehender Keil und liegender Keil) und häufig vorkommende Keilkonfigurationen zu nutze, um den Suchraum zu reduzieren. Zusammenfassend basiert es auf der Annahme, dass sich das zu durchsuchende Objekt in bekannte und grundlegende Formen, Keile der Keilschrift, zerlegen lässt, und die gesuchte Form eine geometrische Anordnung dieser Grundform ist. Dafür erweitern das Konzept des Frequent Pattern Minings indem wir die Geometrie der häufig vorkommenden Muster beachten.

Die k-Means Gruppierung der Keiltypen hat gegenwärtig eine Fehlerrate von 10%. Wir planen die Fehlerrate zu reduzieren indem wir die Parameter der PCA Dimensionreduktion automatisiert lernen und optimieren. Das Bilden der geometrischen Schnittmenge ist ein zeitaufwändiger Prozess. Wir arbeiten an einer Methode diesen Algorithmus zu beschleunigen indem wir Keilkonfigurationen aus dem Wörterbuch entfernen, die nicht zur Suche beitragen. Weitere mögliche Anwendungsbereiche für unser Verfahren sind Chinesische Zeichen, Heraldik, Maya Schriftzeichen und die kodikologische Untersuchung der Anordnung von Textpassagen eines Keilschrifttextes.

Bibliography

Bogacz, Bartosz / Gertz, Michael / Mara, Hubert (2015): "Character Retrieval of Vectorized Cuneiform Script", in: *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR)*, Nancy, France.

Bogacz, Bartosz / Massa, Judith / Mara, Hubert (2015): "Homogenization of 2D & 3D Document Formats for Cuneiform Script Analysis", in: *Proceedings of the 2015 Workshop on Historical Imaging and Processing*, Nancy, France 115-122.

Borger, Rykle (22010): *Mesopotamisches Zeichenlexikon* (= Alter Orient und Altes Testament – Veröffentlichungen zur Kultur und Geschichte des Alten Orients und des Alten Testaments 305). Münster: Ugarit-Verlag.

Han, Jiawei / Cheng, Hong / Xin, Dong / Yan, Xifeng (2007): "Frequent pattern mining: current status and future directions", in: *Data Mining and Knowledge Discovery* 15, 1: 55-86.

Kaniewski, David / Van Campo, Elise / Guiot, Joel / Le Burel, Sabine / Otto, Thierry / Baeteman, Cecile (2013): "Environmental Roots of the Late Bronze Age Crisis", in: *PLoS One* 8, 8 http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071004 [letzter Zugriff 07. Februar 2016].

Kanungo, Tapas / Mount, David M. / Netanyahu, Nathan S. / Piatko, Christine D. / Silverman, Ruth / Wu, Angela Y. (2002): "An efficient k-means clustering algorithm: Analysis and implementation. Pattern Analysis and Machine Intelligence", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7: 881-892.

Mara, Hubert / Krömker, Susanne (2013): "Vectorization of 3D-Characters by Integral Invariant Filtering of High-Resolution Triangular Meshes", in: *Proceedings of 12th International Conference on Document Analysis and Recognition (ICDAR / IAPR)*, Washington D.C., USA 62–66.

Mara, Hubert / Krömker, Susanne / Jakob, Stefan / Breuckmann, Bernd (2010): "GigaMesh and Gilgamesh - 3D Multiscale Integral Invariant Cuneiform Character Extraction", in: *Proceedings of VAST10 - International Symposium on Virtual Reality, Archaeology and Cultural Heritage*, Palais du Louvre, Paris, France 131-138.

Mohri, Mehryar / Rostamizadeh, Afshin / Talwalkar, Ameet (2012): Foundations of Machine Learning (= Adaptive Computation and Machine Learning series). Cambridge, Massachusetts: MIT Press.

Schölkopf, Bernhard / Smola, Alexander / Müller, Klaus-Robert (1997): "Kernel Principal Component Analysis", in: Gerstner, Wulfram / Germond, Alain / Hasler, Martin / Nicoud, Jean-Daniel (eds.): *Artificial Neural Networks*. Proceedings of the 7th International

Conference Lausanne, Switzerland (ICANN'97). Berlin / Heidelberg: Springer 583-588.

Soden, Wolfram von (1994): *The ancient Orient*. An introduction to the study of the ancient Near East. Michigan: Wm. B. Eerdmans Publishing Co.

Formate als Sackgassen: Handlungsempfehlungen

Bohl, Benjamin W.
bohl@edirom.de
Zentrum für Musik- und Filminformatik, Hochschule Ostwestfalen-Lippe

Berndt, Axel, Dr.
berndt@hfm-detmold.de
Zentrum für Musik- und Filminformatik, Hochschule Ostwestfalen-Lippe

Senft, Björn
bsenft@s-lab.uni-paderborn.de
Software Quality Lab, Universität Paderborn

Im Kontext des *Zentrum - Musik - Edition - Medien* beschäftigen sich die Autoren mit der Modellierung und Codierung musikalischer Phänomene. Formate zur Codierung von Musik reichen von ASCII-basierten Codierungen (Humdrum, ABC-Notation) über SGML-basierte Formate (SMDL) und XML-basierte Formate (MusicXML, MEI) bis hin zu technischen Steuerdaten (MIDI und spezifische Formate für Sequencer-Programme) oder Audiodaten (FLAC, MP3) (vgl. Selfridge-Field 1997).[1] Es sind die spezifischen Anforderungen und Ansprüche, der Fokus auf die Darstellung bestimmter (musikalischer) Phänomene, und die Ausrichtung auf einen bestimmten Nutzungskontext, die jedes Format prägen und ihm seine Berechtigung geben (vgl. Veit 2006). Jedoch stellt der Informationsaustausch zwischen den verschiedenen Nutzergruppen, mit ihren spezifischen Formatvorlieben und den jeweils relevanten -erfordernissen, ein essentielles Problem dar. Ausgehend von dieser, zwar im Beispiel musikspezifischen, im Kern jedoch allgemeingültigen Problemstellung, sollen im Folgenden Handlungsempfehlungen entwickelt werden, die zur Planung und Einschätzung digital arbeitender Projekte herangezogen werden können.

„Während MusicXML als Austauschformat inzwischen größte Verbreitung gefunden hat, versucht MEI ausdrücklich, musikeditorische Anforderungen zu erfüllen.“ (Kepper 2009: 216). Im Vergleich zu MusicXML und vielen anderen Codierungsformaten für Musik (vgl. Selfridge-Field 1997), ermöglicht das XML-basierte Codierungsformat der Music Encoding Initiative (MEI) eine umfassende metadatliche Beschreibung (vgl. Richts / Herold 2014), sowie die Codierung editorischer Sachverhalte (vgl. Roland / Kepper 2014). Aufgrund dieses Alleinstellungsmerkmals hat es sich im Bereich der digitalen Musikedition etabliert und findet sich inzwischen auch im Recommended Formats Statement der Library of Congress (vgl. Library of Congress 2015). Somit ist es nachvollziehbar, dass zunehmend mehr Editionsprojekte auf MEI zurückgreifen, nicht zuletzt, um dem *Digital Turn* und seiner Forderung nach einer nachhaltigen Bereitstellung von Forschungsdaten zur Nachnutzung nachzukommen.

Denkbare Nutzungsszenarien in diesem Kontext sind u. a. im Music Information Retrieval (MIR), in der Interpretationsforschung sowie in der weiteren Verarbeitung von Musikdaten (Musikgenerierung und -adaption), im Notendruck und in der Musikproduktion verortet. Jede Disziplin bringt ihre eigenen Anforderungen an die Modellierung der Informationen mit.

- Im MIR werden einfache Strukturen benötigt, etwa CSV-Daten, um ein schnelles Parsen für Echtzeit-Anwendungen zu gewährleisten.
- In der Interpretationsforschung spielt die Analyse von Audiodaten eine wichtige Rolle. MEI ist dafür nicht spezifisch genug, enthält keine Audiodaten und keine Möglichkeit, solche Analyseergebnisse zu repräsentieren.
- Für die Musikgenerierung und -adaption werden Tondaten und Steuerdaten vorausgesetzt. Im Falle von MEI sind erstere unvorteilhaft strukturiert, deshalb aufwendig zu prozessieren. Steuerdaten lassen sich nur unzureichend einbinden. Auch die Musikproduktion arbeitet vornehmlich mit elektronischen Steuerdaten, sowie Audiodaten.
- Für das Layout des graphischen Notenbildes, ist die logische Struktur der musikalischen Informationen zweitrangig. Aufgrund der zu geringen und unvollständigen Unterstützung durch Notationsprogramme (etwa mittels Importer) ist MEI für den Notendruck derzeit irrelevant.

Seine durchaus beabsichtigten Uneindeutigkeiten und die Möglichkeiten der Anreicherung mit editionsspezifischen Informationen machen MEI zu einem für die digitale Musikedition mächtigen, für die exemplarisch beschriebenen weiteren Nutzungszenarien jedoch unpraktikablen Format. Dies birgt die Gefahr, dass trotz der Bereitstellung der in MEI codierten Editionsdaten das Ende der Nutzungskette bereits erreicht ist.

Ähnliche „Sackgasseneffekte“ lassen sich auch in anderen Nutzungskontexten und deren Formaten beobachten. Zu deren Überwindung sind mehrere grundlegende Szenarien denkbar: die Erweiterung eines bestehenden Formates, die gleichzeitige Nutzung

mehrerer Formate (ggf. gekapselt in einem Container-Format), oder die Konvertierung in andere Formate. Jeder Ansatz ist mit spezifischen Vor- und Nachteilen verbunden, die im Folgenden diskutiert werden sollen.

Lösungsansätze

Erweiterung

Eine naheliegende Lösung mag in der Erweiterung des Formats bestehen. Das beinhaltet die Modellierung, Formalisierung und Implementierung der neuen Elemente. Während dies für punktuelle Phänomene noch praktikabel sein mag, ziehen umfangreichere Erweiterungen eine immer größere Komplexität der Datenstruktur nach sich. Insbesondere dann, wenn eine bereits die spezifischen Erfordernisse der einen Anwendungsdomäne widerspiegelnde Struktur mit einer weiteren, einer ganz anderen Domäne Rechnung tragenden Struktur, überlagert wird. Dies kann im Falle vom MEI bereits jetzt beobachtet werden, wie beispielsweise durch das Nebeneinander verschiedener Zeitdarstellungen (symbolische / musikalische Zeit, Aufführungszeit) und einzelne, jedoch unvollständige Querbezüge zum MIDI-Standard.[2] Gegebenenfalls kann ein grundsätzlich neues Datenformat dabei entstehen.

Die Replikation relevanter Informationen in einen neuen und entsprechend anders strukturierten Bereich des Formats würde hingegen in einem wenig (speicher-)effizienten Format mit zahlreichen Redundanzen resultieren. Generell sind Redundanzen aufwendig zu pflegen. Hierbei auf Referenzen zurückzugreifen kann sowohl die Gefahr von Inkonsistenzen, als auch den Wartungsaufwand verringern.

Zudem muss sich ein “allen gerecht werden wollendes” Format neben den spezialisierten, etablierten Formaten durchsetzen können. Dass dies gelingt, ist höchst fraglich, da zunächst alle Verarbeitungsverfahren und Werkzeuge, die für die etablierten Formate bereits bestehen, neu implementiert oder zumindest angepasst werden müssen. Ferner sind die etablierten Formate und ihre zugehörigen Werkzeuge gerade dank ihrer Spezialisierung auch für ihren jeweiligen Anwendungskontext optimiert und unterstützen die effiziente Arbeit mit den Daten. Ein weniger spezialisiertes Format ist daher oft ineffizienter, nicht nur hinsichtlich seines Speicherbedarfs, sondern auch hinsichtlich des benötigten Rechen- bzw. Verarbeitungsaufwandes.

Manuelle parallele Datenhaltung

Wenn die Konkurrenz zu etablierten Formaten vermieden werden soll, was im Sinne der Nachhaltigkeit generell zu empfehlen ist, bietet sich die parallele Bereitstellung der Daten in mehreren Formaten an. In Abhängigkeit der zu adressierenden Anwendungsszenarien und den damit einhergehenden Anwenderprofilen wird eine Auswahl der relevanten Formate getroffen. Die Daten werden nun parallel in jedem dieser Formate gepflegt. Das kann unter Zuhilfenahme der dafür existierenden Werkzeuge geschehen, sodass kein Software-technischer Entwicklungsaufwand anfällt. Jedoch entsteht ein Mehraufwand in der Datenpflege, denn die allen Formaten gemeinen Inhalte (Redundanzen) müssen synchron gehalten werden. Jedes Format hat ferner seinen eigenen Anwendungskontext mit entsprechenden, spezialisierten (nichtredundanten) Inhalten. Automatismen, welche dem Anwender diesen Synchronisationsaufwand abnehmen, sind im Allgemeinen nicht vorhanden; die Arbeit geschieht „manuell“. Die richtige Verwendung und Pflege der Daten und Werkzeuge erfordert eine entsprechende Bearbeitungsdisziplin der Editoren. Dies stellt eine Gefahr für die Konsistenz des Datensatzes dar und birgt die Gefahr des Zerfalls des Datensatzes in einzelne unzusammenhängende, weil nicht synchrone, Datenobjekte.

Konvertierung

Möchte man den aus der parallelen Datenhaltung resultierenden manuellen Mehraufwand vermeiden, bietet die Nutzung von Konvertern eine Erleichterung. Der Nutzer arbeitet, so lange es seiner Fragestellung genügt, in ein und demselben Format und konvertiert es erst bei Bedarf in andere Formate. In einer entsprechenden Arbeitsumgebung kann dies durch Automatismen unterstützt werden, welche bei Veränderungen an einem Objekt die Synchronisation mit den Parallelobjekten durchführen. Konverter können innerhalb bestehender Anwendungsprogramme in Form von Importern die formatübergreifende Arbeit erleichtern.

Sofern jedoch die Formate nicht äquivalent sind, wovon im Allgemeinen ausgegangen werden muss, kann die Konvertierung mit Informationsverlust verbunden sein, vor allem dann, wenn die betreffenden Informationen im Zielformat der Konvertierung grundsätzlich nicht repräsentierbar sind. So kann die Erstellung eines Datenobjektes durch Konvertierung lediglich der Startpunkt sein, an welchem die dem Ausgangs- und Zielformat gemeinsamen Inhalte übernommen werden und von wo aus die formatspezifischen Inhalte dann vom Nutzer einzupflegen sind. Sollten für das Zielformat relevante Daten im Ausgangsformat fehlen, so sind auch diese vom Nutzer zu ergänzen. Die gleiche Art von Informationsverlust ist auch bei der Rückkonvertierung zu bedenken. Für den Anwender steigt also der Pflegeaufwand für die in mehreren Formaten vorgehaltenen Daten in dem Maße,

in dem konvertierungsbedingter Informationsverlust und -ergänzung manuell ausgeglichen werden müssen. Die Konvertierung automatisiert lediglich die Pflege der redundanten Inhalte, d. h., die Schnittmenge der Datensammlung in den verschiedenen Formaten. Denkbar ist es in einigen Fällen, die nicht in der Schnittmenge enthaltenen Informationen separat zu den Datenformaten zu speichern, um sie bei der Rückkonvertierung wieder einzupflegen. Eine weitere Voraussetzung für eine (zumindest in weiten Teilen) automatisierte Rückkonvertierung stellt das Wissen über die Transformationshistorie dar.

Handlungsempfehlung

Die bisherigen Ausführungen lassen bereits erkennen: Eine bequeme Lösung gibt es nicht. Jeder der genannten Lösungsansätze findet in der Praxis bereits mehrfach Anwendung, jeweils mit den entsprechenden Vor- und Nachteilen. Diese gilt es abzuwägen, will man sich im Rahmen eines konkreten Projektes für einen Ansatz entscheiden. Dabei werden die folgenden vier Kriterien von maßgeblicher Bedeutung sein.

Nachhaltigkeit: Sollen die Daten längerfristig und über das Projekt hinaus nutzbar sein?

Wenn dies gewünscht ist, sollten die Ergebnisse in den etablierten Formaten der zur Nachnutzung angedachten Nutzergruppen gespeichert werden. Ein eigens im Projekt entwickeltes Format oder Derivat kann, wenn es sich nicht etabliert und keine dem technischen Fortschritt folgenden Aktualisierungen garantiert, keine Nachhaltigkeit sichern.

Rückfluss: Findet ein uni- oder bidirektionaler Austausch zwischen den verschiedenen, vom Projekt adressierten Nutzergruppen statt?

Ein eigens für das Projekt entwickeltes Datenformat wird den aus dem Projektkontext heraus gerichteten Austausch erschweren. Der Rückfluss wird ohne entsprechende Konverter für das eigene Format kaum praktikabel sein.

Der immer wieder auszugleichende Informationsverlust im Konverteransatz wird in einem unidirektionalen Szenarium kaum ein Problem darstellen, denn ohne den Rückkonvertierungsschritt entfällt die mehrfache Einpflege der nichtredundanten Inhalte. Die Übernahme der redundanten Inhalte wird hingegen auch beim Rückfluss erleichtert.

Die parallele Datenhaltung wird für den bidirektionalen Austausch am praktikabelsten sein, weil sie konzeptionell vorsieht, alle Inhalte in den bevorzugten Formaten der adressierten Nutzergruppen vorzuhalten und Änderungen in allen Repräsentationen zu synchronisieren.

Synchronisationsaufwand: Wie hoch darf der Aufwand zur Datensatzpflege sein?

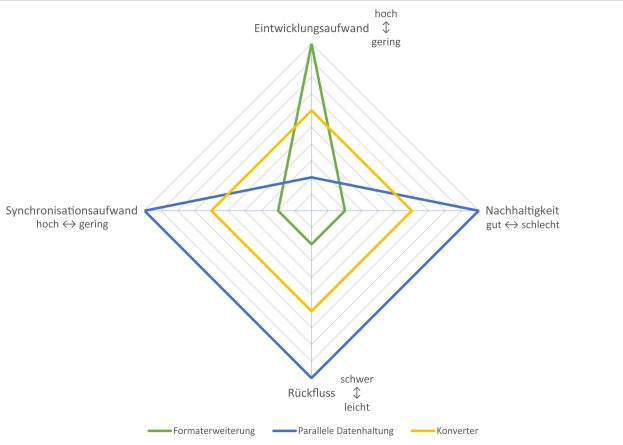
Der manuelle Aufwand zur Datensatzpflege ist bei der Vorhaltung der Daten in mehreren Formaten ohne Automatisierungen höher als bei den anderen vorgeschlagenen Lösungen. Der Rückgriff auf ein im Projekt praktisch ausschließlich verwendetes eigenes Format (eigene Formatanpassung), minimiert den Synchronisationsaufwand. Konverter stellen einen Mittelweg dar, denn die redundanten Informationen können (semi-)automatisch synchronisiert werden, lediglich die formatspezifischen Inhalte erfordern manuellen Pflegeaufwand.

Entwicklungsaufwand: Wieviel Entwicklungskapazitäten sind im Projekt vorgesehen?

Die Entwicklung eines eigenen Datenformats oder von Derivaten existierender Formate zieht auch die Entwicklung von Verarbeitungsverfahren und Werkzeuge nach sich, setzt also einen insgesamt hohen Bedarf an (Software-)Entwicklungskapazitäten voraus.

Auch Konverter verlangen Entwicklungskapazitäten, wenn auch (abhängig von der Menge der unterstützten Datenformate) in bescheidenerem Umfang, zumal für viele etablierte Formate bereits ausgereifte Konverter existieren.

Für die manuelle Pflege der Daten, parallel in mehreren Formaten, sind keine Entwicklungsarbeiten notwendig; hierfür genügen die bestehenden Editoren für die jeweiligen Formate.



Die vorstehende Abbildung veranschaulicht die Gewichtungen der vorgestellten Lösungsansätze in einem Starplot. Dies soll eine Orientierungshilfe zur Projektplanung sein und als Grundlage von Handlungsempfehlungen dienen. Selbstverständlich gibt es weiterführende, die obigen Ansätze kombinierende Möglichkeiten, die etwa im Rahmen von Datenbanksystemen oder integrierten Arbeitsumgebungen bestimmte Aufgaben vereinfachen können. Sofern diese Systeme nicht bereits bestehen, ist dies mit einem gesteigerten Programmieraufwand verbunden, der von inhaltsorientierten Projekten kaum zu leisten ist. Daraus motiviert sich der Bedarf für die Bereitstellung von projektübergreifenden und langfristigen Forschungsinfrastrukturen, eine Thematik, der sich das *Zentrum - Musik - Edition - Medien* mit der Erforschung nachhaltiger Entwicklungskonzepte im Bereich der digitalen Musikedition widmet. Für die

langfristige Bereitstellung einer solchen Infrastruktur bietet die aktuelle Förderpolitik in Deutschland jedoch nur selten die nötigen Grundlagen.

Notes

1. Bei Selfridge-Field (1997) ist eine umfassende Auflistung und Beschreibung unterschiedlicher Codierungsformate für Musik zu finden. Das Buch kann gewissermaßen als Standardwerk in diesem Bereich angesehen werden.
2. Dieser Ansatz eines immer größer werdenden Systems kann ebenfalls bei etablierten Softwaresystemen festgestellt werden. Solche komplexen Systeme werden in der Folge immer schwieriger zu warten und können ihren eigentlichen Zweck immer schlechter erfüllen. Daher gibt es bei diesen Systemen den Trend zur Modularisierung und Spezialisierung (vgl. Krahn / Rumpe 2005).

Bibliographie

Kepper, Johannes (2009): "XML-basierte Codierung musikwissenschaftlicher Daten – Zu den Voraussetzungen einer digitalen Musikedition", in: *it – Information Technology* 51: 216–221.

Library of Congress (2015): *Preservation: Recommended Formats Statement* https://www.loc.gov/preservation/resources/rfs/textmus.html [letzter Zugriff 08. Januar 2016].

Krahn, Holger / Rumpe, Bernhard (2005): *Evolution von Softwarearchitekturen*. Informatik-Bericht 2005-04. Braunschweig: TU Braunschweig http://www.se-rwth.de/~rumpe/publications20042008/Evolution-von-Software-Architekturen.pdf [letzter Zugriff 08. Januar 2016].

Richts, Kristina / Herold, Kristin (2014): *Daten- und Metadatenformate in den Fachdisziplinen: Musikwissenschaft* https://wiki.de.dariah.eu/display/publicde/3.3+Musikwissenschaft [letzter Zugriff: 04. Januar 2016].

Roland, Perry / Kepper, Johannes (eds.) (2014): *Music Encoding Initiative Guidelines*. Release 2013. Revision 2.1.1. Charlottesville / Detmold: Music Encoding Initiative Council http://github.com/music-encoding/music-encoding/releases/download/MEI2013_v2.1.1/MEI_Guidelines_2013_v2.1.1.pdf [letzter Zugriff: 04. Januar 2016].

Selfridge-Field, Eleanor (1997): *Beyond MIDI*. The handbook of musical codes. Cambridge: MIT Press Ltd.

Veit, Joachim (2006): "Musikwissenschaft und Computerphilologie – eine schwierige Liaison?", in: *Jahrbuch für Computerphilologie* 7: 67–92 http://computerphilologie.uni-muenchen.de/jg05/veit.html [letzter Zugriff 30. September 2015].

Algorithmische Visualisierungen: Ausdruck von Routinen und Denkstilen in den Digital Humanities

Bubenhofer, Noah
bubenhofer@cl.uzh.ch
Universität Zürich, Schweiz

Zu den wichtigsten Arbeitsinstrumenten der Digital Humanities gehören die algorithmische Verarbeitung von Daten, die statistische Modellierungen von Zusammenhängen in Daten und visuelle Analysemethoden, um Daten verstehen zu können. Diese Instrumente folgen bestimmten Routinen wissenschaftlicher Praxis: Sie verwenden erprobte Algorithmen und halten sich an Standards der Datenmodellierung. Gleichzeitig konstituieren sie aber auch die wissenschaftliche Praxis mit – sie sind, um mit Ludwick Fleck zu sprechen, Mittel zur Profilierung eines Denkstils innerhalb eines wissenschaftlichen Denkkollektivs. Dazu gehören nicht nur die erwähnten Arbeitsinstrumente, sondern auch sprachliche Mittel: Fachbegriffe, Metaphern, Stile.

Um die Verknüpfung von Arbeitsinstrumenten und wissenschaftlichen Routinen genauer zu verstehen, müssen die Praktiken und Kulturen, in denen diese Instrumente erstellt werden, reflektiert werden. Welchen Einfluss hat die Wahl einer bestimmten Programmiersprache zur Implementierung eines Algorithmus auf die Digital-Humanities-Praxis? Beispielsweise die Verwendung des „postmodernen“ Perl (Wall 1999) statt Python? Welchen Einfluss hat die Programmiersprache auf die Art der algorithmisch erstellen Visualisierung? Beispielsweise die Verwendung von D3.js, P5.js, R oder der Software Excel (Bubenhofer 2015)? Welchen wissenschaftlichen Paradigmen entspringen populäre statistische Modellierungen? Beispielsweise Topic Modelling oder Support Vector Machines? Und mit welchen sprachlichen Mitteln werden die angewandten Instrumente in den wissenschaftlichen Diskurs eingebracht und legitimiert?

Wissenschaftsgeschichtliche Ansätze von Ludwik Fleck (Fleck 1983, 2011) oder auch Thomas S. Kuhn (Kuhn 1996) lassen sich fruchtbar verknüpfen mit Überlegungen der Software Studies (Fuller 2003; Mackenzie 2006; Manovich 2013; Cox / McLean 2012), die den kulturellen Kontext und die soziale Praxis als Einflussfaktoren von Software-Erstellung und -Nutzung betonen. Ebenso existiert eine Diskussion um die Rolle von Algorithmen, statistischen Modellierungen oder

generell Software-„Tools“ in Forschungsprozessen der Digital Humanities (Berry 2014; Bubenhofer / Scharloth 2015; Kath et al. 2015; Rieder / Röhle 2012). Die reiche Praxis der Informationsvisualisierung und Visual Analytics für Fragen der Digital Humanities führt ebenso nicht nur zu methodischen, sondern auch methodologischen und theoretischen Diskussionen (Chen et al. 2008; Keim et al. 2010). Auch die Rolle von Denkstilen in Wissenschaftsdiskursen und ihre Manifestation auf sprachlicher Ebene wird in neuerer Zeit intensiver reflektiert (Czachur 2013; Fix 2011; Schiewe 1996).

Der Vortrag möchte vor diesem Hintergrund Code, Modelle, Visualisierungen und Sprache als Mittel und Instrument im Kontext wissenschaftlicher Routinen in den Digital Humanities reflektieren. Inwiefern drücken sich in den gewählten Programmiersprachen, Algorithmen, Visualisierungstypen, statistischen Modellen und sprachlich gefassten Interpretationen unterschiedliche Denkstile der Digital Humanities aus? Wo liegen die Chancen, aber auch die Gefahren, diese Denkstile zu reproduzieren? Welche Auswirkungen haben die Wahl und der reflektierte oder nicht reflektierte Umgang mit den Instrumenten auf wissenschaftliche Innovation in den Digital Humanities?

Dazu wird zunächst der Einsatz und die Typen von Visualisierungen in den textorientierten Digital Humanities analysiert und dann die technischen aber auch kulturellen Entstehungsbedingungen der algorithmischen Visualisierungen untersucht.

Bibliographie

Berry, David M. (2014): *Critical Theory and the Digital*. London, Oxford, New York, New Delhi, Sydney: Bloomsbury.

Bubenhofer, Noah (2015): "Coding Cultures: Über den Zusammenhang von Programmiersprachen und Denkstilen", in: *Sprechtakel*. Linguistische Notizen https://www.bubenhofer.com/sprechtakel/2015/08/08/coding-cultures-ueber-den-zusammenhang-von-programmiersprachen-und-denkstilen/ [letzter Zugriff 17. August 2015].

Bubenhofer, Noah / Scharloth, Joachim (2015): "Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate", in: *Zeitschrift für Germanistische Linguistik* 43, 1: 1–26.

Chen, Chun-houh / Härdle, Wolfgang / Unwin, Antony (eds.) (2008): *Handbook of data visualization* (= Springer handbooks of computational statistics). Heidelberg: Springer.

Cox, Geoff / McLean, Alex (2012): *Speaking Code*. Coding as Aesthetic and Political Expression. Cambridge, Mass.

Czachur, Waldemar (2013): "Ludwik Flecks Denkstilansatz als Inspiration für die Diskurslinguistik", in: *Zeitschrift des Verbandes Polnischer Germanisten* 2: 141–150.

Fix, Ulla (2011): *Denkstile und Sprache*. Die Funktion von „Sinn-Sehen“ und „Sinn-Bildern“ für die „Entwicklung einer wissenschaftlichen Tatsache“ http://home.uni-leipzig.de/fix/Fleck.pdf [letzter Zugriff 04. März 2014].

Fleck, Ludwik / Werner, Sylwia / Zittel, Claus (eds.) (2011): *Denkstile und Tatsachen*: Gesammelte Schriften und Zeugnisse. Berlin: Suhrkamp.

Fleck, Ludwik / Schäfer, Lothar / Schnelle, Thomas (eds.) (1983): *Erfahrung und Tatsache: gesammelte Aufsätze*. Frankfurt am Main: Suhrkamp.

Fuller, Matthew (2003): *Behind the blip: essays on the culture of software*. New York: Autonomedia.

Kath, Roxana / Schaal, Gary S. / Dumm, Sebastian (2015): "New Visual Hermeneutics", in: *Zeitschrift für germanistische Linguistik* 43, 1: 27–51.

Keim, Daniel A. / Kohlhammer, Jörn / Ellis, Geoffrey et al. (2010): *Mastering the information age* - solving problems with visual analytics. Goslar: Eurographics Association.

Kuhn, Thomas S. (1996): *Structure of Scientific Revolutions*. University of Chicago Press.

Mackenzie, Adrian (2006): *Cutting Code: Software And Sociality* (Digital Formations). Bern / Berlin / Frankfurt am Main / New York / Paris / Wien: Peter Lang.

Manovich, Lev (2013): *Software Takes Command*. New York / London: INT edition.

Rieder, Bernhard / Röhle, Theo (2012): "Digital Methods: Five Challenges", in: Berry, David M. (eds.): *Understanding Digital Humanities*. Basingstoke: Palgrave 67–84.

Schiewe, Jürgen (1996): *Sprachenwechsel - Funktionswandel - Austausch der Denkstile*. Die Universität Freiburg zwischen Latein und Deutsch. Tübingen: Niemeyer.

Wall, Larry (1999): *Perl, the first postmodern computer language* http://www.wall.org/~larry/pm.html [letzter Zugriff 19. August 2015].

Digital Humanities in Bewegung: Ansätze für die computergestützte Filmanalyse

Burghardt, Manuel
manuel.burghardt@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Wolff, Christian
christian.wolff@ur.de

Lehrstuhl für Medieninformatik, Universität Regensburg

Einleitung: Digital Humanities und Filmanalyse

Während sich die „Vermessung der Kultur“ (Lauer 2013) in den textorientierten Geisteswissenschaften in den letzten Jahren rasant entwickelt hat (vgl. etwa Konzepte wie *Culturomics*, *Distant Reading*, etc.), so befindet sich die „Vermessung ästhetischer Erscheinungen“ (Flückinger 2011) für den Bereich der Filmwissenschaft und Filmanalyse noch in den Anfängen. Flückinger (2011: 44) spricht in diesem Zusammenhang gar von einem Spannungsfeld zwischen Empirie und Ästhetik, welches sich zwangsläufig ergeben muss, wenn man „die eigentümliche Unschärfe, die allen künstlerischen Werken eignet, in messbare Einheiten zerlegen will“. Dabei lassen sich quantitative Ansätze in der Filmanalyse mindestens bis in das Jahr 1912 [1] zurückverfolgen und auch aktuelle Lehrbücher zur Filmanalyse beschreiben gleichermaßen *weiche* (qualitative) und *harte* (quantitative) Kategorien und Methoden (Korte 2004: 15). Bei quantitativen Ansätzen steht vor allem die Analyse von Dauer und Auftretenshäufigkeit einzelner Einstellungen in einem Film im Mittelpunkt (vgl. Salt 2006, Kap. „The Numbers Speak“). So stellt etwa die online verfügbare Datenbank *Cinemetrics* (Cinemetrics o. J.) entsprechende Informationen zur Länge und Verteilung einzelner Einstellungen für mehrere tausend Filme bereit und ermöglicht so vergleichende Analysen von Filmen aus unterschiedlichen Genres und Epochen.

Während die Segmentierung der Filme in der *Cinemetrics*-Datenbank von der Community manuell vorgenommen wird, gibt es auch Beispiele für Forschungsarbeiten, bei denen die quantifizierbaren Parameter automatisch erhoben werden. Hoyt, Ponot und Roy (2014) präsentieren etwa einen Prototyp namens *ScripThreads*, der in der Lage ist, Filme der *American Film Scripts Online*-Datenbank zu parsen und die Handlungsentwicklung eines Films anhand der Szenen und Figuren zu visualisieren. Ein Beispiel für die vergleichende Analyse von Filmmetadaten findet sich im *Cinegraph*-Projekt von Chris Weaver (2014). Hier können Filme anhand unterschiedlicher Metadaten (z. B. Filmname, Veröffentlichungsdatum, Bewertung, Genre, Oscars, Darsteller, etc.) miteinander verglichen und in einer interaktiven Darstellung zueinander in Beziehung gesetzt werden.

Daneben finden sich im Netz eine ganze Reihe experimenteller Tools, die nicht immer einen wissenschaftlichen Anspruch haben, aber gut illustrieren, welche weiteren Aspekte von Filmen automatisch analysierbar sind: Beispielhaft sei etwa das Python-Tool *VideoGrep* (Lavigne 2014) genannt, welches das Durchsuchen von Filmdialogen nach bestimmten Schlüsselwörtern ermöglicht, um auf Basis der Treffer dann einen automatischen Zusammenschnitt („supercut“) all der Szenen, in denen das gesuchte Wort vorkommt, zu erstellen. Die Anwendung *Pretentious-O-Meter* (Beard 2015) analysiert automatisch, wie groß die Bewertungslücke zwischen Nutzerbewertungen und professionellen Filmkritiken eines Films ist und visualisiert dies in einem Kontinuum, welches von „mass-market“ bis „very pretentious“ reicht. Weitere Ansätze der automatischen Filmanalyse finden sich für die Farbverwendung in Filmen: Frederic Brodbeck (2011) visualisiert in seinem Filme als kreisförmig angeordnete Timelines, in denen u. a. die jeweils dominanten Farben zu sehen sind. Ein weiteres Projekt visualisiert Filme als zusammengestauchte Einzelframes, um so farbige *MovieBarcodes* (MovieBarcodes o. J.) zu erstellen.

Auch auf der DHd 2015 wurde das Thema der Quantifizierung filmischer Strukturen über Filmbild, Filmschnitt und Filmstil bereits auf methodischer Ebene thematisiert (Heftberger 2015) und Howanitz (2015) präsentierte eine erste *Distant Watching*-Studie für das „Fern-Sehen“ memetischer YouTube-Videos, deren „Schnittkurven“ er auf Frame-Ebene analysiert. In diesem Beitrag knüpfen wir thematisch an die genannten DHd-Vorträge an und diskutieren grundlegende Möglichkeiten der computergestützten Filmanalyse, die über die Quantifizierung von Einstellungen und Szenen hinausgehen. Dabei sollen weitere automatisch quantifizierbare Parameter zur Diskussion gestellt werden, um so neue Perspektiven und Zugänge zur computergestützten Filmanalyse aufzuzeigen und das Thema noch stärker in den Digital Humanities zu verankern. Um die Grenzen und Möglichkeiten dieser Ansätze besser illustrieren zu können, wurde eine Reihe von Prototypen erstellt, die nachfolgend kurz vorgestellt werden.

Prototypen für die computergestütze Filmanalyse

In diesem Abschnitt werden drei unterschiedliche Prototypen beschrieben, die jeweils auf unterschiedliche quantifizierbare Aspekte von Filmen abzielen und damit die Untersuchung ganz unterschiedlicher Fragestellungen erlauben. Die Tools greifen allesamt auf im Web frei verfügbare Informationen zu Filmen zurück: So stehen etwa über die Plattformen *OpenSubtitles* oder die *Internet Script Movie Database* maschinenlesbare Dialoge von Filmen und Serien in großem Umfang zur Verfügung. Zusätzlich können detaillierte Metadaten sowie auch nutzergenerierte Bewertungen und Kommentare zu Filmen über Plattformen wie *IMDb* (Internet Movie Database) abgerufen werden. Darüber hinaus soll als weiterer quantifizierbarer Parameter, der direkt aus den Filmen extrahiert werden kann, die Farbverwendung [2] in die Analysen mit einbezogen werden. Alle nachfolgend beschriebenen Prototypen wurden jeweils mit Standard-

Webtechnologien (HTML / CSS / JavaScript) und bestehenden Python-Bibliotheken umgesetzt.

SubVis – Analyse der Filmsprache

Das *SubVis*-Tool analysiert über *OpenSubtitles* verfügbare Dialoge von beliebigen, zunächst allerdings nur englischsprachigen Filmen anhand typischer linguistischer Parameter wie Wortfrequenzen oder POS-Tagging und visualisiert die Ergebnisse in einem interaktiven Web-Interface. Zusätzlich kann die Auftretenshäufigkeit einzelner Zeichen oder längerer Sprachsequenzen (= jeweils ein eingeblendeten Untertitel) für beliebig definierbare Analyseintervalle (z. B. jeweils für 5 Minuten-Sequenzen) in einer Timeline dargestellt werden, um bspw. auf einen Blick zu sehen, an welchen Stellen im Film besonders viel oder wenig gesprochen wird (vgl. Abbildung 1).

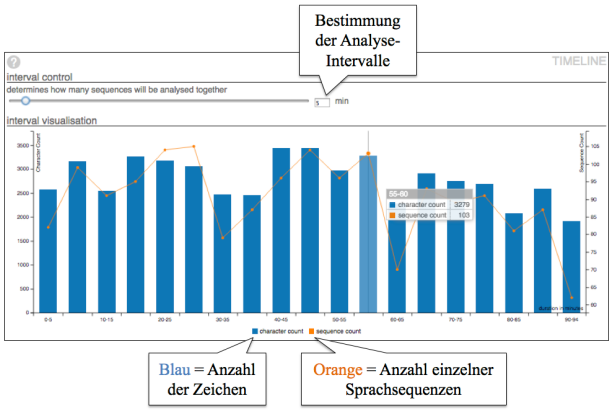


Abb. 1: Beispielhafte Visualisierung der Zeichen- und Sprachsequenzhäufigkeiten für jeweils fünfminütige Teilabschnitte des Films „Anchorman: The Legend of Ron Burgundy“.

Beispielhafte Fragestellungen, die mit dem Tool untersucht werden können:

- Gibt es für die Filme unterschiedlicher Regisseure jeweils typische Schlüsselwörter?
- Kann man für Filme aus unterschiedlichen Genres beobachten, dass an bestimmten Stellen (z. B. Anfang oder Schluss) besonders viel oder wenig gesprochen wird?
- Wird in Filmen aus den 1980er Jahren insgesamt mehr gesprochen als in Filmen der 1990er Jahre?

Series Analysis Tool (SAT) – Analyse von TV-Serien anhand von Nutzerbewertungen, Figuren und Sprache

Das *Series Analysis Tool* (*SAT*) ermöglicht die Analyse von Serien und einzelnen Episoden. Dabei werden verschiedene Parameter in einer Timeline-Darstellung visualisiert. Ein wesentlicher Analyseaspekt ist dabei die Bewertung einzelner Episoden durch die IMDb-Community, sodass auf einen Blick erkennbar ist, ob eine Serie im Laufe der Zeit besser oder schlechter bewertet wird, oder ob es einzelne Episoden gibt, die auffallend positiv oder negativ bewertet wurden. Zusätzlich liest das Tool das Figureninventar für jede Episode aus und erlaubt es, die Darstellung nach bestimmten Figuren zu filtern. So kann schnell erkannt werden, ob das Auftreten bestimmter Figuren ggf. Einfluss auf die Bewertung einzelner Episoden hat. Weiterhin wurde die Sprache der Serien hinsichtlich Sentiment- und Emotionswörtern analysiert (vgl. Abbildung 2). Als Datengrundlage dient ein bestehendes Korpus (Tiedemann 2012), in dem alle auf *OpenSubtitles* in englischer Sprache verfügbaren Untertitel von TV-Serien und Filmen bis zum Jahr 2013 enthalten sind. Dabei kam für die Sentiment Analyse das *AFINN*-Lexikon (Nielsen 2011) und für die Identifikation acht grundlegender Emotionen (Angst, Wut, Freude, etc.) das *NRC Emotion Lexicon* (Mohammad / Turney 2010) zum Einsatz. Sowohl die Sentiment-Scores (positiv / negativ) als auch die Emotionsmarker können für jede Episode in die Visualisierung mit einbezogen werden, um so potenzielle Korrelationen zu den Nutzerbewertungen aufzuzeigen.

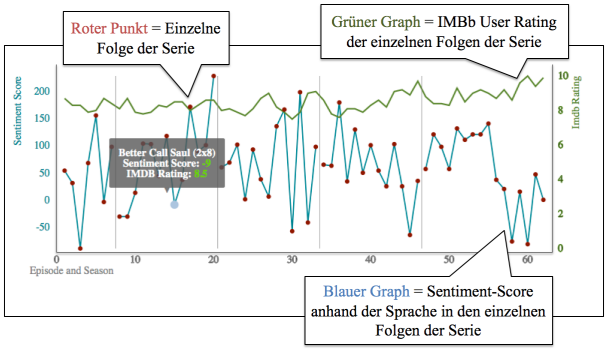


Abb. 2: Beispielhafte Visualisierung der Serie „Breaking Bad“, mit paralleler Darstellung der Benutzerbewertungen sowie der Sentiment-Analyse der Dialoge für jede einzelne Episode.

Beispielhafte Fragestellungen, die mit dem Tool untersucht werden können:

- Gibt es generelle Trends bei der Bewertung von Serien mit zunehmender Zahl von Staffeln?
- Wirken sich Sentiment- und Emotionsmarker der Dialoge positiv oder negativ auf die Bewertung einer Episode aus?

- Wirkt sich das Auftreten bestimmter Nebenfiguren positiv oder negativ auf die Bewertung einer Episode aus?

MovieColors – Analyse von Filmen anhand von Farbe und Sprache

Der Prototyp *MovieColors* erlaubt die computergestützte Analyse von Filmen anhand der Parameter Farbe und Sprache. Dabei wird zunächst der Film in einzelne Frames zerlegt. Mithilfe eines Clustering-Algorithmus werden dann die jeweils dominanten Farben extrahiert. Anhand dieser Farbinformation können charakteristische Farbprofile – ähnlich wie im eingangs erwähnten *MovieBarcodes*-Projekt – für den gesamten Film erstellt werden. Zusätzlich wird die Sprache des Films über dessen Untertitel anhand von Wortfrequenzen und grundlegenden Sentiment-Werten (positiv / negativ) analysiert.

Die Visualisierungskomponente des Tools erlaubt es, Farbinformation und Sprachanalyse in einer parallelen Ansicht darzustellen, um so potenzielle Korrelationen zwischen dem Sentiment der Sprache und besonders markanten Schlüsselwörtern sowie auch der Farbverwendung identifizieren zu können (vgl. Abbildung 3). Zusätzlich kann jeder Frame einzeln angezeigt werden, zusammen mit dem entsprechenden Untertitel sowie einer Analyse der dominanten Farben im jeweiligen Bild.

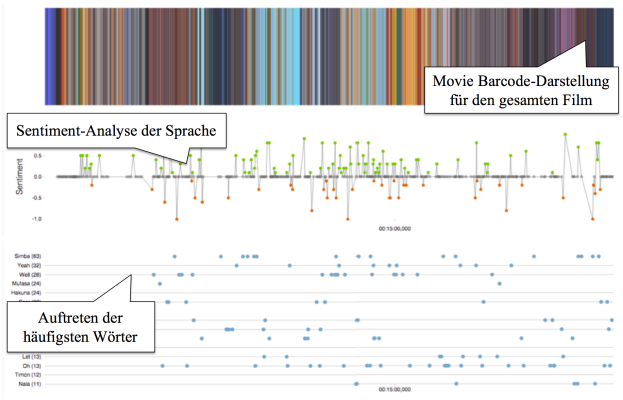


Abb. 3: Analyse des Films „König der Löwen“, mit Darstellung des Farbprofils (oben), der Sentiment-Analyse (Mitte) sowie der häufigsten Wörter (unten) entlang der Zeitachse des Films.

Beispielhafte Fragestellungen, die mit dem Tool untersucht werden können:

- Gibt es charakteristische Farbprofile für Filme aus verschiedenen Genres oder Epochen?
- Korrelieren bestimmte Farben mit positiven oder negativen Sentiment-Scores, also etwa dunkle Farben bei negativer Sprache?
- Korrelieren bestimmte Farben mit Schlüsselwörtern, also etwa schwarz und lila immer dann, wenn der Bösewicht des Films auftritt?

Ausblick

Die in diesem Beitrag vorgestellten Prototypen beschreiben erste Versuche, Filme computergestützt anhand unterschiedlicher, automatisch quantifizierbarer Parameter zu analysieren. Im Austausch mit Kollegen aus der Medienwissenschaft werden die Tools in den nächsten Monaten praktisch erprobt und je nach Fragestellung iterativ angepasst und gegebenenfalls um weitere Funktionen ergänzt. Sobald die Prototypen weiter ausgearbeitet sind, sollen sie auch der Community über den DH-Regensburg-Blog zugänglich gemacht werden. Gleichzeitig sind weitere Prototypen angedacht, bei denen als zusätzliche Analyseparameter Gesichtserkennung (vgl. Arandjelovic / Zisserman 2005) sowie auch die Auswertung der Audiospur (vgl. *Zulko* 2014) umgesetzt werden sollen.

Danksagungen

Alle hier beschriebenen Prototypen wurden im Rahmen des Projektseminars „Digital Humanities“, im Masterstudiengang Medieninformatik an der Universität Regensburg, angefertigt. Besonderer Dank für die engagierte Umsetzung der Tools gebührt Hanns Meißner und Michael Stahl (*SubVis*), Robert Jackermeier, Florian Ludwig und Alexander Uitz (*SAT*) sowie Michael Kao (*MovieColors*).

Notes

1. Vgl. den Vortrag von Tsivian (2014) auf der 1. Cinemetrics Conference, Chicago (Neubauer Collegium 2014).
2. Zur historischen Verwendung von Farbe im Film vgl. auch die Online-Datenbank "Timeline of Historical Film Colors" von Flückinger (2011-2013).

Bibliographie

Arandjelovic, Ognjen / Zisserman, Andrew (2005): "Automatic face recognition for film character retrieval in feature-length films", in: *Proceedings of the Computer Vision and Pattern Recognition Conference (IEEE)* 860-867.

Beard, Niall (2015): Pretentious-O-Meter http://pretentious-o-meter.co.uk/ [letzter Zugriff 04. Februar 2016].

Brodbeck, Frederic (2011): *Cinemetrics*. Bachelor graduation project at the Royal Academy of Arts (KABK), Den Haag [letzter Zugriff 04. Februar 2016].

Cinemetrics (o. J.): http://www.cinemetrics.lv/ [letzter Zugriff 04.Februar 2016]

Flückiger, Barbara (2011): "Die Vermessung ästhetischer Erscheinungen", in: *Zeitschrift für Medienwissenschaft* 5, 2: 44-60.

Flückinger, Barbara (2011-2013): *Timeline of Historical Film Colors* http://zauberklang.ch/filmcolors/ [08. Januar 2016].

Heftberger, Adelheid (2015): "Filmbild , Filmschnitt , Filmstil – die Quantifizierung und Visualisierung von filmischen Strukturen", in: *Book of Abstracts, DHd 2015*.

Howanitz, Gernot (2015): „Distant Waching: Ein quantitativer Zugang zu YouTube-Videos“, in: *Book of Abstracts, DHd 2015*.

Hoyt, Eric / Ponot, Kevin / Roy, Carrie (2014): „Visualizing and Analyzing the Hollywood Screenplay with ScripThreads“, in: *Digital Humanities Quarterly* 8, 4.

IMDb (o. J.): *Internet Movie Database*. http://www.imdb.com/ [letzter Zugriff 04. Februar 2016].

IMSDb (o. J.): *Internet Script Movie Database*. http://www.imsdb.com/ [letzter Zugriff 04. Februar 2016].

Korte, Helmut (2004): *Einführung in die Systematische Filmanalyse*. Berlin: Erich Schmid Verlag.

Lauer, Gerhard (2013): "Die digitale Vermessung der Kultur", in: Geiselberger, Heinrich / Moorstedt, Tobias (eds.): *Big Data – Das neue Versprechen der Allwissenheit*. Berlin: Suhrkamp 99-116.

Lavigne, Sam (2014): *Videogrep*. Automatic Supercuts with Python http://lav.io/2014/06/videogrep-automatic-supercuts-with-python/ [letzter Zugriff 04. Februar 2016].

Mohammad, Saif M. / Turney, Peter D. (2010): "Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon", in: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* 26-34.

MovieBarcode (o. J.): http://moviebarcode.tumblr.com/ [letzter Zugriff 04. Februar 2016].

Neubauer Collegium (2014): *UChicago Cinemetrics Conference* https://www.youtube.com/watch?v=6ZXj67bygEc [letzter Zugriff 04. Februar 2016].

Nielsen, Finn Å. (2011): "A new ANEW: evaluation of a word list for sentiment analysis in microblogs", in: *Proceedings of the ESWC2011 Workshop on „Making Sense of Microposts: Big things come in small packages*“ 93-98.

OpenSubtitles (o. J.) [letzter Zugriff 04. Februar 2016].

Salt, Barry (2006): *Moving into Pictures*. London: Starwood.

Tiedemann, Jörg (2012): "Parallel Data, Tools and Interfaces in OPUS", in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* 2214-2218.

Weaver, Chris (2014): *Cinegraph* http://www.cs.ou.edu/~weaver/improvise/examples/cinegraph/index.html [letzter Zugriff 04. Februar 2016].

Zulko (4.7.2014): "Automatic Soccer Highlights Compilations With Python" (Blogpost), in: *__del__ (self)* Eaten by the Python http://zulko.github.io/blog/2014/07/04/automatic-soccer-highlights-compilations-with-python/ [letzter Zugriff 08. Januar 2016].

Die Kunst als Ganzes. Heterogene Bilddatensätze als Herausforderung für die Kunstgeschichte und die Computer Vision.

Dieckmann, Lisa

lisa.dieckmann@uni-koeln.de
Universität zu Köln, Kunsthistorisches Institut

Bell, Peter

bell@uni-heidelberg.de
Ruprecht-Karls-Universität Heidelberg, Heidelberg Collaboratory for Image Processing (HCI)

Den aufwändig nach bestimmten Schemata modellierten Datensätzen fachspezifischer Bilddatenbanken und der Verwendung von Ontologien, Thesauri und Normdaten steht die anschwellende Bilderflut des digitalen und vernetzten Zeitalters gegenüber. Auch im Bereich kulturhistorischer Artefakte liegen viele digitale Abbildungen in völlig unterschiedlichen Formen, mal mit Text und Metadaten, mal mehr oder weniger strukturiert und erschlossen vor. Dieser positiven Entwicklung einer Demokratisierung des Forschungsmaterials durch die digitale, vernetzte und a-hierarchische Bereitstellung, muss jedoch mit Verfahren begegnet werden, welche die heterogenen Bilder nach bestimmten Kriterien automatisiert erschließen und strukturieren und das Retrieval im Sinne von Precision und Recall optimieren.

Mit welchen Verfahren lässt sich die große Menge von Bildinhalten auffindbar und erschließbar machen? Wie kann der Heterogenität der Bild- und Metadaten begegnet werden? Und: Wie kann dennoch der Individualität von historisch-hermeneutischer Forschung Rechnung getragen werden? Neben automatisierten textanalytischen

Verfahren, mit denen die Daten strukturiert, kategorisiert oder auch angereichert werden, können Bilder semantisch über rein bildanalytische Verfahren erschlossen werden. Darin besteht die größere und eigentliche Herausforderung. Denn das Bild ist im Vergleich zum Text schwerer in sinnhaltige und abgrenzbare Entitäten zerlegbar und seine Ikonographie nicht ohne weiteres über visuelle automatisierte Verfahren rekonstruierbar. Darin liegt der „semantic gap".

Der Vortrag thematisiert die Erschließung von größeren und nur teilweise strukturierten Mengen von Bildern und Metadaten durch automatisierte Verfahren. Der Schwerpunkt liegt hierbei auf Verfahren zur visuellen Analyse. Anhand des heterogenen und verteilten Bilderpools des prometheus-Bildarchivs (prometheus 2001-2016) wird derzeit mit der Computer Vision Group der Universität Heidelberg (CompVis 2015) ein Projekt vorbereitet, in welchem die bereits sehr elaborierten Ansätze Anwendung finden.

Das prometheus-Bildarchiv (Universität zu Köln, Kunsthistorisches Institut) ist ein verteiltes digitales Bildarchiv, das über 80 strukturell und inhaltlich verschiedene Bilddatenbanken der Kunstgeschichte, Archäologie und weiterer bildbasierter Disziplinen und damit über 1,4 Mio. Bilder miteinander verknüpft.

Die Computer Vision Group des Heidelberg Collaboratory for Image Processing an der Ruprecht-Karls-Universität Heidelberg widmet sich der Grundlagenforschung zum automatischen Bildverstehen. Sie entwickelt Algorithmen zur Erschließung von Bildbestandteilen (Segmentierung), diskriminativer Objekterkennung und Szenenvergleich. Sie hat bereits in mehreren Projekten mit der Kunstgeschichte zusammengearbeitet.

Im vorzustellenden Projekt sollen Algorithmen zum automatischen Sehen und maschinellen Lernen speziell zur Erschließung und zur Recherche an kulturellem Erbe optimiert und entwickelt werden. Komplementär zur üblichen Textsuche werden die NutzerInnen des Bildarchivs nach Bildern und Bildpartien suchen können. Darüber hinaus sollen die Bilder aufgrund verschiedener Ähnlichkeiten sortiert und vorgeschlagen werden.

Für die Computer Vision bietet das prometheus-Bildarchiv ein großes Potential und viele skalierbare Aufgaben, da viele der Bilder in einem stilistischen oder semantischen Zusammenhang stehen, der sich visuell oder über die Metadaten erschließen lässt. Es ist somit keine Anwendungsaufgabe, sondern dezidiert informatische Grundlagenforschung. Bisher existieren nur wenige Versuche und Prototypen, wobei es sich im Wesentlichen um die Adaption von bestehenden Computer Vision Ansätzen auf kunsthistorische Spezialprobleme handelt.

Die skizzierte automatische Bildanalyse geht verschiedene Problemfelder an. Zum einen kann die Abbildung als Ganzes analysiert werden. Dabei sollen nicht nur Duplikate, sondern auch Reproduktionen und Variationen eines Kunstwerks aufgefunden werden, wie z. B. Kunstwerke, die im Laufe der Geschichte immer wieder rezipiert wurden (Laookon-Gruppe, Apoll von Belvedere). Die automatische Bildanalyse muss daher verschiedene Dimensionen von Ähnlichkeit im Blick behalten.

Es können aber auch einzelne Bildpartien und Motive eines Bildes analysiert werden. In vielen Fällen sind diese Motive nicht verschlagwortet (Kreuz, Schädel, Pferd) und erst die visuelle Analyse der Bildinhalte liefert die gewünschten Ergebnisse. Auch die spezifische Formbehandlung oder der Duktus des Künstlers – Phänomene, die quasi nie textlich beschrieben werden, können so in einen Zusammenhang gebracht werden. Schließlich kann die Bildanalyse auch syntaktische Unebenheiten und variierende Schreibweisen (Bildtitel, Künstlernamen) ausgleichen. Als alternative Suchstrategie kann die Bildsuche diese Übersetzungsschwierigkeiten umgehen.

Ein Kernproblem ist die methodische Herausforderung, Ähnlichkeit in verschiedenen Dimensionen zu definieren. Hier bieten sich einerseits zahlreiche stilkundliche und kennerschaftliche Ansätze an, die jedoch in ihrer Reichweite umstritten sind und noch nicht auf diese Weise technisch evaluiert wurden. Andererseits lassen sich semantische Taxonomien (ICONCLASS) und kontextbezogene Metadaten hinzuziehen, deren Verhältnis zum visuellen Befund erst bestimmt werden muss. Daraus ergibt sich auch ein großer Bedarf an methodischer Reflektion und interdisziplinärer Diskussion zwischen den beiden Bildwissenschaften. Weiterhin sind technische Schwierigkeiten durch die sehr große Bildmenge und die wissenschaftlichen Standards entsprechende tiefe Erschließung gegeben. Dies kann nur gewährleistet werden, wenn die visuelle Suchanfrage prägnant genug ist, aber auch die Variationsweite definiert wird. Da dies nur in wenigen Fällen mit einem Suchbereich funktioniert, muss ein interaktives Modell geschaffen werden, in dem sich Mensch und Maschine dialogisch dem gewünschten Ergebnis nähern. Die verschiedenen Dimensionen von Ähnlichkeit und kontextueller Bezogenheit sind ein anspruchsvolles Grundlagenforschungsproblem der Computer Vision, das nicht mit vorgefertigten Algorithmen oder bestehenden Usecases adaptiv gelöst werden kann.

Die NutzerInnen sollen mehrere Suchfelder markieren und kombinieren können, sowie die Ergebnisse bewerten, um mit Positiv- und Negativbeispielen eine neue Suche auszulösen. Dazu bedarf es eines ergonomischen Userinterfaces, in dem die NutzerInnen die neuartige Suche leicht erlernen und nutzen können. Der Algorithmus lernt fortwährend an den durchgeführten Suchen und Feedbacks, welche Suchen für die NutzerInnen relevant sind und erschließt dadurch kontinuierlich den Bilddatensatz.

Computer Vision kann also, indem direkt auf die Bildinformationen zugegriffen wird, Beschreibungen vornehmen und Verbindungen zwischen Kunstwerken aufzeigen, die vom menschlichen Auge nicht oder nur unter größtem Zeitaufwand gesehen werden können.

Als Vorschlagssystem findet es nicht nur genau die Partien nach denen der Anwender sucht, sondern weist auch auf Bilder mit ähnlichen Formen, Texturen oder Farbwerten hin und visualisiert die ähnlichen Bilder in Form übersichtlicher Synopsen. Die Arbeit mit digitalen Bildrepositorien wird nicht nur effektiver, sondern auch assoziativer, durch das Vorschlagen neuer Verbindungen zwischen Kunstwerken. Dieser visuelle Zugang zu wissenschaftlichen Bilddatenbanken bereichert nicht nur die historischen Bildwissenschaften um ein Forschungsinstrument, sondern erleichtert auch den benachbarten Disziplinen (Archäologie, Kunstpädagogik etc.) einen intuitiven Zugriff auf das Material jenseits fachlicher Terminologie.

Bibliographie

Bell, Peter / Ommer, Björn (2015): "Training Argus, Ansätze zum automatischen Sehen in der Kunstgeschichte", in: *Kunstchronik* 68, 8: 414-420.

Bell, Peter / Ommer, Björn / Schlecht, Joseph (2013): "Nonverbal Communication in Medieval Illustrations Revisited by Computer Vision and Art History", in: *Visual Resources*. An International Journal of Documentation 29, 1-2: 26-37.

Bell, Peter / Dieckmann, Lisa / Ommer, Björn / Takami, Masato (2015): "Passion Search. Prototype of an unrestricted image search of the crucifixion", in: http://hci.iwr.uni-heidelberg.de/COMPVIS/projects/suchpassion/ [letzter Zugriff 05. Januar 2016]

Chung, Joon Son / Arandjelović, Relja / Bergel, Giles/ Franklin, Alexandra/ Zisserman, Andrew (2014): "Re-presentations of Art Collections", in: Agapito, Lourdes / Bronstein, Michael M. / Rother, Carsten (eds.): *Computer Vision - ECCV 2014 Workshops.* Heidelberg / New York / Dordrecht / London: Springer 85-100.

CompVis(2015): *Computer Vision Group*. Ruprecht-Karls-Universität Heidelberg http://hci.iwr.uni-heidelberg.de/COMPVIS/ [letzter Zugriff 08. Januar 2016].

Crowley, Elliot J. / Zisserman, Andrew (2014): "In Search of Art", in: Agapito, Lourdes / Bronstein, Michael M. / Rother, Carsten (eds.): *Computer Vision - ECCV 2014 Workshops.* Heidelberg /New York / Dordrecht / London: Springer 54-70.

Dieckmann, Lisa (2015): "prometheus – das verteilte digitale Bildarchiv für Forschung & Lehre e. V.", in: Euler, Ellen / Hagedorn-Saupe, Monika/ Maier, Gerald/ Schweibenz, Werner/ Sieglerschmidt, Jörn (eds.): *Handbuch Kulturportale*. Online-Angebote aus Kultur und Wissenschaft. Berlin / Boston: DeGruyter 223-229.

Johnson, C. Richard Jr. / Wang, James Z. et al. (2008): "Image Processing for Artist Identification - Computerized Analysis of Vincent van Gogh's Painting Brushstrokes", in: *IEEE Signal Processing Magazine* 25, 4: 37-48.

Kohl, Jeanette / Srinivasan, Ramya / Roy-Chowdhury, Amit / Rudolph, Conrad (2013): "Quantitative Modeling of Artists Styles in Renaissance Face Portraiture", in: Second International Workshop on Historical Document Imaging and Processing http://www.ee.ucr.edu/~amitrc/publications/icdar2013.pdf [letzter Zugriff 05. Januar 2016]

prometheus (2001-2016)*prometheus – Das verteilte digitale Bildarchiv für Forschung & Lehre*. Kunsthistorisches Institut der Universität zu Köln http://prometheus-bildarchiv.de/index [letzter 08. Januar 2016].

Vaughan, William (1997) "Computergestützte Bildrecherche und Bildanalyse", in: Hubertus Kohle (ed.): *Kunstgeschichte digital*. Eine Einführung für Praktiker und Studierende. Berlin: Reimer 97-105.

Die Corpusanalyse multimodaler Erzählungen am Beispiel graphischer Romane

Dunst, Alexander
alexander.dunst@gmail.com
Universität Paderborn, Deutschland

Hartel, Rita
rst@upb.de
Universität Paderborn, Deutschland

Einleitung

Dieser Vortrag präsentiert Analysen und Visualisierungen eines derzeit im Aufbau befindlichen Corpus an graphischen Romanen (oder „Graphic Novels“, einer Unterform des Medium Comics) und stellt den für die Annotation dieser multimodalen Erzählform entwickelten Editor vor, der zeitgerecht zur DHd-Jahrestagung in Leipzig für den Download zur Verfügung stehen wird. Während sich die Analyse literarischer Text-Corpora bereits seit mehreren Jahren im Fokus der Digitalen Geisteswissenschaften befindet, stehen Bestrebungen zur Erforschung visueller Erzählformen wie Theater, Comics, Film, Fernsehen, sowie Computerspiele, oft eine Randerscheinung in den DH dar und vor einer Reihe ungelöster Herausforderungen. Diese bestehen sowohl in technischer - etwa in Bezug auf die automatisierte Erkennung visueller Objekte und die Annotation komplexer Bild-Text- Kombinationen – als auch in methodischer Hinsicht. Angesichts der Dominanz visueller Erzählformen seit dem frühen 20. Jahrhundert,

sowie noch verstärkt in der Gegenwartskultur, stellt dies eine außerordentliche Forschungslücke dar.

In einer kurzen Einleitung wird der Vortrag den derzeit im Aufbau befindlichen Corpus sowie die Zielsetzungen der vom deutschen Bundesministerium für Bildung und Forschung (BMBF) geförderten Nachwuchsgruppe „Hybride Narrativität: Digitale und Kognitive Methoden zur Erforschung Graphischer Literatur“ erläutern. Darauf folgt die Vorstellung der für die Annotation entwickelten XML-Beschreibungssprache und des graphischen Editors. Im zweiten Teil des Vortrages stellen wir einige Methodenkombinationen vor, die es ermöglichen sollen, die Bild--#Text--#Verbindungen multimodaler Kulturformen, sowie deren Beitrag zur spezifischen Narrativität graphischer Romane, zu verstehen.

GNML-Editor: Werkzeuge zur (Halb-)Automatischen Annotation

Während die Analyse von Textcorpora oft bereits automatisiert möglich ist, bleibt eine automatische Analyse multimodaler Narrative derzeit eine Zukunftsvision. Im Fall von Comics und gezeichneter, sowie aus anderen Gründen nicht perspektivischer, Bilder gelingt die Objekt-Identifikation (etwa die Wiedererfassung eines vorab bekannten Charakters) nur mit viel Trainingsaufwand und recht hohen Fehlerzahlen. Auch bei der automatischen Erkennung Handschriften-ähnlicher Fonts versagen übliche Standard-OCRs. Daher führt der Weg über eine Annotation des Bild--# Materials mit anschließender Analyse der Annotationen und Bild-Daten. Hierzu wird im Rahmen unseres Forschungsprojektes die XML-Sprache „Graphic Narrative Markup Language“ (GNML) entwickelt, welche die visuellen und textuellen Aspekte Graphischer Literatur beschreibt. GNML baut auf der „Text Encoding Initiative“ (TEI), und damit auf etablierten Standards, auf. Basierend auf den GNML-Annotationen können die in der graphischen Literatur enthaltenen Texte analysiert werden, Auswertungen der Bildinhalte vorgenommen, oder deren Kombination analysiert werden.

Um die Fehleranfälligkeit bei der Annotation gering zu halten wird ein graphischer GNML-Editor entwickelt. Dieser unterstützt Fachwissenschaftler bei der effizienten Annotation mit Mechanismen wie Autovervollständigung von Charakter-Namen oder integrierten Rechtschreibprüfungen. Durch eine halb-automatische Erfassung wird die Annotation beschleunigt und so erst der Aufbau eines größeren Corpus ermöglicht. Teil des Editors ist eine Erkennung der Panel-Strukturen, sowie Werkzeuge, welche die Eckpunkte einer Sprechblase oder eines Textkästchens automatisch ermitteln. Ergänzt werden diese Werkzeuge um eine effiziente Charakter-Erfassung.

Da sich die Konzepte des Editors (visuelle Objekte mit graphischen und textuellen Eigenschaften) nicht nur auf Comics beschränken sondern auch auf andere Bild-Text- Kombinationen anwendbar sind, lässt sich der Editor auf eine Obermenge solcher Formate erweitern. Diese Generalisierung erlaubt es, eine XML-basierte Annotationssprache zu hinterlegen und automatisch einen entsprechenden Editor zu generieren, sowie Daten in der hinterlegten Annotationssprache zu erfassen. Damit kann der Editor auch in der Annotation anderer multimodaler Medien Anwendung finden.

Analysen und Visualisierungen eines Corpus Graphischer Romane

Der zweite Teil des Vortrags stellt Ansätze vor, die exemplarisch für einige strukturelle Bestandteile von Comics (und insbesondere des graphischen Romans) Methoden der Bild- und Textanalyse miteinander verbinden. Für die digitale Literaturwissenschaft entwickelte Zugänge wie das Topic Modelling sind für solche Kulturformen aufgrund ihrer Bildlastigkeit nur von beschränkter Relevanz. Ansätzen zur computergestützten Bilderkennung und der Analyse großer Bildmengen fehlt hingegen bisher oft das narrative Erkenntnisinteresse. Erschwerend kommt noch hinzu, dass die Konzepte der Narratologie meist für literarische Texte entwickelt wurden und den Spezifika multimodalen Erzählens häufig nicht gerecht werden.

In einer ersten Analyse des derzeit noch in Erstellung befindenden Gesamtkorpus von rund 300 graphischen Romanen vergleichen wir die historische Entwicklung der visuellen und Textebenen ihrer Buchcovers. Dazu gehören sowohl grammatische und semantische Auswertungen der Romantitel mit Hilfe des Stanford Parser (vgl. De Marneffe et al. 2006), als auch der farblichen und stilistischen Gestaltung. In einem weiteren Schritt widmen wir uns detaillierteren Analysen eines ersten Sub-Corpus, der aus den zehn meist zitierten Titeln des Gesamtcorpus besteht. Zwar lassen sich aufgrund der geringen Zahl hier keine Genre-Vergleiche anstellen, oder stichhaltig historische Entwicklungen nachverfolgen. Beispielhaft können allerdings narrative Entwicklungen dargestellt werden: so kombinieren wir Netzwerkanalysen der Figuren mit deren visueller Prominenz und zugeordnetem Textanteil, sowie mit stilistischen Analysen dieser Figurentexte. Weiters stellen wir, im Anschluss an Arbeiten von Lev Manovich (vgl. u. a. Manovich 2012), explorative Visualisierungen aller Einzelseiten im Gesamtverlauf der Erzählung vor.

Abschließend wendet sich der Vortrag der Frage zu, ob die Text-Bild-Verbindungen multimodaler Narrative mit solchen Methodenkombinationen aus der digitalen Literatur- und Bildwissenschaft zu erfassen sind, oder sich durch die Operationalisierung alternativer Ansätze aus der intermedialen Narratologie, etwa Rick Altmans Konzept des „Following“ (vgl. Altman 2008), eigenständige Analysemethoden entwickeln lassen.

Bibliographie

Manovich, Lev (2012): „How to Compare One Million Images?", in: Berry, David (ed.): *Understanding Digital Humanities* Basingstoke: Palgrave Macmillan 249-278.

Altman, Rick (2008): *A theory of narrative*. Columbia University Press.

De Marneffe, Marie-Catherine / MacCartney, Bill / Manning, Christopher D. (2006): „Generating Typed Dependency Parses from Phrase Structure Parses", in: *Proceedings of Language Resources and Evaluation* (LREC) 6: 449-454.

Wer bist Du, Nutzer? Eine Studie zur Nutzung dreier Korpus-Plattformen für mündliche Daten

Fandrych, Christian
fandrych@rz.uni-leipzig.de
Universität Leipzig, Deutschland

Frick, Elena
frick@ids-mannheim.de
IDS Mannheim, Deutschland

Hedeland, Hanna
hanna.hedeland@uni-hamburg.de
Universität Hamburg, Deutschland

Iliash, Anna
annailiash.fm@gmail.com
Universität Leipzig, Deutschland

Jettka, Daniel
daniel.jettka@uni-hamburg.de
Universität Hamburg, Deutschland

Meißner, Cordula
cordula.meissner@uni-leipzig.de
Universität Leipzig, Deutschland

Schmidt, Thomas
thomas.schmidt@ids-mannheim.de
IDS Mannheim, Deutschland

Wallner, Franziska
f.wallner@rz.uni-leipzig.de
Universität Leipzig, Deutschland

Weigert, Kathrin
kw59kahe@studserv.uni-leipzig.de
Universität Leipzig, Deutschland

Einleitung

Im Laufe der letzten Jahre sind weltweit mehrere Angebote entstanden, die mündliche Korpora – also Sammlungen von Audio- oder Video-Aufnahmen gesprochener Sprache mit zugehörigen Annotationen und Metadaten – der wissenschaftlichen Gemeinschaft zur Verfügung stellen. Exemplarisch seien CLAPI (Bert et al. 2010) und ESLO (Baude / Dugua 2011) für das Französische, die ORAL-Serie im Tschechischen Nationalkorpus (Kren 2015), oder auch das erst kürzlich gestartete BNC Spoken2014 genannt. Außer für genuin korpuslinguistische Untersuchungen sind diese Ressourcen auch für die Gesprächsanalyse, die Sprachvermittlung, die Kommunikationsforschung und viele weitere geisteswissenschaftliche Disziplinen von Interesse.

Da die Angebote relativ neu sind und ihre Nutzer die neuen Möglichkeiten des digitalen Zugriffs gerade erst erkunden, wissen wir noch relativ wenig darüber, wer solche mündlichen Korpora wie und für welche Zwecke nutzt. Dies war der Anlass für die hier beschriebene Nutzerstudie, die von Mitarbeitern dreier solcher Angebote im deutschsprachigen Raum (DGD, GeWiss, HZSK, siehe unten) durchgeführt wurde. Die Nutzerstudie besteht aus einer webbasierten Umfrage, qualitativen („kontextuellen") Interviews mit „Power"-Usern sowie Think-Aloud-Experimenten mit Neueinsteigern. In diesem Beitrag konzentrieren wir uns auf die Auswertung der Umfrage.

Korpus-Plattformen

Die Datenbank für Gesprochenes Deutsch (DGD , Schmidt 2014a) am IDS Mannheim bietet Zugriff auf 23 mündliche Korpora des Archivs für Gesprochenes Deutsch, darunter große Variationskorpora wie „Deutsche Mundarten" (Zwirner-Korpus) und „Deutsche Umgangssprache" (Pfeffer-Korpus) sowie Gesprächskorpora wie das neue Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, Schmidt 2014b). Die Plattform erlaubt das explorative Browsen, ein systematisches Querying sowie einen Download von Audio-Daten mit zugehörigen Transkripten und Metadaten. Seit dem ersten Release im Dezember 2012 haben sich über 4000 Studierende, Forschende und Lehrende für eine Nutzung der DGD registriert.

Das Korpus „ Gesprochene Wissenschaftssprache Kontrastiv “ (GeWiss, Slavcheva / Meißner 2014) wurde in einer Kooperation des Herder-Instituts an der Universität Leipzig, der Aston University (Birmingham) und der Universität Wroclaw aufgebaut. Ziel des Projekts ist es, eine empirische Basis für komparative Untersuchungen akademischer Sprache zu schaffen. Das Korpus umfasst zwei Genres gesprochener Wissenschaftssprache: Vorträge von Studierenden und Experten sowie mündliche Prüfungen. Der Großteil der Aufnahmen dokumentiert die Verwendung des Deutschen durch Muttersprachler des Deutschen, Englischen, Polnischen und Bulgarischen. Hinzu kommen Vergleichsdaten in Italienisch, Englisch und Polnisch von jeweiligen Muttersprachlern. In dieser Zusammensetzung ermöglicht das Korpus Untersuchungen auf verschiedenen Ebenen wie Lexik, Grammatik, Phonetik, Struktur, Funktion, Stil und Diskurs. Das GeWiss-Korpus hat zurzeit etwa 400 registrierte Nutzer.

Der Großteil der am Hamburger Zentrum für Sprachkorpora (HZSK , Hedeland et al. 2014) gehosteten mündlichen Korpora stammt aus dem Sonderforschungsbereich 538 Mehrsprachigkeit. Diese 27 Korpora dokumentieren in vielfältiger Weise verschiedene Aspekte individueller oder gesellschaftlicher Mehrsprachigkeit in unterschiedlichsten Sprachkonstellationen. Sie umfassen u. a. mehrere Spracherwerbskorpora und Korpora aus mehrsprachigen Kommunikationssituationen (wie z. B. Dolmetschen). Die Daten werden interessierten Studierenden, Forschenden und Lehrenden über das HZSK-Repositorium (Jettka / Stein 2014) als Korpora im EXMARaLDA-Format (Schmidt / Wörner 2014) zur Verfügung gestellt. Weitere Korpora wurden in jüngster Vergangenheit in die Bestände des HZSK integriert. Etwa 600 Nutzer weltweit haben sich bislang für eine Nutzung dieser Datenbestände angemeldet.

Umfrage

Die Umfrage wurde in Kooperation der drei Projektpartner zunächst mit 10 Testnutzern pilotiert und anschließend in ihrer endgültigen Form mit Hilfe der Software LamaPoll implementiert. Sie besteht aus insgesamt 128 Fragen, die in einen allgemeinen Teil mit Fragen zu persönlichen Daten (Alter, Sprachkenntnisse etc.) und zu relevanten Vorkenntnissen (Suchsprachen, Transkriptionserfahrung etc.) sowie drei angebotsspezifische Teile zu den jeweiligen Plattformen unterteilt sind.

Ein Aufruf zur Teilnahme wurde an etwa 5000 registrierte Nutzer der drei Angebote geschickt. Die Umfrage war anschließend für einen Monat offen. 669 Nutzer folgten dem Aufruf, 401 davon füllten den Fragebogen komplett aus. Dies entspricht einer Rücklaufquote von 8%. Im Folgenden diskutieren wir exemplarisch Ergebnisse zu ausgewählten Teilen der Umfrage.

Allgemeiner Teil

Nach den persönlichen Angaben im allgemeinen Teil ist der typische Nutzer eine Nutzerin (67%) zwischen 21 und 30 Jahren (54%), hat Deutsch als Muttersprache (66%), lebt und arbeitet in Deutschland (71%) und befindet sich im Studium bzw. ist graduiert (59% gegenüber 40% auf Doktorandenniveau oder darüber).

Auf die Frage „Welche Bereiche interessieren Sie?“ (Mehrfachauswahl war möglich), wurde wie folgt geantwortet:

Germanistische Linguistik	238	59,35%
Deutsch als Fremdsprache	199	49,63%
Korpuslinguistik	196	48,88%
Gesprächsforschung	195	48,63%
Spracherwerb	172	42,89%
Soziolinguistik	154	38,40%
Pragmatik	145	36,16%
Fremdsprachenunterricht	132	32,92%
Konstrastive Linguistik	122	30,42%
Dialektologie	114	28,43%
Phonetik	93	23,19%
Computerlinguistik	84	20,95%
Wissenschaftssprache	83	20,70%
Lexikographie	67	16,71%
Korpustechnologie	65	16,21%
Sonstiges (bitte angeben)	46	11,47%

Abb. 1: Frage 6 – „Welche der folgenden Bereiche interessieren Sie? (Mehrfachantwort möglich)“

Die Antworten zeigen, dass die Interessen der Nutzer sich über das gesamte Spektrum der zur Auswahl stehenden Teildisziplinen verteilen. Keine der Optionen wurde von weniger als 10% ausgewählt, so dass wir einstweilen auch keine der betreffenden Nutzergruppen als irrelevant für die weitere Entwicklung der Angebote ausschließen können. Unter den häufiger genannten Antworten sind mit etwa DaF, Gesprächsforschung und Pragmatik mehrere Nutzergruppen, für die trotz ihrer traditionell empirischen Ausrichtung die Arbeit mit digitalen Sprachdatenbanken sicherlich noch nicht als der Normalfall gelten kann. Dediziert „technisch“ ausgerichtete Disziplinen wie Computerlinguistik und

Korpustechnologie rangieren hingegen am unteren Ende der Liste.

Die Teilnehmer wurden weiterhin nach Vorkenntnissen befragt, die für die Arbeit mit mündlichen Korpora relevant sind:

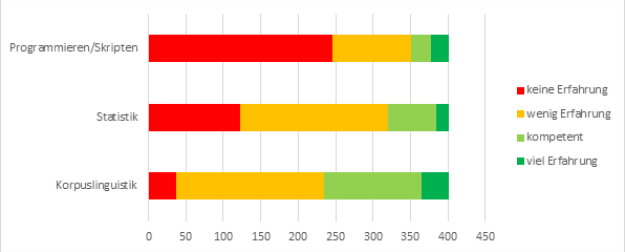


Abb. 2: Frage 10 – „Bitte beurteilen Sie Ihre Erfahrung in folgenden Bereichen“

Eine große Mehrheit der Teilnehmer gibt an, über keine oder wenig Erfahrung in Programmieren / Skripten und Statistik zu verfügen (88% bzw. 80%). Eine etwas größere Minderheit (41%) beurteilt ihre Kenntnisse in Korpuslinguistik positiv.

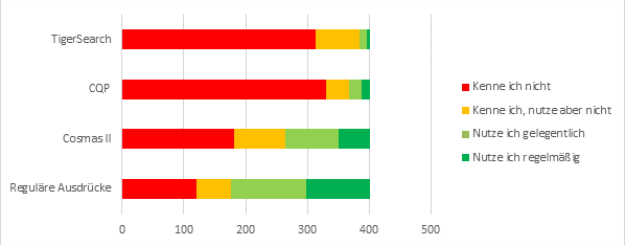


Abb. 3: Frage 11 – „Welche der folgenden Suchabfragesprachen kennen / nutzen Sie?“

Reguläre Ausdrücke sind der einzige formale Mechanismus, der von einer Mehrheit (56%) gelegentlich oder regelmäßig genutzt wird. Während COSMAS II – die Suchabfragesprache für die schriftlichen IDS-Korpora – noch bei 34% gelegentliche oder regelmäßige Anwendung findet, sind CQP und TigerSearch – als zwei weitere für die deutschsprachige Korpuslinguistik relevante Suchabfragesprachen den meisten Teilnehmern (82% bzw. 78%) unbekannt.

In Bezug auf die Vorerfahrungen mit Transkription stellt sich das Gesamtbild deutlich anders dar.

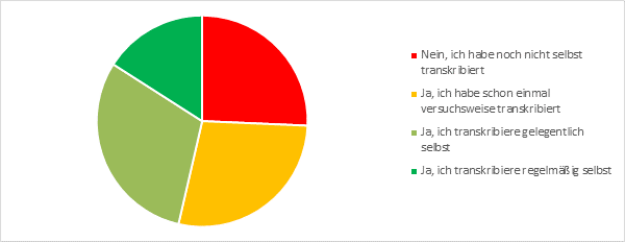


Abb. 4: Frage 13 – „Verfügen Sie über eigene Transkriptionserfahrung?“

Knapp die Hälfte der Befragten (46%) transkribiert gelegentlich oder regelmäßig selbst. Unter diesen Teilnehmern gaben etwas mehr als die Hälfte (56%) an, Standard-Office-Software (typischerweise MS Word, 82%) für die Transkription zu nutzen, etwa ebenso viele (55%), mit spezialisierter Transkriptionssoftware zu arbeiten.

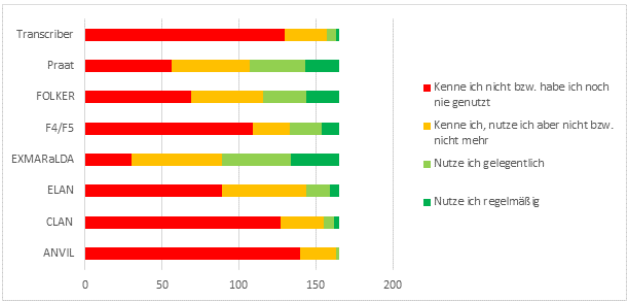


Abb. 5: Frage 16 – „Mit welchem / en spezialisierten Transkriptionseditor / en arbeiten Sie?“

EXMARaLDA (regelmäßige Nutzung: 19%, gelegentlich: 27%), Praat (13% bzw. 22%) und FOLKER (13% bzw. 17%) sind bei letzteren die am häufigsten genutzten Tools.

Angebotsspezifischer Teil

Nach dem allgemeinen Teil wurde Nutzern die Wahl gelassen, zu welchen der drei Angebote sie im weiteren Verlauf der Umfrage befragt werden wollten. Da sich eine Mehrzahl (261 Teilnehmer) hier für die DGD entschied, stellen wir im Folgenden einige exemplarische Auswertungen für diesen Teil der Umfrage vor.

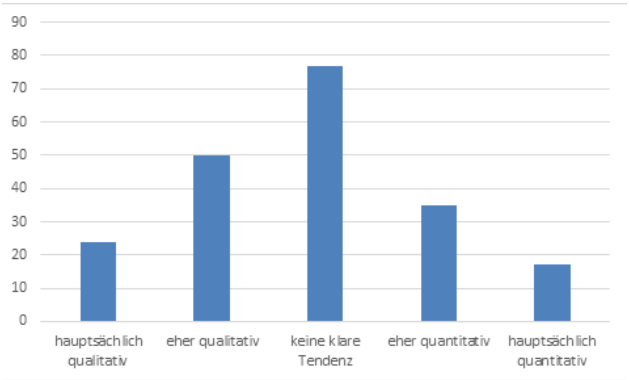


Abb. 6: Frage 33 – „Wie lässt sich Ihre methodische Herangehensweise am besten beschreiben, wenn Sie mit der Datenbank für gesprochenes Deutsch (DGD) arbeiten?“

Bei der Frage nach der Anwendung von qualitativen oder quantitativen Analysemethoden positionierte sich

der größte Anteil der Befragten (38%) in der Mitte des Spektrums. Bei den übrigen Befragten zeigte sich eine leichte Tendenz zu qualitativen Herangehensweisen (37% vs. 25%).

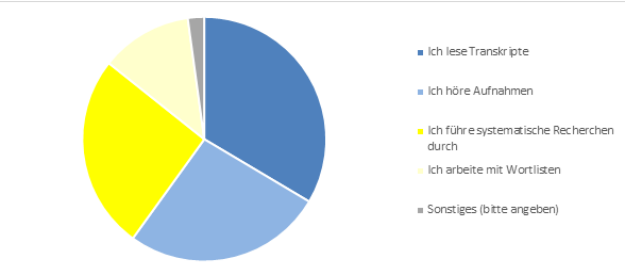


Abb. 7: Frage 34 – „Was ist Ihre Haupttätigkeit, wenn Sie mit der Datenbank für gesprochenes Deutsch (DGD) arbeiten? (Mehrfachantwort möglich)“

Dies spiegelt sich auch in den Antworten auf die Frage nach der Hauptaktivität beim Arbeiten mit der DGD wieder: Hier beurteilten die Befragten die manuell-intellektuelle Inspektion der Daten (Transkripte lesen, Audio anhören) als geringfügig relevanter als Methoden, die auf semi-automatischem Retrieval basieren (Queries, Wortlisten).

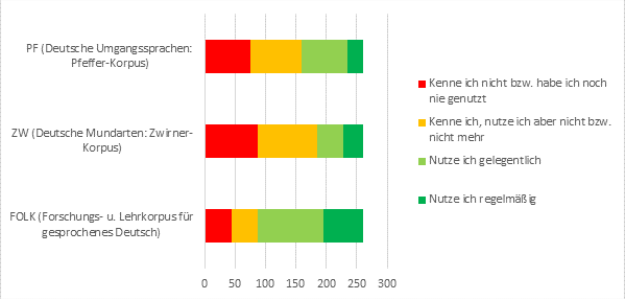


Abb. 8: Frage 37 – „Mit welchen Korpora der DGD arbeiten Sie? (Mehrfachantwort möglich)“

FOLK als das neueste, technisch fortschrittlichste und größte Gesprächskorpus ist auch dasjenige, das am meisten genutzt wird (regelmäßig oder gelegentlich von 25% bzw. 41%), und es besteht auch weiterhin Interesse an den älteren großen Variationskorpora ZW (12%/16%) und PF (10%/12%). Andere in der DGD enthaltene ältere und / oder kleinere Korpora fallen hingegen im Vergleich kaum ins Gewicht.

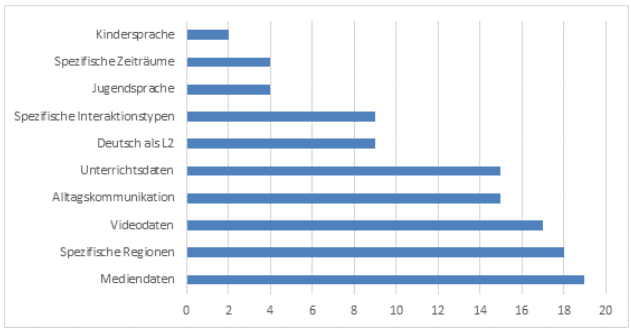


Abb. 9: Frage 54 – „Welche anderen / zusätzlichen Datentypen würden Sie sich in der DGD wünschen?“

Bei der Frage nach Wünschen für zusätzliche Daten oder neue Datentypen in der DGD wurden Mediendaten (z. B. geskriptete oder freie Interaktionen in Fernsehen oder Radio), Videodaten und Unterrichtsdaten auffällig häufig genannt, es gab aber auch mehrfache Nutzerwünsche nach ganz spezifischen Interaktionstypen (z. B. Arzt-Patienten-Kommunikation, Konflikte), nach Daten aus bestimmten Regionen (z. B. Schweiz, ehemalige DDR, Norddeutschland) oder von bestimmten Sprechern (Kinder, Jugendliche oder L2-Lerner) sowie nach Daten aus spezifischen Zeiträumen („nach der Wende“, „die frühesten archivierten Aufnahmen“).

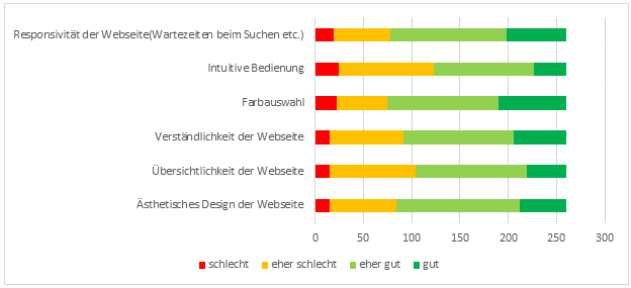


Abb. 10: Frage 52 – „Bitte bewerten Sie die Webseite der DGD“

Das Gesamturteil zur Nutzerfreundlichkeit der DGD-Website fällt positiv aus, wobei die Zufriedenheitswerte allerdings bei eher oberflächlichen Design-Details wie der Farbauswahl (positiv bewertet von über 70%) höher ausfallen als bei letztendlich entscheidenderen Kategorien wie „Intuitive Bedienung“ (52%).

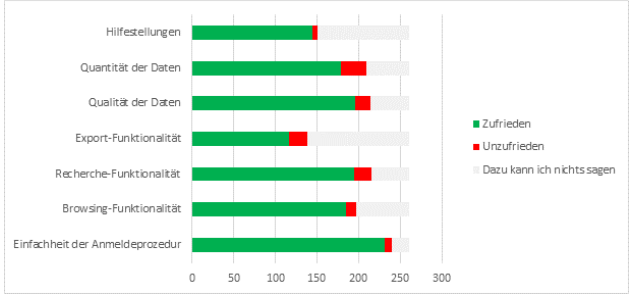


Abb. 11: Frage 49 – „Wie zufrieden sind Sie mit ...?“

Bezogen auf spezifische Teilbereiche der DGD-Funktionalität, wurden Quantität der Daten (11%), Exportoptionen (8%) und Suchfunktionalität (8%) am häufigsten als Bereiche genannt, mit denen Nutzer unzufrieden waren.

(Vorläufige) Schlussfolgerungen

Aus den Ergebnissen der Umfrage lassen sich eine Vielzahl von Informationen über die Hintergründe, Vorkenntnisse und Erwartungen der Nutzer sowie über die Art und Weise, wie sie mit den Plattformen arbeiten, entnehmen. Obwohl wir die Auswertung gerade erst begonnen haben, können wir bereits erste vorläufige Schlüsse ziehen: vielleicht am wichtigsten ist die Erkenntnis, dass die Nutzerschaft der Angebote in Bezug auf Forschungsinteressen und -hintergründe äußerst heterogen ist. Es zeichnet sich auch bereits ab, dass wir weder durchgängig von einem „technisch“ ausgebildeten Nutzer ausgehen können, noch, dass Angebote für mündliche Korpora vornehmlich mit „klassischen“ korpuslinguistischen Methoden genutzt werden. Ausgehend von diesen Erkenntnissen sind wir zuversichtlich, dass uns die vollständige Auswertung der Studie helfen wird, ein deutlich klareres Bild unserer Nutzer zu bekommen, und wir auf dieser Grundlage die Nützlichkeit und Nutzbarkeit der jeweiligen Ressourcen noch merklich verbessern können. Die vollständige Auswertung wird zum Zeitpunkt der Konferenz vorliegen und kann dort dann vorgestellt werden.

Bibliographie

Baude, Olivier / Duga, Céline (2011): "(Re)faire le corpus d’Orléans quarante ans après : quoi de neuf, linguiste?" In: *Corpus* 10: 99-118.

Bert, Michel / Bruxelles, Sylvie / Etienne, Carole / Mondada, Lorenza / Traverso, Véronique (2010): "Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL)", in: *Pratiques - Interactions et corpus oraux* 147-148: 17-34.

Cambridge University Press / Lancaster University (2015): *Spoken British National Corpus* http://languageresearch.cambridge.org/index.php/spoken-british-national-corpus [letzter Zugriff 09. Februar 2016].

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (2014): "Multilingual Corpora at the Hamburg Centre for Language Corpora", in: Ruhi, Sukriye / Haugh, Michael / Schmidt, Thomas / Wörner, Kai (eds.): *Best Practices for Spoken Language Corpora in Linguistic Research*. Cambridge: University Press 208-224.

Herder Institut der Universität Leipzig / Aston University (Birmingham) / Universität Wroclaw (2009-2016): *GeWiss*. Gesprochene Wissenschaftssprache https://gewiss.uni-leipzig.de [letzter Zugriff 09. Februar 2016].

HZSK = Hamburger Zentrum für Sprachkorpora: https://corpora.uni-hamburg.de [letzter Zugriff 09. Februar 2016].

IDS (2012-2016): *DGD*. Datenbank für Gesprochenes Deutsch http://dgd.ids-mannheim.de [letzter Zugriff 09. Februar 2016].

Jettka, Daniel / Stein, Daniel (2014): "The HZSK Repository: Implementation, Features, and Use Cases of a Repository for Spoken Language Corpora", in: *D-Lib Magazine* 20, 9 / 10 http://www.dlib.org/dlib/september14/jettka/09jettka.html [letzter Zugriff 09. Februar 2016].

Kren, Michal (2015): "Recent developments in the Czech National Corpus", in: *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)* 1-4.

Schmidt, Thomas (2014a): "The Database for Spoken German - DGD2", in: *Proceedings of the Ninth International conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland* 1451-1457.

Schmidt, Thomas (2014b): "The Research and Teaching Corpus of Spoken German - FOLK", in: *Proceedings of the Ninth International conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland* 383-387.

Schmidt, Thomas / Wörner, Kai (2014):"EXMARaLDA", in: Durand, Jacques / Gut, Ulrike / Kristoffersen, Giert (eds.): *The Oxford Handbook of Corpus Phonology*. Oxford: OUP 402-419.

Slavcheva, Adriana / Meißner, Cordula (2014): "Building and maintaining the GeWiss corpus – perspectives on the construction, sustainability and further enrichment of spoken corpora. A showcase." In: Ruhi, Sukriye / Haugh, Michael / Schmidt, Thomas /Wörner, Kai (eds.): *Best Practices for Spoken Language Corpora in Linguistic Research*. Cambridge: University Press. 20-35.

Automatische Textanalysen in der Geschichtswissenschaft – Auswertung, Interpretation und Relevanz

Fiedler, Maik
fiedler@gei.de
Georg Eckert Institut für internationale Schulbuchforschung, Deutschland

Weiß, Andreas
weiss@gei.de
Georg Eckert Institut für internationale Schulbuchforschung, Deutschland

Heuwing, Ben
heuwing@uni-hildesheim.de
Institut für Informationswissenschaft & Sprachtechnologie, Universität Hildesheim

Schnober, Carsten
schnober@ukp.informatik.tu-darmstadt.de
Ubiquitous Knowledge Processing Lab, Deutsches Institut für Internationale Pädagogische Forschung / Technische Universität Darmstadt

Motivation und Fragestellung

In jedem Digital-Humanities-Projekt (DH) stellt sich die Frage von neuem: Welche „Relevanz haben die Modellierung, Vernetzung und Visualisierung für die Geistesartefakte selbst und für den Gewinn reproduzierbarer wissenschaftlicher Erkenntnisse über sie? Im Projekt „Welt der Kinder“ (WdK) wurden diese Punkte mit Hilfe von Topic Modeling und Text-Mining-Werkzeugen mit in der Geschichtswissenschaft anerkannten Thesen in einem kontrollierten Verfahren überprüft. Es handelt sich bei WdK mit seinem repräsentativen Textkorpus von über 3000 historischen Schulbüchern um ein bisher weltweit einzigartiges Projekt, das für künftige ähnliche Vorhaben vorbildhaft sein will.

Aus Sicht der klassischen Geschichtswissenschaften gibt es bei der Bearbeitung großer Datenmengen häufig Argwohn gegenüber der Sekundäranalyse maschinell generierter Ergebnisse, verstärkt durch mangelndes Wissen über fachfremde Methodik. Dies lässt die Ergebnisse der DH oft als zweifelhaft oder nicht neuwertig erscheinen. Zusätzlich können Verzerrungen durch die Zusammensetzung einer Textsammlung entstehen, durch die Dokumentenauswahl und des zu analysierenden Vokabulars sowie aus den darauf aufbauenden Aggregierungen und Visualisierungen (Chuang et al. 2012). Große Datenmengen erfordern ein anderes Vorgehen bei der Auswertung als die traditionell in den Geschichtswissenschaften üblichen Verfahren. Die Methode der automatischen Textanalyse stellen trotzdem eine durch die Forschungsziele beeinflusste subjektive Sichtweise auf die vorhandenen Daten dar (DiMaggio et al. 2013). Wir zeigen an Hand eines in WdK vorgenommen Validierungsexperiments, welche Aushandlungsprozesse notwendig waren, um nachnutzbare und nachvollziehbare Ergebnisse zu erhalten.

Für die Meta-Analyse von klassischen und digitalen geschichtswissenschaftlichen Herangehensweisen ist die Beantwortung folgender Fragen prioritär:

Erstens) Wie können auf klassischem Weg erbrachte Ergebnisse für die DH so codifiziert werden, dass sie nicht nur für Menschen interpretierbar, sondern auch durch die digitalen Werkzeuge reproduzierbar sind? Sinn dieses Verfahrens ist es, Versuchsanordnungen und Analysen so aufzubauen, dass diese nicht immer „bei Null“ beginnen müssen, sondern, wie ein klassischer Fachtext, anerkannte Annahmen und Erkenntnisse implizit transportieren und wiederholen.

Zweitens) Wie kann die Belastbarkeit von Ergebnissen, die mit Hilfe von Methoden der automatischen Textmodellierung auf einem umfangreichen Korpus erbracht worden sind, validiert werden?

Drittens) Wie kann man die Leistung digitaler Methoden für explorative Analysen anwenden, ohne auf ein bereits feststehendes Ziel hinzuarbeiten?

Viertens) Wie müssen die Versuchsanordnung und das Projekt aufgebaut werden, um den Daten zu vertrauen und sie interpretieren sowie kontextualisieren zu können?

Der Vortrag wird den Arbeitsprozess (interdisziplinäre Arbeit an historischen Thesen mit Hilfe digitaler Tools) analysieren, die verschiedenen fachspezifischen Methoden problematisieren sowie schlaglichtartig Wege beleuchten, die zu möglichen Antworten auf die gestellten Fragen führen können.

Werkzeuge

Die Grundlage der Topic-Modelling-basierten Analyse besteht auf im Bereich DH etablierter Methoden wie LDA (*Latent Dirichlet Allocation*; Blei et al. 2003). Dieses Verfahren ordnet Begriffe auf Basis von Kookkurrenz und statistischen Analysen einander zu und extrahiert Topics in Form gewichteter Wortlisten. Diese ergeben für menschliche Benutzer interpretierbare Listen, und erlauben eine automatische Inferenz von Topic-Verteilungen innerhalb eines Dokuments.

Die Validierungsstudie wurde mit einem interaktiven Prototyp durchgeführt, der die Texte im Korpus und Statistiken über die Ergebnismengen zugänglich macht. Suchanfragen können sich auf Metadaten – beispielsweise Jahr und Ort der Veröffentlichung oder Schultyp – Termanfragen und Topic-Verteilungen beziehen. Ergebnisse werden mit Statistiken zur Topic-Intensität und relativen Dokumentenhäufigkeit im Zeitverlauf ausgegeben.

Vorgehen bei der Validierung:

Belastbarkeitsüberprüfungen bauen Vertrauen in datenbasierte, historische Schlussfolgerungen und Annahmen auf. So wird überprüft, ob die statistischen

Modelle existierende Erkenntnisse mehrheitlich bestätigen, und als wie zuverlässig bestätigende oder widerlegende Ergebnisse eingeschätzt werden (DiMaggio et al. 2013; Evans 2014). Die im Experiment bearbeiteten historischen Thesen stellten Sachverhalte dar, die sich quantitativ überprüfen lassen, etwa durch den Vergleich von Topic-Verteilungen (Newman / Block 2006; Yang et al. 2011), und im Nachhinein von Experten für das jeweilige Fachgebiet in Hinblick auf ihre Plausibilität überprüft werden.

Für die Validierungsstudie wurden zu überprüfende Thesen vorab definiert, um Abweichungen von der ursprünglichen Fragestellung zu dokumentieren. Sie sind repräsentativ für reale historische Fragestellungen im Rahmen des Projektes (Kolonien und Auswanderung; Französische Revolution und Befreiungskriege; deutsche Kriegsflotte). Dabei wurden in einem ersten Schritt Begrifflichkeiten und Interpretationen der Fragestellungen in interdisziplinären Arbeitsgruppen diskutiert, um fachliche Verständnisschwierigkeiten auszuräumen. Da die Thesen erschöpfend und präzise mit den vorhandenen Werkzeugen untersucht wurden, bilden auch die Auswertungsstrategien mögliche Vorgehensweisen für die Überprüfung bereits vorliegender Hypothesen ab.

Auswertung

Bei der Analyse der Thesen zeigten sich unterschiedliche Strategien für die einzelnen Schritte der Auswertung. Wichtig hierbei war, ob unterschiedliche Herangehensweisen, vergleichbare Ergebnisse reproduzierten. Die Ergebnisse der einzelnen Arbeitsgruppen widersprachen einander an wenigen Stellen, und gegebenenfalls primär in ihrer Bewertung der Verlässlichkeit der Ergebnisse. Die vorgegebenen geschichtswissenschaftlichen Thesen wurden in den Versuchen mit Topic-Modellen größtenteils bestätigt und zusätzlich mittels Termanfragen validiert.

Das Vorgehen bei den Topic-Modelling-basierten Analysen beinhaltete im ersten Schritt eine Suche nach relevanten Topics an Hand einzelner Terme. Dabei zeigte sich, dass die Topics in Modellen mit einer manuell überschaubaren Topic-Anzahl (50, 100, 200) für spezielle historische Forschungsfragen zu allgemein oder auch zu spezifisch ausfielen. Teilweise wurden daraufhin die Thesen stellvertretend an Hand thematischer Teilgebiete oder übergeordneter Themen untersucht.

Für eine höhere Genauigkeit wurden auch Kombinationen aus Termsuche und Dokumentenfiltern auf Basis automatisch generierter Topics eingesetzt. Für eine Bewertung der Abfragegenauigkeit wurden manuelle Inspektionen der relevantesten Trefferdokumente durchgeführt und Anfragen iterativ neu formuliert. Um für die Validierung eine Vergleichsebene bereitzustellen, wurden zusätzliche Analysen nur auf der Grundlage manuell und mittels historischen Vorwissens gewählter Terme durchgeführt.

Schlussfolgerungen

Zusammengefasst kann zwischen zwei grundlegenden Vorgehensweisen unterschieden werden. In der ersten Variante werden die aufgestellten Thesen konfirmatorisch überprüft. Diese werden dafür formalisiert und in Form von Suchanfragen und zu erwartenden Ergebnissen operationalisiert. Die Ergebnisse werden dann vor allem hinsichtlich der erwarteten Zeitverläufe und relativen Unterschiede zwischen Untermengen interpretiert.

Die explorative Herangehensweise an die Datenanalyse berücksichtigt dagegen auch andere Hinweise aus den Ergebnissen, und sucht nach Erklärungen für beobachtete Auffälligkeiten. Die Aussagekraft der Ergebnisse kann dabei jedoch dadurch eingeschränkt werden, dass die untersuchten Thesen erst mit Kenntnis der Daten formuliert worden sind. Eine Strategie, um diese Unsicherheit auszugleichen, besteht darin, Evidenz für eine Aussage mit mehreren unterschiedlichen Vorgehensweisen zu sammeln.

Diese Ergebnisse zeigen den potentiellen Mehrwert von DH an, da mit Hilfe computerlinguistischer und informationswissenschaftlicher Methoden klassische Thesen aus der Geschichtswissenschaft präzisiert werden konnten. Die Interpretation quantitativer Ergebnisse, etwa als Diagramm visualisiert, konnte sich nach Bedarf auf die vorab definierten Vorannahmen beschränken. Die Einbeziehung größerer zeitlicher Kontexte erforderte teilweise, die dargestellten Verläufe und Tendenzen mit verschiedenen Zeitspannen neu zu interpretieren. Als wichtige Vorgehensweise hat sich hier die Bildung eines gleitenden Durchschnitts über längere Zeiträume erwiesen, um thematische Tendenzen zuverlässiger interpretieren zu können.

Als wichtiger Faktor stellte sich auch die Qualität der OCR-Digitalisierung heraus. Bei Daten aus historischen Quellen (Schriftbild Sütterlin / Fraktur) werden auch mit aktueller Technologie aufgrund der verwendeten Schriftarten teilweise über 10 Prozent der Zeichen falsch erkannt, was bei der Auswertung der maschinell generierten Topics durch die Benutzer zu Problemen bei der Interpretation und Weiterverwendung führt. Daher muss die Frage gestellt werden, wie Daten zukünftig in den Vorverarbeitungsschritten aufbereitet werden, damit Topic Modelling und andere automatische Methoden zu hilfreichen und interpretierbaren Ergebnissen führen.

Neben der Einbeziehung von Topic Models, die auf unterschiedliche Perspektiven optimiert wurden, werden im Rahmen des Projektes andere Herangehensweisen an die statistische Textmodellierung, wie z. B. Clustering-Verfahren, in Hinblick auf ihre Anwendbarkeit und Robustheit vergleichend evaluiert. In diesem Zusammenhang ist es wichtig, thematisch relevante Topics einfach auffindbar zu machen und sie für Anfragen kombinieren zu können. Des Weiteren sollten Topics geordnet nach Themen oder Diskursfeldern und / oder -strängen präsentiert werden sowie in einer leicht lesbaren

Anzeige deren synchrone und diachrone Verteilungen herausstellen, wobei Ungleichverteilungen innerhalb der Untersuchungsmenge und die Zuverlässigkeit statistischer Aggregierungen deutlich gemacht werden müssen.

Bibliographie

Blei, David. M. / Ng, Andrew. Y. / Jordan, Michael I. (2003): "Latent Dirichlet allocation", in: *Journal of Machine Learning Research* 3: 993–1022.

Chuang, Jason / Ramage, Daniel / Manning, Chistopher / Heer, Jeffrey (2012): "Interpretation and Trust: Designing Model-driven Visualizations for Text Analysis", in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *CHI '12*. New York, NY, USA: ACM 443–452.

DiMaggio, Paul / Nag, Manish / Blei, David (2013): "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding", in: *Poetics* 41, 6: 570–606.

Evans, Michael S. (2014): "A Computational Approach to Qualitative Analysis in Large Textual Datasets", in: *PLoS ONE*9, 2, e87908.

Kaplan, Frédéric (2015): "A map for Big Data research in Digital Humanities", in: *Frontiers in Digital Humanities* 2, 1: http://journal.frontiersin.org/article/10.3389/fdigh.2015.00001/abstract [letzter Zugriff 08. Januar 2016].

Newman, David J. / Block, Sharon (2006): "Probabilistic topic decomposition of an eighteenth-century American newspaper", in: *Journal of the American Society for Information Science and Technology* 57, 6: 753–767.

Yang, Tze-I. / Torget, Andrew J. / Mihalcea, Rada (2011): "Topic modeling on historical newspapers", in: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics 96–104.

Das juristische Referenzkorpus (JuReko) - Computergestützte Rechtslinguistik als empirischer Beitrag zu Gesetzgebung und Justiz

Gauer, Isabelle
isabelle.gauer@medienkultur.uni-freiburg.de
Albert-Ludwigs-Universität Freiburg, Deutschland

Hamann, Hanjo
hamann@coll.mpg.de
Max-Planck-Institut für Gemeinschaftsgüter, Deutschland

Vogel, Friedemann
friedemann.vogel@medienkultur.uni-freiburg.de
Albert-Ludwigs-Universität Freiburg, Deutschland

Sprachwissenschaftler_innen und Jurist_innen haben gemein, dass sie mit Texten arbeiten. Der juristische Umgang mit Texten ist allerdings geprägt und überformt von den Verfassungsgeboten der Rechtssicherheit und Vorhersehbarkeit der Interpretation von Normtexten, die eine disziplinäre Standardisierung erfordern: „Im Gegensatz zur grundsätzlich nicht normierbaren Alltagssprache oder zur Offenheit literaturwissenschaftlicher Interpretationen ist die Sprache des Rechts auf weitestgehende Verbindlichkeit, Deutlichkeit und Disziplin (zumindest) angelegt“ (Jeand'Heur 1998: 1287). Juristische Fachtexte lassen sich deshalb nur mit einem stark spezialisierten fachsprachlichen Sach- und (impliziten) Methodenwissen adäquat verstehen (vgl. hierzu Vogel 2012b: 34ff.).

Das spezialisierte Fach(sprach)wissen in der Jurisprudenz hat mindestens drei Funktionen: Erstens soll es juristische Entscheidungsarbeit valide und zuverlässig organisieren; zweitens soll es die Komplexität der Lebenswelt auf ‚rechtsrelevante' und verfahrenssichere, also in juristischen Kategorien verarbeitbare, Ausschnitte reduzieren; drittens stiftet es binnendisziplinäre Identität (Ingroup): Wer die Sprache und die ‚Denke' der Jurisprudenz nicht beherrscht, hat vor Gericht schlechte Karten.

All diese in der Regel für Laien nicht erkennbaren Funktionen stehen hinter sog. „ Subsumtionen“, also der juristischen Auslegungsmethode. Damit ist kein rechtspositivistisches ‚Anwenden' eines objektiv oder subjektiv vorgegebenen ‚Gesetzesinhalts' gemeint. Die juristische „Auslegung“ von Normen ist vielmehr ein komplexer Prozess der Ko(n)textualisierung von Lebenswelt (zu beurteilender Sachverhalt, „Fall“) und Textwelt (inter- und intratextuelle Verknüpfung von Norm- und dogmatischen Texten). Lebens- und Textwelt sind dabei nicht lediglich ‚gegeben' und „im Sinne eines kybernetischen Informationsübertragungsmodells“ im Hinblick auf ‚die' Norm zu „decodieren“ (so noch Baden 1977: 14ff.; vgl. dazu kritisch Busse 2005). Sie ‚geben' dem hermeneutisch tätigen Rechtsarbeiter vielmehr sinnlich wahrnehmbare Hinweisreize (Gumperz 1982: 131f.), die gemeinsam mit bereits bestehendem, institutionalisiertem juristischen Norm(sprach)wissen in mentalen Modellen Sinn-voll gemacht werden können (Hörmann 1980). Rechtsnormen sind also keine absoluten Entitäten, sondern Ergebnis konstruktiver Textarbeit mit

unterschiedlichen versprachlichten Eingangsdaten und Geltungsansprüchen (Müller et al. 1997; Felder 2003).

Seit rund 30 Jahren widmet sich die Rechtslinguistik als gemeinsame Teildisziplin von Rechts- und Sprachwissenschaft diesen Vertextungsverfahren im Recht (vgl. Vogel 2016). Juristische und linguistische Untersuchungen erfolgten dabei bislang ausschließlich mittels qualitativer Zugänge und auf Basis weniger hundert Texte. Die Ergebnisse geben wichtige Einblicke in die Mikroprozesse unseres sprachbasierten Rechtssystems, sei es vor Gericht, in der Verwaltung oder in der Gesetzgebung (Überblick bei Felder / Vogel 2016). In der frühen Rechtskybernetik und heutigen Rechtsinformatik hingegen wird das Recht meist als logisch operierendes Ontologiesystem zu formalisieren versucht, das die semantisch Struktur seiner realen performativen Bearbeitung jedoch vernachlässigt (vgl. zur Kritik am „Subsumtionsautomaten 2.0“ Kotsoglou 2014; Vogel 2015). Erst neuere Ansätze einer „evidenzbasierten Jurisprudenz“ (Hamann 2014) und rechtstheoretisch fundierten Korpuslinguistik in den USA (Mouritsen 2010, 2011) sowie in Deutschland (Vogel 2012a; Vogel et al. 2015; Hamann 2015) versprechen praxisnahe Analysen und Einsichten in die ‚Makroökonomik‘ juristischer Fachsprache und -kommunikation.

An dieser Stelle setzt ein seit 2014 laufendes und von der Heidelberger Akademie der Wissenschaften finanziertes Projekt zur Konzeption und Auswertung eines „Juristischen Referenzkorpus“ (JuReko) an (Vogel / Hamann 2015). Ziel des Projektes, das den Kern der „International Research Group Computer Assisted Legal Linguistics“ (CAL ^{2} 2014-2016) bildet, ist im ersten Schritt der Aufbau eines kontrollierten, zunächst statischen Fachtext-Korpus, das alle wichtigen Textsorten aus Judikative, Legislative und Rechtswissenschaft umfasst (v.a. Aufsätze aus juristischen Fachzeitschriften, Entscheidungstexte und Normtexte; Zielgröße: rund eine Milliarde fortlaufender Wortformen). Die Textdaten werden zunächst im html-Format gewonnen und anschließend in mehreren Konvertierungschritten TEI-P5-konform kodiert. Dafür kommen xsl-Transformationen zum Einsatz, die auf die unterschiedlichen Webseitenstrukturen angepasst werden. Im Anschluss werden die Texte mit Part-of-Speech und weiteren Annotationen und Metadaten angereichert, wobei die speziellen Anforderungen einer rechtslinguistischen Textanalyse und -verarbeitung im Vordergrund stehen.

Das Korpus bildet im zweiten Schritt die Grundlage für die Erprobung neuer computerlinguistischer Methoden zur Analyse insbesondere juristischer Semantik bzw. Dogmatik sowie zur Beschreibung von Wortschätzen und grammatischen Mustern in verschiedenen Rechtsbereichen auf Basis geeigneter Metriken. In Zusammenarbeit mit Praktikern aus Gesetzgebung und Rechtsprechung werden weitere Untersuchungsprojekte abgeleitet und vorbereitet. Hierzu zählt etwa die Entwicklung von Werkzeugen für die rechtslinguistisch wie korpusstatistisch-empirisch fundierte Optimierung der Gesetzesredaktion.

Der Vortrag stellt das Infrastrukturvorhaben „JuReko“ vor und diskutiert Möglichkeiten und Grenzen des durch die Projektgruppe entwickelten Ansatzes der „Computergestützten Rechtslinguistik“ als komplementären Beitrag zur qualitativen, juristischen Hermeneutik. Dabei wird anhand von Beispielen sowohl auf die textlinguistischen als auch technischen Details des Projektes eingegangen. Im Ausblick steht die Erweiterung des JuReko um Rechtstexte des britischen Case Law als Ausgangspunkt für ein Europäisches Rechtskorpus (European Law Corpus) und damit eine weltweit einzigartige Grundlage für rechts(sprach)kulturvergleichende Studien.

Bibliographie

Baden, Eberhard (1977): *Gesetzgebung und Gesetzesanwendung im Kommunikationsprozess.* Studien zur jur. Hermeneutik u. zur Gesetzgebungslehre. Baden-Baden: Nomos-Verlagsgesellschaft.

Busse, Dietrich (2005): „Ist die Anwendung von Rechtstexten ein Fall von Kommunikation? Rechtslinguistische Überlegungen zur Institutionalität der Arbeit mit Texten im Recht“, in: Lerch, Kent D. (ed.): *Die Sprache des Rechts*. Recht Vermitteln: Strukturen, Formen und Medien der Kommunikation im Recht. 3 Bände. Berlin: Walter De Gruyter 23–54.

CAL^{2} (2014-2016): *International Research Group: Computer Assisted Legal Linguistics*. University of Freiburg http://www.cal2.eu/ [letzter Zugriff 08. Januar 2016].

Felder, Ekkehard (2003): *Juristische Textarbeit im Spiegel der Öffentlichkeit*. Berlin / Boston: De Gruyter.

Felder, Ekkehard / Vogel, Friedemann (eds.) (2016): *Handbuch Sprache im Recht*. Berlin / Boston: De Gruyter Mouton

Gumperz, John Joseph (1982): *Discourse strategies*. Cambridge: University Press.

Hamann, Hanjo (2014): *Evidenzbasierte Jurisprudenz. Methoden empirischer Forschung und ihr Erkenntniswert für das Recht am Beispiel des Gesellschaftsrechts*. Tübingen: Mohr Siebeck.

Hamann, Hanjo (2015): „Der "Sprachgebrauch" im Waffenarsenal der Jurisprudenz. Die Rechtspraxis im Spiegel der quantitativ-empirischen Sprachforschung“, in: Vogel, Friedemann (ed.): *Zugänge zur Rechtssemantik*. Interdisziplinäre Ansätze im Zeitalter der Mediatisierung zwischen Introspektion und Automaten. Berlin / New York: Walter De Gruyter 184-204.

Hörmann, Hans (1980): „Der Vorgang des Verstehens“, in: Kühlwein, Wolfgang (ed.): *Sprache und Verstehen*. Tübingen: Narr 17–29.

Jeand'Heur, Bernd (1998): „Die neuere Fachsprache der juristischen Wissenschaft seit der Mitte des 19. Jahrhunderts unter besonderer Berücksichtigung von Verfassungsrecht und Rechtsmethodik“, in: Hoffmann, Lothar / Burkhardt, Armin / Ungeheuer, Gerold / Wiegand, Herbert Ernst / Steger, Hugo / Brinker, Klaus (eds.): *Fachsprachen: ein internationales Handbuch der Fachsprachenforschung und Terminologiewissenschaft*. (= Handbücher zur Sprach- und Kommunikationswissenschaft 14.1). Berlin: De Gruyter 1286–1295.

Kotsoglou, Kyriakos N. (2014): „Subsumtionsautomat 2.0. Über die (Un-)Möglichkeit einer Algorithmisierung der Rechtserzeugung“, in: *Juristenzeitung* 69, 9 451–457.

Mouritsen, Stephen C. (2010): „The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning“, in: *Brigham Young University Law Review* 1915–1980 http://www.lawreview.byu.edu/archives/2010/5/10Mouritsen.pdf [letzter Zugriff 07. November 2012].

Mouritsen, Stephen C. (2011): „Hard Cases and Hard Data: Assessing Corpus Linguistics as an Empirical Path to Plain Meaning“, in: *The Columbia*. Science and Technology Law Review 8: 156–205 http://www.stlr.org/cite.cgi?volume=13&article=4 [letzter Zugriff 07. November 2012].

Müller, Friedrich / Christensen, Ralph / Sokolowski, Michael (1997): *Rechtstext und Textarbeit* (= Schriften zur Rechtstheorie). Berlin: Duncker & Humblot.

Vogel, Friedemann (2012a): „Das Recht im Text. Rechtssprachlicher Usus in korpuslinguistischer Perspektive“, in: Felder, Ekkehard / Müller, Marcus / Vogel, Friedemann (eds.): *Korpuspragmatik*. Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin / Boston: De Gruyter 314–353.

Vogel, Friedemann (2012b): *Linguistik rechtlicher Normgenese*. Theorie der Rechtsnormdiskursivität am Beispiel der Online-Durchsuchung (= Sprache und Wissen 9). Berlin / Boston: De Gruyter.

Vogel, Friedemann (2015): „Zwischen Willkür, Konvention und Automaten: Die interdisziplinäre Suche nach Bedeutungen in Recht und Gesetz“, in: Vogel, Friedemann (ed.): *Zugänge zur Rechtssemantik*. Interdisziplinäre Ansätze im Zeitalter der Mediatisierung zwischen Introspektion und Automaten. Berlin / New York: Walter De Gruyter.

Vogel, Friedemann (2016): „Rechtslinguistik: Zur Bestimmung einer Fachrichtung“, in: Felder, Ekkehard / Vogel, Friedemann (eds.): *Handbuch Sprache im Recht* (= Handbücher Sprachwissen 12). Berlin / Boston: De Gruyter Mouton.

Vogel, Friedemann / Christensen, Ralph / Pötters, Stephan (2015): *Richterrecht der Arbeit – empirisch untersucht. Möglichkeiten und Grenzen computergestützter Textanalyse am Beispiel des Arbeitnehmerbegriffs*. Berlin: Duncker & Humblot.

Vogel, Friedemann / Hamann, Hanjo (2015): „Vom corpus iuris zu den corpora iurum – Konzeption und Erschließung eines juristischen Referenzkorpus (JuReko) “, in: *Jahrbuch der Heidelberger Akademie der Wissenschaften für 2014*. Heidelberg: Winter.

Operationalisierung von Forschungsfragen in CLARIN-D - Der Anwendungsfall Ernst Jünger

Goldhahn, Dirk
dgoldhahn@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

Eckart, Thomas
teckart@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

Heyer, Gerhard
heyer@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

Einleitung

CLARIN (*C*ommon *La*nguage *R*essources and Technology *In*frastructure) ist eine Forschungsinfrastruktur, deren Umsetzungssphase im Jahr 2016 erfolgreich abgeschlossen sein wird (Krauwer 2014). Ziel von CLARIN ist der Aufbau einer Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften, wobei insbesondere linguistische Daten, Werkzeuge und Dienste in einer integrierten, interoperablen und skalierbaren Infrastruktur für die Fachdisziplinen der Geistes- und Sozialwissenschaften bereitgestellt werden sollen. Im nachfolgenden Beitrag wollen wir ausschnittsweise skizzieren, welche Probleme CLARIN adressiert, wie die konzeptionelle Lösung und deren technische Umsetzung aussieht und in welcher Form eine Interaktion mit der Nutzercommunity stattfindet. Einige der gesetzten Ziele und gewählten Vorgehensweisen sind dabei allgemeingültig und wären somit zumindest teilweise auf andere Infrastrukturprojekte übertragbar.

Als konkretes Beispiel für die Nutzung einer solchen Forschungsinfrastruktur wird im Folgenden ein Usecase vorgestellt, der zur Beantwortung einer

realen Forschungsfrage der Germanistik verschiedene Bestandteile der Infrastruktur CLARIN nutzt (Goldhahn 2015). Dabei werden verteilte Daten und Werkzeuge genutzt, um Ressourcen zu finden, zweckmäßig aufzubereiten, zu analysieren und die Ergebnisse zu visualisieren.

Forschungsfrage

Ernst Jüngers politische Publizistik der Jahre 1919 bis 1933 liegt in einer philologisch aufbereiteten und annotierten Edition (Berggötz 2001) vor. Die Relevanz dieser Texte liegt in der Vielzahl behandelter Themen begründet, die relevant für die Entwicklung Deutschlands in den zwanziger und frühen dreißiger Jahren sind. Dies umfasst unter anderem Fronterfahrungen, Konsequenzen des verlorenen Krieges sowie das Thema der nationalen Neuorientierung. Dabei ändern Jüngers Texte in den 15 Jahren ihrer Erstellung deutlich thematische Prioritäten und linguistische Form (Gloning 2016).

Schlüsselfragen, die aus linguistischer und diskurshistorischer Perspektive bezüglich dieses Korpus bestehen, umfassen eine mögliche Korrelation der Sprachverwendung auf Wortebene mit den konkreten Themen, die in den Texten behandelt werden. Dabei sollte das lexikalische Profil Jüngers über die Dimension Zeit charakterisiert und mit den lexikalischen Profilen zeitgenössischen Materials (wie zum Beispiel Zeitungstexte der 1920er oder Werke anderer Autoren der gleichen Zeit) abgeglichen werden.

Operationalisierung

Um diese Forschungsfragen systematisch zu beantworten, müssen sie zuerst operationalisiert werden. Wichtige Aspekte dieses Prozesses sind:

- Daten: Textkollektionen, die für Forschungsfrage genutzt werden können (sowohl für Analyse- als auch Referenzkorpora)
- Algorithmen: Methoden, um die gewünschten Analysen durchzuführen und durch ihre Kombination zu komplexeren Anwendungen und Prozesses zu verbinden
- Ergebnisse und Visualisierungen: Präsentation und Zugriffsmöglichkeiten auf die Analyse- und Rohdaten

Fokus der Operationalisierung wird auf der Nutzung der CLARIN Infrastruktur liegen, um relevante Daten und Algorithmen zu suchen und die Analyse durchzuführen. Dabei werden zuerst Texte gesucht, die für die Forschungsfrage von Relevanz sind. Das Korpus von Ernst Jüngers politischer Publizistik der Jahre 1919 bis 1933, das unter anderem auch die Veröffentlichungsdaten aller Texte enthält, dient dabei als Startpunkt.

Für den eigentlichen Vergleich wird eine konkrete Analysemethode benötigt. Eine Möglichkeit ist hier die Nutzung einer sogenannten Differenzanalyse (Heyer et al. 2008). Dabei können Unterschiede zwischen Jüngers Texten unterschiedlicher Jahre oder zwischen Jüngers Texten und Referenzkorpora untersucht werden.

Dies erlaubt uns die:

- Quantifizierbarkeit von Korpusähnlichkeit,
- Identifikation von Vokabularunterschieden und
- weitere Analysen hervorstechender Ergebnisse.

Referenzdaten

Eine Voraussetzung für die Durchführung einer Differenzanalyse ist die Verfügbarkeit von Referenzmaterial. Für die Suche nach entsprechenden Textdaten bietet sich das bereits erwähnte CLARIN Virtual Language Observatory an. Durch die Einschränkung der vorhandenen Ressourcen des VLO über facettierte und Volltextsuche auf Korpora in deutscher Sprache des 20. Jahrhunderts stellt sich das DWDS Kernkorpus als relevante Ressource heraus (Abbildung 1).

Abb. 1: Suche nach Referenztexten unter Verwendung des Virtual Language Observatory.

Das DWDS Korpus (Geyken 2006) wurde an der Berlin-Brandenburgischen Akademie der Wissenschaften zwischen 2000 und 2003 erstellt.

Der Hauptzweck des DWDS Kernkorpus ist der Einsatz als empirische Basis eines großen monolingualen Wörterbuches des 20. Jahrhunderts. Das Kernkorpus besteht aus ungefähr 100 Millionen laufenden Wörtern und ist weitgehend über Zeit und vier Genres balanciert. Über die DWDS Webservices wurden Texte aller Genres extrahiert.

Kombination zu Workflows - Vorverarbeitung

Voraussetzung für die Durchführung einer Differenzanalyse ist die Aufbereitung des Rohmaterials. Dabei müssen insbesondere die Wortfrequenzen der zugrunde liegenden Texte extrahiert werden. Damit sind vor allem Satzsegmentierung und Tokenisierung wichtige Vorverarbeitungsschritte. Darüber hinaus ist die Nutzung eines POS-Taggers zur Generierung von Wortartinformationen für erweiterte Analysen hilfreich.

Für derartige Verarbeitungen ist die bereits erwähnte verteilte Umgebung WebLicht (Hinrichs et al. 2010) ein wichtiges Hilfsmittel. Abbildung 2 stellt einen Überblick über eine WebLicht-basierte Prozesskette dar. Sie importiert die Plaintext-Dateien, konvertiert diese in ein internes Format (das Text Corpus Format TCF), extrahiert Sätze und Wörter, annotiert Wortarten und zählt die Häufigkeit aller vorkommenden Wörter.

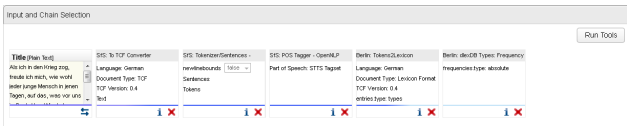


Abb. 2: Vorverarbeitungskette in WebLicht.

Diese Verarbeitung wurde auf der Basis der Ernst Jünger Texte für die Jahre 1919 bis 1933 durchgeführt. Als Resultat stehen die Worthäufigkeiten für jedes einzelne Jahr dieser Zeitspanne zur Verfügung. Darüber hinaus wurden die Referenztexte des DWDS in 15 Jahresscheiben zerlegt und jeweils für jedes Genre ein Teilkorpus erstellt. Diese 60 Einzelressourcen wurden anschließend mittels der bereits erläuterten Prozesskette aufbereitet.

Kombination zu Workflows - Analyse

Die eigentliche Analyse wurde im Anschluss mithilfe der Webanwendung Corpus Diff[1] durchgeführt. Diese Webumgebung ermöglicht die vergleichende Analyse verschiedener Textkorpora, genauer, deren Vokabulars. Die einfach zu benutzende Oberfläche erlaubt das Anlegen verschiedener Analyseprozesse für eine parallele Verarbeitung. Die Berechnung der Korpusähnlichkeit erfolgt dabei ausschließlich auf der Basis von Wortlisten die jeweils ein Textkorpus repräsentieren. Die Oberfläche erlaubt die Auswahl aus verschiedenen Ähnlichkeitsmaßen, die alle auf der Kosinusähnlichkeit von Wortvektoren basieren (Goldhahn 2013). Das Ergebnis ist ein normalisierter Wert zwischen 0 (keine Ähnlichkeit der Wortlisten) und 1 (Vokabulare mit identischer Häufigkeitsverteilung). Die Anwendung basiert komplett auf RESTful Webservices, die alle benötigten Informationen bereitstellen: einen Überblick über alle vorhandenen Korpusrepräsentationen und die vollständigen Wortlisten für jedes Korpus.

Die Nutzung von Worthäufigkeitslisten hat verschiedenen Vorteile: Wortlisten sind verdichtete Repräsentationen des Inhalts eines Korpus, die aufgrund ihrer geringen Größe einfach zu verarbeiten sind. Darüber hinaus unterliegen diese Informationen keinen Einschränkungen durch das Urheberrecht, da kein Zugriff auf die eigentlichen Volltexte benötigt wird. Dies bedeutet, dass in den meisten Fällen selbst für Ressourcen mit sehr restriktiven Lizenzbedingungen ein Austausch dieser Daten unbedenklich ist.

Über die Weboberfläche kann ein Nutzer alle relevanten Einstellungen vornehmen: Auswählen einer Korpusmenge, des zu nutzenden Ähnlichkeitsmaßes und wie viele der häufigsten Wörter für die Analyse genutzt werden sollen (s. Abbildung 3). Als Resultat wird dem Benutzer eine Matrixdarstellung der paarweisen Korpusähnlichkeit mit verschiedenen Farbschemata präsentiert. Diese Farbschemata werden zur Betonung ähnlicher und somit zusammengeclusteter Korpora genutzt. Ein Dendogram stellt darüber hinaus eine Visualisierung der Korpusähnlichkeiten auf der Basis eines Single-Linkage-Clusterings für alle genutzten Wortlisten dar. Beide Visualisierungen, Matrix und Dendogram, sind Mittel zur Identifikation interessanter Korpuspaare mit ungewöhnlich hoher oder niedriger Vokabularähnlichkeit. Die beschriebene Analyse kann genutzt werden, um eine diachrone Analyse der Änderungen über die Zeit durchzuführen, aber auch um Korpora unterschiedlichen Genres oder unterschiedlicher Herkunft miteinander zu vergleichen.

Abb. 3: Konfiguration eines Korpusvergleichs-Prozesses.

Durch die Auswahl zweier Korpora können detailliertere Informationen über die Unterschiede ihrer Vokabulare angezeigt werden. Dies beinhaltet vor allem auch Listen von Wörtern, die in einem der Korpora signifikant häufiger oder sogar exklusiv auftreten. Beides sind wertvolle Hilfsmittel um Wörter zu identifizieren, die spezifisch für die jeweilige Ressource sind. Darüber hinaus sind diese Ergebnisse Ausgangspunkt für tiefere hermeneutische Analysen durch die jeweiligen Fachwissenschaftler.

Ist der Nutzer an einem konkreten Wort interessiert, kann die Entwicklung seiner Häufigkeit über den Untersuchungszeitraum durch ein Liniendiagramm angezeigt werden. Dies ist üblicherweise relevant für wichtige Schlüsselterme der jeweiligen Texte oder Wörter, die in den vorherigen Analyseschritten als relevant herausgearbeitet wurden. Dabei kann die diachrone Entwicklung der Nutzungshäufigkeit des Wortes über verschiedene Genres hinweg einfach dargestellt werden.

Beispielergebnisse

Abbildung 4 (links) stellt die Ähnlichkeitsmatrix und das Dendogram für Ernst Jüngers Texte der Jahre 1919 bis 1933 dar. Unter anderen ist hier auch das Korpuspaar der Texte von 1920 und 1927 interessant, da hier eine besonders geringe Ähnlichkeit vorliegt. Bei der Analyse hervorstechenden Vokabulars fällt hier unter anderem die deutlich prominentere Nutzung des Wortes „Feuer“ in den Texten von 1920 auf (Abbildung 4, rechts).

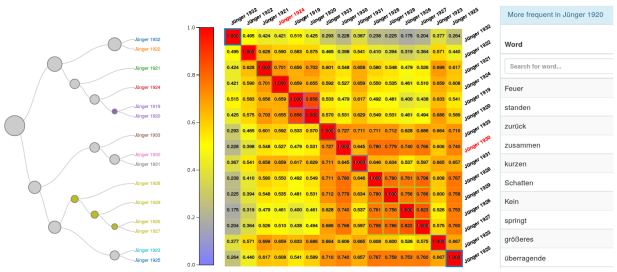


Abb. 4: Ähnlichkeitsmatrix und Dendogram für Ernst Jünger Texte der Jahre 1919-1933 (links), Liste der Wörter mit höherer relativer Worthäufigkeit für das Jahr 1920 im Vergleich mit 1927 (rechts).

Das Beispiel „Feuer“ (hier vor allem in seiner militärischen Bedeutung) zeigt die Nützlichkeit dieser Visualisierung. Sowohl in der Verwendung durch Ernst Jünger über 15 Jahre hinweg als auch im Vergleich mit Zeitungstexten der gleichen Periode, können Unterschiede in dessen Verwendung identifiziert werden (s. Abbildung 5) und sind damit ein idealer Einstiegspunkt für die tiefere Analyse durch Fachwissenschaftler.

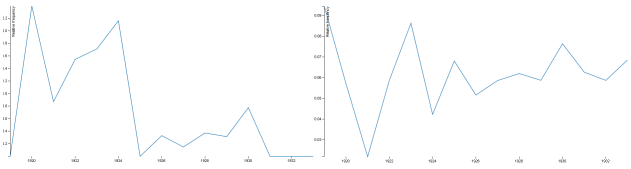


Abb. 5: Relative Häufigkeit des Wortes “Feuer” in Texten von Ernst Jünger(links) und in Zeitungstexten(rechts) von 1919 bis 1933.

Ein zweites Beispiel für diese Form der Analyse ist das Wort „Krieg“, das ebenfalls eine interessante Häufigkeitsverteilung aufweist. Die Verwendung dieses Wortes reflektiert das Nachwirken und die Allgegenwärtigkeit der Kriegserfahrungen in Texten dieser Zeit. Dabei ist die relative Häufigkeit in der Publizistik Ernst Jüngers deutlich höher als in Zeitungstexten.

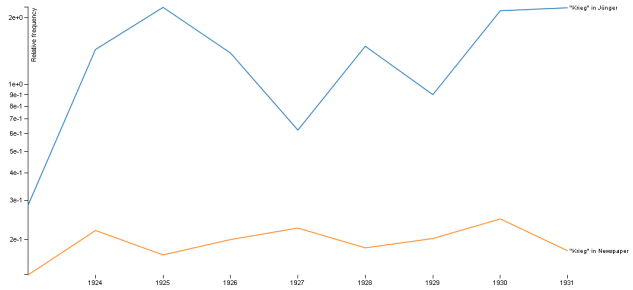


Abb. 6: Relative Häufigkeit des Wortes “Krieg” in Texten von Ernst Jünger und in Zeitungstexten von 1923 bis 1931.

Zusammenfassung

Anhand eines konkreten Anwendungsfalls der Germanistik wurde dargestellt wie sich die Infrastrukturbestandteile zu einem umfangreichen Workflow kombinieren lassen. Dabei wurden auf der Basis verteilter Ressourcen mit Hilfe einer Metadatensuchmaschine relevante Daten und Werkzeuge identifiziert und anschließend über eine föderierte Prozesskette aufbereitet. Die Analyse dieser Daten erfolgte über eine benutzerfreundliche Weboberfläche, die auch erweiterte Visualisierungsmöglichkeiten anbietet.

Notes

1. Erreichbar unter http://corpusdiff.informatik.uni-leipzig.de .

Bibliographie

Berggötz, Sven Olaf (2001): *Ernst Jünger*. Politische Publizistik 1919 bis 1933. Stuttgart: Klett-Cotta.

CLARIN-D: *Forschungsinfrastruktur für Sprachressourcen in den Geistes- und Sozialwissenschaften* http://www.clarin-d.de/de/ [letzter Zugriff 16. Februar 2016].

Geyken, Alexnder (2006): "A reference corpus for the German language of the 20th century", in: Fellbaum, Christiane (ed.): *Collocations and Idioms*. Linguistic, lexicographic, and computational aspects. London: Continuum Press 23-40.

Gloning, Thomas (in Vorbereitung): "Ernst Jüngers Publizistik der 1920er Jahre. Befunde zum Wortgebrauchsprofil", in: Benedetti, Andrea / Hagestedt, Lutz (eds.): *Totalität als Faszination*. Systematisierung des Heterogenen im Werk Ernst Jüngers. Berlin / Boston: de Gruyter.

Goldhahn, Dirk (2013): *Quantitative Methoden in der Sprachtypologie*. Nutzung korpusbasierter Statistiken. Dissertation, Universität Leipzig http://

www.qucosa.de/fileadmin/data/qucosa/documents/13055/Thesis_Goldhahn_pflichtexemplare_druck.pdf .

Goldhahn, Dirk / Eckart, Thomas / Gloning, Thomas / Dreßler, Kevin / Heyer, Gerhard (2015): "Operationalisation of Research Questions of the Humanities within the CLARIN Infrastructure – An Ernst Jünger Use Case", in: CLARIN Annual Conference 2015 in Wroclaw, Poland.

Heyer, Gerhard / Quasthoff, Uwe / Wittig, Thomas (2008): *Text Mining*. Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. W3L-Verlag.

Hinrichs, Marie / Zastrow, Thomas / Hinrichs, Erhard (2010): "WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure", in: *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010), Malta.

Krauwer, Steven / Hinrichs, Erhard (2014): "The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2014) 1525–1531.

Nutzung und Kombination von Daten aus strukturierten und unstrukturierten Quellen zur Identifikation transnationaler Lebensläufe

Gradl, Tobias
tobias.gradl@uni-bamberg.de
Universität Bamberg, Deutschland

Henrich, Andreas
andreas.henrich@uni-bamberg.de
Universität Bamberg, Deutschland

Biographien erscheinen als nahezu ubiquitärer Forschungsgegenstand in den unterschiedlichsten geisteswissenschaftlichen Disziplinen. Neben der qualitativen Betrachtung wurden aus diesem Grund auch Verfahren zur quantitativen Analyse biographischer Texte entwickelt, die zumeist die Identifikation und Extraktion relevanter Merkmale aus unstrukturiertem Text behandeln. So präsentieren beispielsweise Bamman und Smith eine Methode zur unüberwachten Erkennung biographischer Daten in unstrukturiertem Text (Bamman / Smith 2014). Blessing und Kuhn präsentieren mit ihrem Konzept und webbasiertem Prototypen zur Emigrationsanalyse eine konkrete Anwendung der quantitativen Analyse und Aggregation biographischer Daten (Blessing / Kuhn 2014).

Auf Basis der Machbarkeitsstudie *»Cosmobilities« – Grenzüberschreitende Lebensläufe in den europäischen Nationalbiographien des 19. Jahrhunderts* motivieren wir in diesem Vortrag die Notwendigkeit einer kombiniert qualitativen und quantitativen Betrachtung unterschiedlicher Quellen biographischer Daten – einer Aufgabe, der nach unserer Einschätzung aktuell eher wenig Priorität zugewiesen wird. Als Schwerpunkte vertiefen wir anschließend zwei für die Korrelation und Integration relevanter Daten wichtige Aspekte: Zum einen wird mit der *kontextspezifischen Kombination biographischer Daten* ein *iterativer Ansatz* vorgestellt, der bei der Verknüpfung von Einzelergebnissen der quantitativen Verfahren unterstützt und die Berücksichtigung qualitativer Resultate zulässt. Der zweite Schwerpunkt des Vortrags fokussiert auf die *Unterstützung des für die Erstellung biographischer Profile notwendigen Verarbeitungsprozesses* durch Komponenten der DARIAH-DE Infrastruktur, welche die Erweiterung des Prototypen um zusätzliche strukturierte und unstrukturierte Datenquellen erleichtern.

Fachwissenschaftlicher Kontext

Historiker des Leibniz-Instituts für Europäische Geschichte Mainz und der Ludwig-Maximilians-Universität München untersuchten im Rahmen von *Cosmobilities* exemplarisch, inwiefern biographische Texte über transnationale Bezüge einer Person hinwegtäuschen. Eine Besonderheit der Transnationalität [1] besteht darin, dass sich diese oft erst durch Betrachtung unterschiedlicher Quellen als solche zu erkennen gibt: Durch ihre nationale Prägung beschreiben biographische Texte – insbesondere in den Nationalbiographien – eine Person aus einer nationalen Perspektive und vernachlässigen oder verschweigen Einflüsse der Person auf andere Nationen oder Kulturkreise.

Transnationalität in Lebensläufen: Ein Beispiel

Betrachten wir als Beispiel den 1847 geborenen, jüdischen Bankier Jakob Heinrich Schiff. Nach Geburt und Kindheit in Frankfurt migrierte dieser zunächst im Alter von 18 Jahren und – nach drei Jahren in Hamburg und Frankfurt – 1875 ein weiteres Mal in die USA.

Der rund 950 Wörter umfassende Eintrag zu Jakob Schiff in der deutschsprachigen Wikipedia gibt Aufschluss über die Transnationalität in seinem Leben und betont insbesondere auch berufliche Stationen als Bankier. Der mit rund 2.350 Wörtern umfassendere, englischsprachige Artikel unterscheidet sich vor allem durch die differenzierte Betrachtung des Philanthropen und Geschäftsmanns und seine weitreichende finanzielle Unterstützung Japans im

Krieg gegen Russland 1904-1905. Obwohl beide Artikel jeweils die wesentlichen Aspekte seines Lebens umfassen, enthalten diese auch Informationen, die dem jeweils Anderen fehlen: So erwähnt nur der deutsche Eintrag Schiffs Brüder und beschreibt seine Rolle als Gründungsmitglied der Johann Wolfgang Goethe-Universität. Im englischsprachigen Beitrag fehlen diese Informationen, während aber eine detaillierte Auflistung der von ihm unterstützten, in den Vereinigten Staaten ansässigen Einrichtungen vorgelegt wird.

Historiker können für eine fundierte Auseinandersetzung mit dem Leben von Jakob Schiff auf einen Eintrag der Datenbank von *Immigrant Entrepreneurship* zurückgreifen. In dieser führt das 1987 gegründete Deutsche Historische Institut Washington (DHI) fundierte, redaktionell geprüfte Einträge zu Deutsch-Amerikanischen Unternehmern. Schiff ist dort mit einem über 10.000 Worte umfassenden Artikel verzeichnet. Und obwohl der Artikel eine historisch differenzierte Analyse seines Lebens und Wirkens liefert: einige in der Wikipedia verfügbare Informationen (z. B. Informationen über die Brüder und seine Stiftung des orientalischen Seminars an der Universität Frankfurt) fehlen auch hier.

Wikipedia als biographische Quelle

Das Beispiel Jakob Schiffs erlaubt zwei direkte Rückschlüsse: Erstens, dass oft erst durch die Kombination nationaler Perspektiven ein übergreifender Eindruck über eine transnationale Biographie entstehen kann. Zweitens kann die Wikipedia zwar aufgrund ihrer Intention und Ideologie nicht als Quelle historischer Forschung dienen; für die Identifikation und initiale Analyse der Transnationalität von Biographien bietet die Wikipedia jedoch den Vorteil einer – insbesondere gegenüber den Nationalbiographien – oft weitaus geringeren nationalen Prägung. Vor allem jedoch stehen Wikipedia-Artikel in den verschiedensten Sprachen frei und ohne Zugriffshürden zur Verfügung, worin ein bedeutender Vorteil für die Anwendung quantitativer Verfahren liegt: Allein die deutschsprachige Wikipedia beinhaltet etwa 560.000 Einträge zu Personen. In Kontrast hierzu stellen die ebenfalls beachtlichen Bestände der Allgemeinen Deutschen Biographie (ADB) rund 26.500 Einträge zu Personen bis einschließlich des 19. Jahrhunderts, sowie die Neue Deutsche Biographie (NDB) derzeit knapp 22.000 Einträge.

Für erste quantitative Betrachtungen werden daher bewusst zunächst die Artikel der Wikipedia und die strukturierten Daten aus Wikidata verwendet, um eine breite Datenbasis zu schaffen. Durch die angestrebte Kombinierbarkeit und Selektierbarkeit von Quellen wird die Implementierung später auch Möglichkeiten bieten, Analysen auf historisch fundierte Quellen einzuschränken oder diese z. B. auch mit den Ergebnissen aus der Wikipedia zu vergleichen.

Qualitative Unterstützung der Forschung

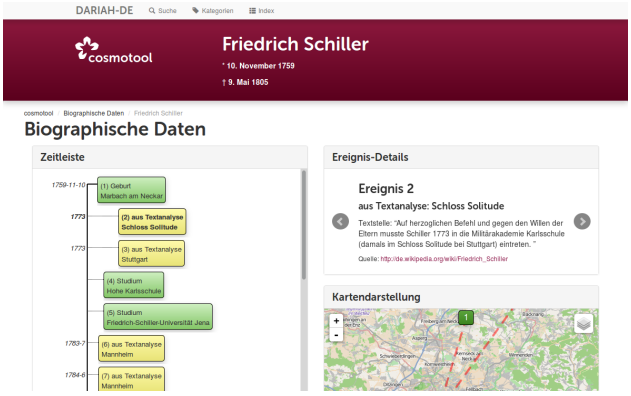
Ein erster entwickelter Prototyp umfasst neben rund 1,8 Millionen aus Wikidata abgeleiteten, biographisch relevanten Daten auch Ergebnisse der quantitativen Analyse biographischer Texte aus der Wikipedia. Durch die Zusammenführung von Ereignissen aus unterschiedlichen und idealerweise auch mehrsprachigen Quellen werden die biographischen Profile schrittweise erweitert und verfeinert.

Kontextspezifische Kombination von Daten

Durch die Kombination unterschiedlicher Quellen kann aber nicht nur eine größere Menge an Ereignissen erkannt werden, auch die Qualität der abgeleiteten Profile kann gesteigert werden. Angaben zu Zeitpunkten, Orten und interagierenden Personen werden in unstrukturierten Texten durch die Anwendung computerlinguistischer Verfahren zwar erkannt, entsprechende Algorithmen können aber Bezeichnungen und Zusammenhänge oft nicht zweifelsfrei auflösen. Wenn nun die Analyse von Texten unterschiedlicher Sprachen und Herkunft Korrelationen erkennt, die einer gegenseitigen Plausibilitätsprüfung standhalten, so kann für entsprechende Ereignisse mit einer höheren Wahrscheinlichkeit angenommen werden, dass diese auch richtig erkannt wurden.

Ein einfaches Beispiel: Die Abbildung zeigt einen Überblick über erkannte Ort / Zeit-Korrelationen im Lebenslauf Friedrich Schillers. Schiller wurde nach den Angaben in Wikidata 1759 geboren. Gegen diese Information können nun die Ergebnisse von Volltextanalysen so geprüft werden, dass algorithmisch erkannte Ereignisse für das Leben Schillers in den Jahren 1710 oder 1880 als unplausibel erkannt werden. Die Farbgebung der Ereignisse in der Zeitleiste deutet

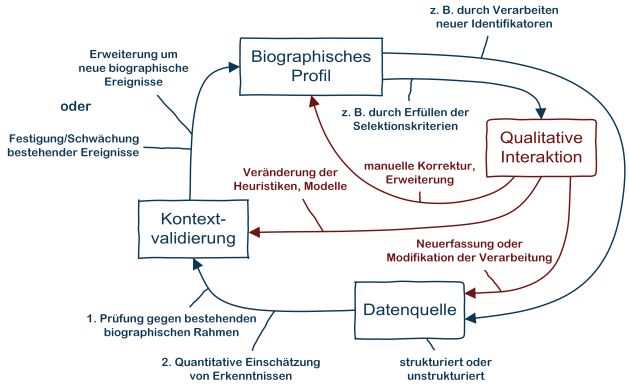
die Sicherheit der Einträge an: grün steht hierbei für gesicherte Erkenntnisse, gelbe Knoten deuten auf ein unbelegtes Ereignis aus der quanitativen Textanalyse hin. Der steigendem Abstand der Knoten von der Zeitleiste spiegelt eine steigende Unsicherheit der Ereignisse im Kontext des biographischen Rahmens wider.



Iterativer Verarbeitungsprozess

Die Umsetzung des Prototypen basiert auf einem generischen Framework für die Korrelation, Verarbeitung und Transformation von Daten, welches ursprünglich für die generische Suche von DARIAH-DE entwickelt wurde und dieser auch zu Grunde liegt. Das Framework zeichnet sich insbesondere dadurch aus, dass eine Phase der deskriptiven Datenmodellierung von der Spezifikation der Verarbeitungslogik getrennt wird (vgl. Gradl / Henrich 2014). Im Wesentlichen wird dadurch erreicht, dass geisteswissenschaftliche Experten die Forschungsdaten ihrer jeweiligen Disziplin um expliziertes Wissen zum Erstellungskontext der Daten anreichern können. Andere Forschende können auf Basis der angereicherten Datenbeschreibung nun Verarbeitungsregeln so spezifizieren, dass die erweiterten Daten in den gewünschten Verwendungskontext transformiert werden. [2]

An konkreten Beispiel der Verarbeitung biographischer Daten resultiert die Anwendung des Frameworks und des zu Grunde liegenden Konzepts in einer iterativen, kontextspezifischen Verarbeitungslogik, die in der folgenden Anwendung skizziert wird und das Zusammenspiel zwischen qualitativer Forschung und quantitativen Verfahren am Beispiel des *Cosmobilities* Prototypen verdeutlicht.



So haben Forscher die Möglichkeit an drei Stellen des Prozesses manuell einzuwirken und die quantitative Verarbeitung zu beeinflussen: Zunächst werden durch die Erfassung einer Datenquelle bzw. der Beschreibung ihrer Datenstrukturen (deskriptive Datenmodellierung) biographische Daten und Texte erfasst. Erkenntnisse, die durch eine angewendete Transformation der Daten extrahiert werden können werden ggf. in den Kontext bestehender biographischer Rahmenbedingungen gesetzt und in biographische Profile übernommen.

Auf eben diese qualitativen Einschätzung können Forscher an zwei wesentlichen Stellen einwirken: Einerseits besteht die Möglichkeit, die Einordnung biographischer Daten durch die Beschreibung von Modellen und Heuristiken zu beeinflussen. Eine vereinfachte Heuristik wird in der folgenden Abbildung dargestellt. Hier würde beispielsweise ein Versterben der Mutter zu einem Eintrag im biographischen Profil des Kindes führen, welcher den Aufenthaltsort des Kindes, insofern dieses zu diesem Zeitpunkt höchstens 16 Jahre alt war, mit einer hohen Wahrscheinlichkeit mit dem Sterbeort der Mutter korreliert. An Stelle einer solchen einfachen Heuristik könnten auch komplexere, epochenspezifische Betrachtungen, wie z. B. den Lebensalterdarstellungen von Wirag (Wirag 1994) oder Anwendungen von Lebensstufenmodellen (z. B. von Grayerz 2010) nach Anforderungen der jeweiligen Forscherperspektive stehen.

```
// Child born
getClaimsForRelatives(h, h.getChild(), true, false, 0, 0, 0.9, "Kind geboren");

// Child died
getClaimsForRelatives(h, h.getChild(), false, true, 0, 30, 0.9, "Kind verstorben");
getClaimsForRelatives(h, h.getChild(), false, true, 31, 40, 0.7, "Kind verstorben");

// Spouse died
getClaimsForRelatives(h, h.getSpouse(), false, true, 0, 0, 0.7, "Partner verstorben");

// Parents died
getClaimsForRelatives(h, h.getMother(), false, true, 0, 16, 0.9, "Mutter verstorben");
getClaimsForRelatives(h, h.getFather(), false, true, 0, 16, 0.9, "Vater verstorben");
```

Die zweite Möglichkeit der qualitativen Beeinflussung besteht in der konkreten Veränderung des biographischen Rahmens, also die manuelle Erfassung oder Korrektur wesentlicher Eckpunkte wie Geburts- und Sterbedaten der einzelnen Person oder auch seiner nächsten Verwandten. Ein weiterer Iterationszyklus folgt schließlich, wenn ein verändertes Profil die definierten Selektionskriterien einer

Forscherin erfüllt und in deren Fokus rückt bzw. wenn ein nun erweitertes Profil neue Hinweise auf weitere Datenquellen beinhaltet. Solche Daten können IDs in Datenbanken sein, aber auch die Vervollständigung eines Geburtsname / Geburtsdatum-Tupels, auf dessen Basis die Suche nach weiteren biographischen Texten fortgesetzt werden kann.

Ausblick

Weitere Entwicklungsschritte sind notwendig um den beschriebenen Verarbeitungszyklus im Rahmen des Prototypen vollständig abzubilden und die Interaktion zwischen qualitativen Verfahren und der qualitativen Forschung anbieten zu können.

Parallel hierzu werden derzeit auch Möglichkeiten zur Aggregation individueller Profile untersucht, um Rückschlüsse über die Transnationalität von Personengruppen anbieten und entsprechende Internationalitätskriterien ableiten zu können.

Notes

1. Für eine differenzierte historische Betrachtung des Themas verweisen wir an dieser Stelle auf das Werk von Deacon, Russel und Woolacott (2010).
2. Weitere theoretische Überlegungen finden sich in Gradl / Henrich (2014); eine Ausarbeitung, die sich mit diesem Konzept technisch weiterführend auseinandersetzt wird derzeit vorbereitet.

Bibliographie

Bamman, David / Smith, Noah A. (2014): "Unsupervised Discovery of Biographical Structure from Text", in: *Transactions of the Association for Computational Linguistics* 2: 363-376.

Blessing, André / Kuhn, Jonas (2014): "Textual Emigration Analysis (TEA)", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation* 2089-2093.

Deacon, Desley / Russel, Penny / Woolacott, Angela (2010): *Transnational Lives*. Biographies of Global Modernity. 1700-present. Basingstoke / Hampshire: Palgrave Macmillan.

Gradl, Tobias / Henrich, Andreas (2014): "A novel approach for a reusable federation of research data within the arts and humanities", in: *Digital Humanities 2014*. Book of Abstracts, Ecole polytechnique federale de Lausanne; Lausanne: Université de Lausanne 382–384 http://dh2014.org/program/abstracts/ [letzter Zugriff 09. Oktober 2015].

Grayerz, Kaspar von (2010): *Passagen und Stationen*. Lebensstufen zwischen Mittelalter und Moderne. Göttingen: Vandenhoeck & Ruprecht.

Heilbrunn, Bernice (2011-201): "Jacob H. Schiff", in: Hoyt, Giles R. (ed.): *Immigrant Entrepreneurship*. German-American Business Biographies 1720 to the Present. Vol. 3. Washington: German Histprical Institute http://immigrantentrepreneurship.org/entry.php?rec=41 [letzter Zugriff 07. Februar 2016].

Lei, Tao / Long, Fan / Barzilay, Regina / Rinard, Martin (2013): "From Natural Language Specifications to Program Input Parsers", in: *The 51st Annual Meeting of the Association for Computational Linguistics* 1294-1303.

Wikipedia (22.11.2015): "Jacob H. Schiff" https://en.wikipedia.org/wiki/Jacob_Schiff [letzter Zugriff 07. Februar 2016].

Wikipedia (07.02.2016): "Jakob Heinrich Schiff" https://de.wikipedia.org/wiki/Jakob_Heinrich_Schiff [letzter Zugriff 07. Februar 2016].

Wirag, Klaus T. (1994): *Cursus Aetatis*. Lebensalterdarstellungen vom 16. bis zum 18. Jahrhundert. München: Univ. Diss.

Judaica recherchieren – Unterstützung bei der Realisierung forschungsspezifischer Suchlösungen durch die generische Suche von DARIAH-DE

Gradl, Tobias
tobias.gradl@uni-bamberg.de
Universität Bamberg

Lordick, Harald
lor@steinheim-institut.org
Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte

Henrich, Andreas
andreas.henrich@uni-bamberg.de
Universität Bamberg

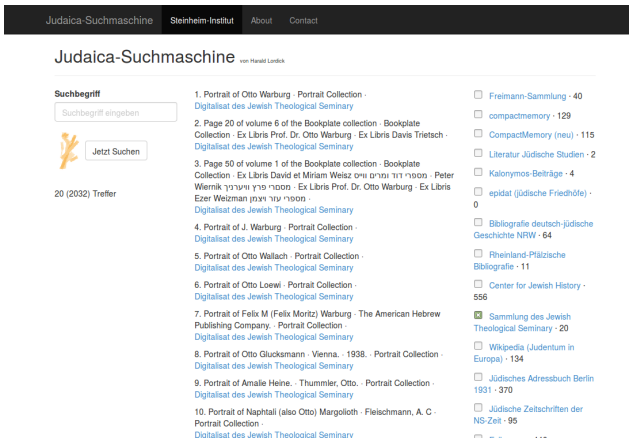
Einleitung

Jenseits der standardisierten und institutionalisierten Bereitstellung von Forschungsquellen und -literatur bedarf es weiterführender Recherchekonzepte und Anwendungen, die Geisteswissenschaftler_innen mit ihren spezifischen Forschungen und den *für sie*

verfügbaren und relevanten Daten so weit wie möglich entgegenkommen. Integrative Suchlösungen wie OAIster und Europeana bieten Zugriff auf eine Vielzahl von Kollektionen und unterstützen breite, fachunabhängige Suchanfragen im strukturellen Rahmen integrativer Schemata (Hagedorn 2013; Peroni et al. 2013). Der semantisch tiefe und forschungspezifische Zugang zu Ressourcen steht bei solchen Suchportalen dabei zumeist nicht im Fokus. Aus diesem Grund stehen den Suchportalen spezifische Lösungen gegenüber, die speziell an die Bedürfnisse des einzelnen Forschers und seine aktuellen Fragestellungen angepasst sind. Ein Beispiel bildet hierbei die am Steinheim-Institut entwickelte Judaica-Suchmaschine (Lordick 2013). In diesem Beitrag stellen wir mit der generischen Suche von DARIAH-DE eine weitere breite Suchlösung vor, welche jedoch auch Funktionen spezifischer Ansätze realisiert, wie etwa die individuelle Aggregation und Filterung von Sammlungen, sowie differenzierte Möglichkeiten der Zugriffskontrolle. Auf Basis des Anwendungsfalls der Judaica-Suchmaschine zeigen wir exemplarisch die Unterstützung individueller Suchbedürfnisse im Rahmen der generischen Suche und verdeutlichen dabei, wie neue, fachspezifische Suchen verfügbar gemacht werden können - ohne dass die hierfür notwendigen, technischen Aspekte durch den einzelnen Forscher umzusetzen sind.

Die Judaica-Suchmaschine als Anwendungsfall

Als Use-Case dient die langjährige Auseinandersetzung mit den Möglichkeiten und Grenzen einer übergreifenden *Judaica-Suchmaschine* im Steinheim-Institut. Sie war ursprünglich als Allegro-C-Katalog gestartet, der verschiedene institutsinterne Datenbanken in eine gemeinsame Datenbank mit dem Ziel der effizienteren fachspezifischen Recherche zusammenführte. Bald darauf wurde begonnen, auch passende externe Datenangebote einzubinden. Die Suchmaschine enthält zur Zeit ca. 500.000 Datensätze aus 20 filterbaren Katalogen, ist auch für mobile Geräte geeignet (Lordick 2014) und XML-basiert.



Über Standardschnittstellen lassen sich beispielsweise integrieren: die Freimann-Sammlung mit 8.772 Titeln (Teil von Judaica Frankfurt), CompactMemory 46.637 (seit 2014 ebenfalls Judaica Frankfurt), Center for Jewish History (New York) 65.667 oder Jewish Theological Seminary 9.257. Die ca. 15.000 digitalisierten Seiten der Sammlung Jüdische Zeitschriften der NS-Zeit der Deutschen Nationalbibliothek sind zwar 2012 abgeschaltet worden. Die Suchmaschine enthält aber noch die 34.346 erschließenden Metadatensätzen dazu. Sie haben nun 'nur noch' bibliografischen Charakter, der um so wertvoller bleibt. Gleiches gilt für die Exilpresse digital: Deutsche Exilzeitschriften 1933–1945 mit nicht weniger als 231.548 Metadatensätzen.

In diesen Kontext gehört auch der Ansatz, eigene Daten ebenfalls über etablierte Standardformate und Protokolle anzubieten, um sie zur Nachnutzung zur Verfügung zu stellen. So wurde ein OAI-PMH Testserver[1] eingerichtet, der digitale Sammlungen des Steinheim-Instituts zusammenführt: Universal-Kirchenzeitung, Kalonymos, Deutsch-jüdische Publizistik. Ebenfalls wurde das Verfahren der Bereitstellung mittels eines statischen OAI-PMH-Repositorys geprüft (Beer et al. 2013). Dieser Weg ist empfehlenswert, hinsichtlich des erforderlichen technischen Knowhows jedoch durchaus anspruchsvoll, erfordert (Zugriff auf) Infrastrukturkomponenten und Serverressourcen und eignet sich insbesondere für 'fertige', abgeschlossene Datensätze.

Viele digitale Quellen lassen sich jedoch unter den gegebenen Umständen gar nicht harvesten, etliche verfügbare fachlich interessante oder auch unentbehrliche Datensätze immerhin mittels ind ividueller Programmierung in die Suchmaschine einbinden: das jüngst erschienene Jüdische Adressbuch Berlin 1931, die NS-Liste der verbannten Bücher, das Kapitel Judentum der Rheinland-Pfälzischen Bibliografie oder die 10.600 Artikel der Kategorie Judentum in Europa der Wikipedia etwa.

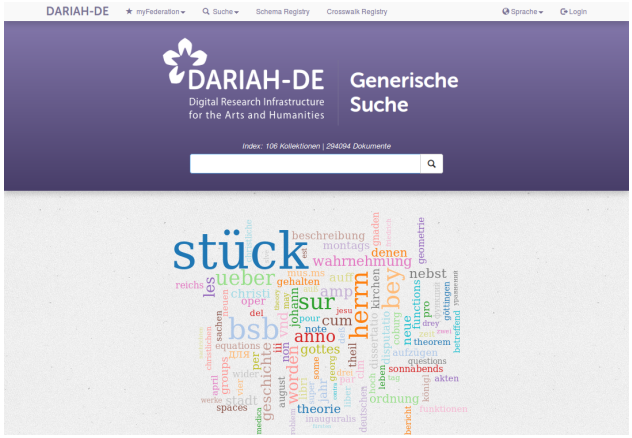
Umsetzung im Rahmen der generischen Suche

Der Umgang mit technischen Problemen zählt nicht zu den typischen Aufgaben von Forschern kultur- und geisteswissenschaftlicher Disziplinen. Eine auf individuelle Bedürfnisse zugeschnittene Integration von Daten oder die Realisierung von Softwarekomponenten ist jedoch für die Umsetzung spezifischer Anwendungen wie der Judaica-Suchmaschine erforderlich. Um Lösungen für technische Probleme in nachhaltiger und für anschließende Forschungsprojekte wiederverwendbarer Form bereitzustellen, beinhalten derzeit laufende Initiativen wie DARIAH oder CLARIN häufig Arbeitspakete zur Umsetzung von Infrastrukturkomponenten.

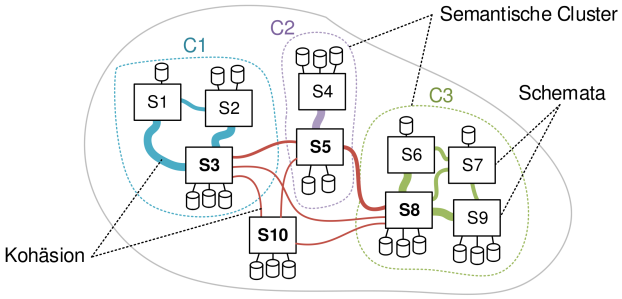
Obwohl die generische Suche von DARIAH-DE (Gradl 2011-2016) nicht als Infrastrukturkomponente, sondern als eigenständiger Dienst entwickelt wurde, basiert auch das Konzept der generischen Suche auf dem Ziel der Umsetzung wiederverwendbarer technischer Funktionalität zur Integration, Verarbeitung und Analyse von Daten. Im Folgenden stellen wir eine im Rahmen der generischen Suche implementierte Funktionalität vor, welche es Forschern ermöglicht, spezifisch angepasste Suchlösungen zu generieren und bereitzustellen - ohne die hierzu erforderlichen technischen Aspekte selbst lösen zu müssen.

Hintergrund des generischen Konzepts

Konzept und Implementierung der generischen Suche resultieren aus der Zielsetzung der Realisierung einer kombinierten und flexibel anpassbar en Breiten- und Tiefensuche (Gradl / Henrich 2014). Dabei wurde zum einen eine universelle und einfache Möglichkeit (vgl. Abbildung) zur übergreifenden Suche in einer Vielzahl digitaler Kollektionen eingerichtet, um Forscher bei der Suche und Analyse relevanter und potenziell unbekannter Ressourcen und Kollektionen zu unterstützen. Neben dieser disziplinunabhängigen Breitensuche erlaubt die generische Suche auch eine kontextspezifische Anpassung ─sowohl im Hinblick auf die Auswahl der zu durchsuchenden Datenquellen, als auch auf das verwendete Schema für die Integration der in diesen enthaltenen, zumeist heterogenen Ressourcen. Die so entstehende Tiefensuche erlaubt eine differenziertere Analyse und Suche in einer forschungsspezifisch begrenzten Datenbasis.



Die flexible Anpassbarkeit der generischen Suche an übergreifende oder spezifische Fragestellungen wird dabei durch das in untenstehender Abbildung exemplarisch angedeutete Konzept der forschungsorientierten Föderation digitaler Kollektionen erreicht (Gradl / Henrich 2014; Gradl et al. 2014).



Anstelle der für übergreifende Integrationssysteme typischen Harmonisierung heterogener Daten (vgl. Batini et al. 1986; Lenzerini 2002) basiert die generische Suche wesentlich auf folgenden Annahmen:

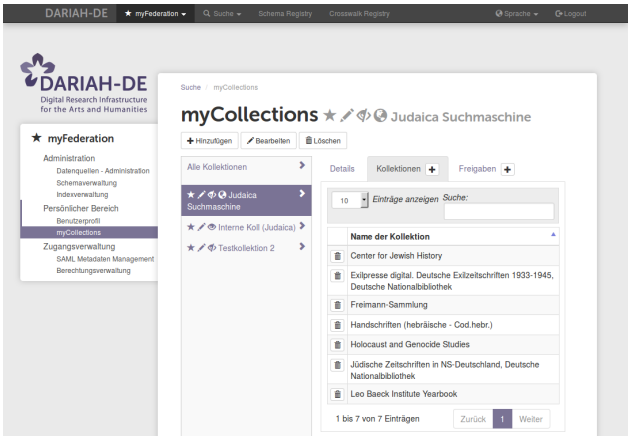
- Durch die semantische Assoziation von Kollektionen und darin verwendeten Schemata durch fachwissenschaftliche Experten können integrative Sichten generiert werden, die den Anforderungen konkreter Forschungsfragen entsprechen. Dies wird insbesondere dadurch erreicht, dass nicht technisch motivierte Integrationsziele wie die Vollständigkeit und Korrektheit eines Integrationsschemas (Batini et al. 1986) im Vordergrund stehen, sondern disziplinspezifische und auch konfliktäre Zusammenhänge modelliert werden können.
- Kohärente Bereiche mit eng assoziierten Schemata *S* und Kollektionen werden in obiger Abbildung durch *semantische Cluster* widergespiegelt und bilden die Voraussetzung für spezifische Tiefensuchen. Für die übergreifende Breitensuche werden wichtige Schemata der einzelnen Cluster (repräsentiert als *S3*, *S5* und *S8*) mit generischen Schemata, wie z. B.

Dublin Core - in der Abbildung symbolisiert durch *S10* - assoziiert.

Die besondere Eigenschaft des Föderationskonzepts besteht darin, dass Daten in ihrer ursprünglichen Form analysiert und indexiert werden. Erst zum Anfragezeitpunkt und in Abhängigkeit von der einer Anfrage zu Grunde liegenden Zusammenstellung von Kollektionen werden die Daten zusammengeführt und integriert.

Abbildung der Cluster in der generischen Suche

Die generische Suche beschreibt einen im Rahmen des DARIAH-DE Projektes entwickelten Dienst, welcher im Hinblick auf seine Datenbasis auf die Einträge der DARIAH-DE Collection Registry zurückgreift (Plutte et al. 2014) und registrierten Benutzern die Aufnahme weiterer Datenquellen erlaubt. Neben der Assoziation von Schemata als wesentlicher Teilaspekt (Gradl / Henrich 2014) bildet die Möglichkeit der Auswahl und Gruppierung relevanter Kollektionen im Rahmen der generischen Suche (als *myCollections*) die Basis für die Erstellung angepasster Suchlösungen.

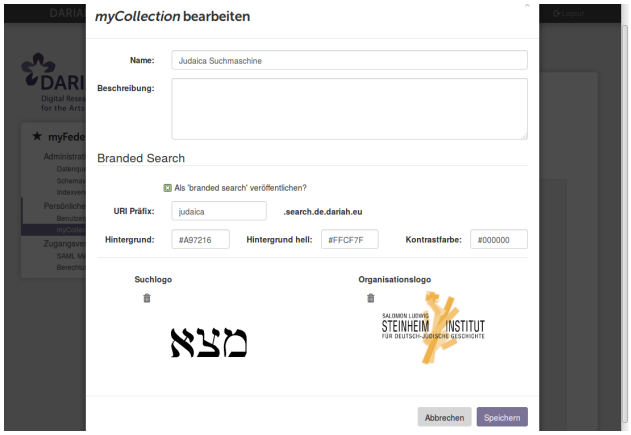


Der Bildschirmausschnitt zeigt exemplarisch drei solcher myCollections und verdeutlicht die derzeit implementierte Funktionalität:

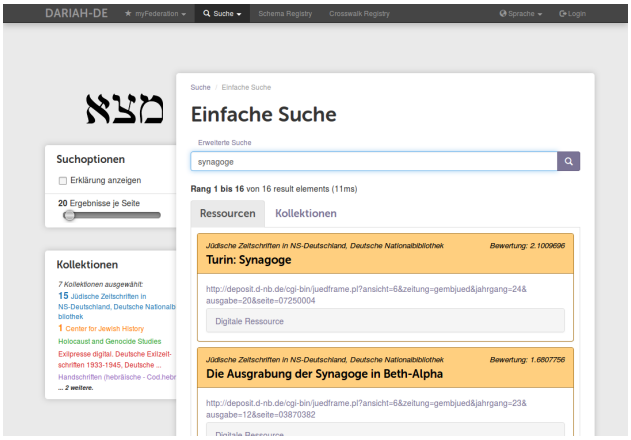
- Einträge von myCollections können durch authentifizierte Anwender angelegt werden und definieren eine Zusammenstellung einzelner Kollektionen.
- Diese myCollections erlauben eine Schnellauswahl von Kollektionen bei der Ausführung von Suchanfragen - sowohl im Rahmen des Benutzerinterfaces, als auch in Form der REST-basierten Schnittstelle.
- Möchte der Benutzer eine myCollection (z. B. im Rahmen eines gemeinsamen Forschungsinteresses) teilen, so können Freigaben für weitere Benutzer oder DARIAH Gruppen erteilt werden.

Spezifischer Zugang in Form einer *Branded Search*

Anknüpfend an den eingeführten Anwendungsfall der Judaica-Suchmaschine wird der Zusammenhang zwischen der generischen Suche von DARIAH-DE und forschungsspezifischen Suchlösungen abgebildet durch die Idee der benutzerdefinierten Zusammenstellung von Kollektionen: Wird eine myCollection - wie in der Abbildung unten dargestellt - als Branded Search ausgezeichnet, so wird dadurch eine eigene Suchoberfläche veröffentlicht, welche sowohl optisch als auch inhaltlich an spezifische Bedürfnisse angepasst und von der eigentlichen generischen Suche abgegrenzt ist.



Die folgenden Bildschirmausschnitte verdeutlichen insbesondere die optische Abgrenzung durch konfigurierbare Farbgebung und die Verwendung von Such- und Organisationslogos. Insbesondere der Vergleich der Wordclouds der Startseiten von generischer Suche und (Gradl / Lordick 2015-2016) deuten jedoch die jeweils unterschiedliche, zu Grunde liegende Datenbasis an: Die in einer Branded Search angebotenen Kollektionen spiegeln bei sämtlichen Such-, Analyse- und Visualisierungaufgaben die von den Erstellern der Suche getroffene Kollektionsauswahl wider.



Im Fall der Veröffentlichung einer Branded Search bleibt diese auch als zugreifbare myCollection für berechtigte Benutzer erhalten und kann im Hinblick auf die Kollektionsauswahl und die Assoziation der Schemata verändert werden ─mit unmittelbaren Auswirkungen auf die entsprechende Branded Search. Von technischen Implementierungen an der Basis der generischen Suche, z. B. der Anbindung weiterer Quellenarten wie Wikipedia können schließlich sämtliche eingerichteten Branded Searches profitieren, sofern dies durch die jeweiligen Forscher gewünscht wird.

Ausblick

Ein ausgeprägt generischer Ansatz muss kein Gegensatz zu den in den Geisteswissenschaften vorherrschenden individuellen Fragestellungen und Forschungsansätzen sein. Indem sie entsprechende Freiheitsgrade, kreatives und kollaboratives Datenmanagement anbietet, erlaubt die generische Suche ihren Nutzern die Erstellung eigener, jeweils individuell ausgelegter Suchmaschinen.

Es ist das Knowhow der Forschenden, das die Relevanz der Daten, die sie zusammenstellen, filtern, teilen, auch ad-hoc zum Zwecke der Recherche bereitstellen, ausmacht. Ein solches Framework, verbunden mit der fachspezifischen Kenntnis der Daten ist eine gute Basis für überraschende Funde und das Aufspüren unerwarteter Zusammenhänge.

Notes

1. am Jülich Supercomputing Center, DARIAH-DE.

Bibliographie

Batini, Carlo / Lenzerini, Maurizio / Navathe, Shamkant Bhalchandra (1986): "A comparative analysis of methodologies for database schema integration", in: *ACM Computing Surveys* 18, 4: 323–364.

Beer, Nikolaos / Herold, Kristin / Kolbmann, Wibke / Kollatz, Thomas / Romanello, Matteo / Rose, Sebastian / Walkowski, Niels-Oliver (2013): *Recommendations for Interdisciplinary Interoperability (R 3.3.1)*. DARIAH-DE report https://dev2.dariah.eu/wiki/download/attachments/14651583/R3.3.1.pdf?version=1&modificationDate=1366904278298&api=v2 [letzter Zugriff 07. Februar 2016].

Europeana Foundation (2008-2015): *Europeana Collections*. Den Haag http://www.europeana.eu/portal/ [letzter Zugriff 07. Februar 2016].

Gradl, Tobias (2011-2016): *DARIAH-DE Generic Search* http://search.de.dariah.eu [letzter Zugriff 07. Februar 2016].

Gradl, Tobias / Henrich, Andreas (2014): "A novel approach for a reusable federation of research data within the arts and humanities", in: *Digital Humanities 2014*. Book of Abstracts382–384 http://dh2014.org/program/abstracts/ [letzter Zugriff 09. Oktober 2015].

Gradl, Tobias / Henrich, Andreas / Plutte, Christoph (2015): "Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Förderation von Kollektionen", in: Constanze Baum / Thomas Stäcker (eds.): *Grenzen und Möglichkeiten der Digital Humanities*. Sonderband der Zeitschrift für digitale Geisteswissenschaften 1 http://dx.doi.org/10.17175/sb001_020 [letzter Zugriff 09. Oktober 2015].

Gradl, Tobias / Lordick, Harald (2015-2016): *Judaica Search*. Branded Search in the DARIAH-DE Generic Search http://judaica.search.de.dariah.eu [letzter Zugriff 07. Februar 2016].

Hagedorn, Kat (2003): "OAIster: a 'no dead ends' OAI service provider", in: *Library Hi Tech* 21, 2: 170–181.

Henrich, Andreas / Gradl, Tobias (2013): "DARIAH(-DE): Digital research infrastructure for the arts and humanities - concepts and perspectives", in: *International Journal of Humanities and Arts Computing* 7: 47–58.

Lenzerini, Maurizio (2002): "Data integration: a theoretical perspective", in: *PODS'02 - Proceedings*

of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems : 233-246 http://dl.acm.org/citation.cfm?doid=543613.543644 [letzter Zugriff 09. Oktober 2015].

Lordick, Harald (2010-2016): *Vieles finden*. Die Judaica-Suchmaschine im Steinheim-Institut. Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte an der Universität Duisburg-Essen http://steinheim-institut.de/vf/ [letzter Zugriff 07. Februar 2016].

Lordick, Harald (2013): "Die judaica-bibliothek im web. ganz real oder noch immer virtuell?", in: *Kalonymos*1: 12–14 http://www.steinheim-institut.de/edocs/kalonymos/kalonymos_2013_1.pdf [letzter Zugriff 09. Oktober 2015].

Lordick, Harald (2014): "Jüdische Geschichte (mobil) recherchieren", in: *Kalonymos*3: 13 http://www.steinheim-institut.de/edocs/kalonymos/kalonymos_2014_3.pdf [letzter Zugriff 9. Oktober 2015].

OAIster (2001-2016): *OAIster*. OCLC WorldCat.org Services. Dublin / Ohio: OCLC Online Computer Library Center http://oaister.worldcat.org/ [letzter Zugriff 07. Februar 2016].

Peroni, Silvio / Tomasi, Francesca / Vitali, Fabio (2013): "Reflecting on the europeana data model", in: Agosti, Maristella / Esposito, Floriana / Ferilli, Stefano / Ferro, Nicola (eds.): *Digital Libraries and Archives*. Berlin / Heidelberg: Springer 228–240 http://link.springer.com/chapter/10.1007%2F978-3-642-35834-0_23 [letzter Zugriff 09. Oktober 2015].

Plutte, Christoph (2011-2014): *DARIAH-DE Collection Registry*. Initial Prototype. http://colreg.de.dariah.eu [letzter Zugriff 07. Februar 2016].

Modelling the Scholarly Domain beyond Infrastructure

Gradmann, Stefan
stefan.gradmann@kuleuven.be
Katholieke Universiteit Leuven, Literary Studies Research Unit, Belgien

Hennicke, Steffen
hennicke@leibniz-gei.de
Georg-Eckert-Institut, Leibniz-Institut für internationale Schulbuchforschung, Deutschland

Tschumpel, Gerold
gerold.tschumpel@student.hu-berlin.de
Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft, Deutschland

Dill, Kristin
kristin.dill@univie.ac.at
Universität Wien, Institut für Germanistik, Österreich

Thoden, Klaus
kthoden@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Berlin, Deutschland

Pichler, Alois
alois.pichler@fof.uib.no
Universität Bergen, Philosophisches Institut, Norwegen

Morbindoni, Christian
christian.morbidoni@gmail.com
Semedia, Università Politecnica delle Marche, Ancona, Italien

Stiller, Juliane
jstiller@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Berlin, Deutschland

This paper presents the findings of research conducted in the EU-funded project Digitised Manuscripts to Europeana [1] (DM2E), which investigated how the modelling of scholarly practices and research processes might inform the development of research environments for digital scholarship in the humanities. In particular, the *Scholarly Domain Model* (SDM) will be presented as a framework for modelling the domain of digital scholarship, which provides the constituents for the systematic enquiry of continuously evolving Virtual Research Environments (VRE) [2] and the emergence of digital practices and methodology within them. The importance of models and a practice of modelling cannot be overestimated for the creation of research environments that advance beyond a level of infrastructure, and achieve sustainability through the focus on the evolving scholarly practices at the heart of the transition that altered the humanities profoundly in the last decades.

International institutions of research funding have contributed tremendous efforts into that transitory process and have encouraged a variety of projects for the advancement of the Digital Humanities [3], focusing on attempts to further the development of infrastructures for digital scholarship in the humanities. In Europe, for example, the European Strategy Forum on Research Infrastructures (ESFRI) has funded several infrastructure projects such as the Digital Research Infrastructures for the Arts and Humanities (DARIAH) and the Common Language Resources and Technology Infrastructure (CLARIN), which have since been complemented by

the Data Service Infrastructure for the Social Sciences and Humanities (DASISH). Each of these infrastructure projects have, in turn, influenced a number of other endeavours on a national, regional, institutional or disciplinary level.

Nevertheless, achieving a constellation of constituents and influential factors that facilitate the prospective sustainability of a Virtual Research Environment as a socio-technical system, is still an unresolved problem. Whereas this certainly is due to many reasons, we believe that among them a deficit of systematic investigation into the actual research practices of humanists and their sustainable representation in the digital realm is of crucial importance. We consider the inclusion of the scholars essential as the actors of a community of practice, who constitute precisely with this scholarly practice the basis for the development of research environments and infrastructures.

In this context, the research gap we identified and attempt to address is the lack of a model, which emphasises the importance of creating a bridge connecting the analogue and digital scholarly practices and, most importantly, underlines the recursive relationship between these scholarly practices and the models and applications reflecting on them. This kind of research falls within what is typically called 'digital humanities' and which we understand as a community of practices, regardless of their particular materiality. We therefore believe that in order to be able to discuss the 'digital humanities' in a way that goes beyond simply discussing infrastructure, and so that the aforementioned challenge can be overcome, we need to start from a modelling process that allows for the systematic and theoretically grounded integration of practices of humanist research approaches in both the analogue and digital world. In this paper, we discuss this undertaking and propose a multi-layered model, the SDM, that exemplifies the constituents of our modelling endeavour. For this reason, the SDM is conceived as an explicit but not definite set of constituents of the domain of digital scholarship in the humanities. In his presentations Manfred Thaller has repeatedly stressed, that the controversy of the 'digital humanities' should rather focus on the scholarly practices in the digital humanities and in particular their prerequisites, the various epistemological implications that the application of digital technology entail, than to be predominated by arguments about labelling (cf. Thaller 2013, 2015a, 2015b; McCarty / Short 2002).

In this regard, Linked Data standards [4] such as the Resource Description Framework (RDF), Resource Description Framework Schema (RDFS), and Web Ontology Language (OWL) constitute a well suited means for the development of the SDM, because they allow the process of modelling to be iterative and continuous since the graph of semantic statements created is extensible. Furthermore, it facilitated the development of the modules of a digital humanities research environment, which has been built around the semantic annotation application Pundit . As we will see, this is also an instance of a still uncommon and emerging way to think of Linked Data as an art with epistemological implications for the practice of modelling the domain of digital scholarship in the humanities (cf. Oldman et al. 2016).

Like other models since, the SDM takes up the notion of Scholarly Primitives (cf. Unsworth 2000) and develops them further. On the basis of the analysis and observation of the practices of digital scholarship, we are endeavouring to acquire a better understanding of the requirements for instructing the development of sustainable infrastructures that enable scholars to harness the potential of digital technology and hence to develop appropriate digital methodologies and practices. This requires to proceed beyond the establishment of static models to the iterative and continuous activity of modelling. Starting from the Scholarly Primitives by Unsworth (2000), the SDM was further constructed and refined by analysing the research literature and related models (cf. Atkins et al. 2003; Project Bamboo 2010; Benardou et al. 2010; Palmer et al. 2009). Furthermore, the conceptual input has been subsequently revised and supplemented by empirical evidence collected through a series of interviews with scholars and researchers from the humanities, and experiments using the Linked Data annotation environment Pundit. Finally, the development of the SDM has continuously been monitored and counselled by an advisory board of Digital Humanists .

Furthermore, the SDM differs from the work done, for example, by DARIAH and the Network of Digital Methods in the Arts and Humanities (NeDiMAH) [5] in so far as it approaches the scholarly domain from a more comprehensive perspective that tries to integrate Primitives and constituents influencing the processes of digital scholarship in the humanities and to reflect on their social construction on different layers of abstraction. Furthermore, we believe that a continuous and recursive process of modelling is ultimately the goal, not the model itself. With reference to Willard McCarty (2003, 2004 and 2005), resonating a distinction, introduced by Geertz (1973) between a model 'of' something and a model 'for' something, the SDM in that sense is a descriptive model 'of' the scholarly domain and as imperfect it may be, it is built for a purpose and it may fail. The benefit of this failure, for that matter, is that it emphasises the importance of modelling in many respects. As for the modelling of scholarly practices of the digital humanities may further the self-reflection and pattern discovery resulting in new models 'of' these practices as well as in models 'for' the conduct of such new modes of digital scholarly activities in the Virtual Research Environments developed concomitantly. The SDM attempts to provide an explicit but not definite set of constituents to initiate a self-reflected development of Research Environments for Digital Scholarship in the Humanities on the basis of scholarly practices going beyond infrastructure.

In sum, the SDM is a framework for better understanding scholarly research practices and the ways digital working modes might evolve in the future. Despite the fact that the SDM has been devised in the context of applications based on Linked Data, the model is independent from particular representations and meant to be applicable as a reference model for the discussion, evaluation and development of digital research infrastructures and environments for the humanities. The SDM allows to create representations of the workflows of digital humanists and to function as a terminological bridge between the scholarly practices of the humanities and digital applications. The goal is to reflect on the social nature of scientific practice, regardless of its materiality, and, based on such reflections, to receive a stable core for its sustainable representation. Only if we better understand how scholars undertake their research and how their functional framework might be adequately translated to the digital environment, we might actually approach the emergence of new digital modes of working.

Notes

1. The project ran from Feb. 2013 until Jan. 2015. URL: http://dm2e.eu/ [last accessed 15. October 2015]; for a more detailed account cf. further Hennicke et al. 2015.
2. Cf. Candela et al. 2013. Respective Research Infrastructure.
3. We understand this term to be grounded in the basis of the translation of the German word for *Geisteswissenschaften* and not in the political sense (also cf. Gold 2012; Terras et al. 2013).
4. Cf. for the following standards http://www.w3.org/standards/techs/rdf and http://www.w3.org/standards/techs/owl [all last accessed 15. October 2015].
5. In particular, the Methods Ontology (NEMO); cf. further Hughes et al. 2016 for a more detailed account.

Bibliographie

Anderson, Sheila / Blanke, Tobias / Dunn, Stuart (2010): "Methodological commons: arts and humanities e-science fundamentals", in: *Philosophical Transactions of the Royal Society A* Mathematical, Physical and Engineering Sciences 368, 1925: 3779-3796 http://rsta.royalsocietypublishing.org/content/368/1925/3779.short [last accessed 15. October 2015].

Atkins, Daniel E. / Droegemeier, Kelvin K. / Feldman, Stuart I. (2003): *Revolutionizing science and engineering through cyberinfrastructure*. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Ann Arbor: University of Michigan Library Press http://www.nsf.gov/cise/sci/reports/atkins.pdf [last accessed 15. October 2015].

Project Bamboo (2010): *Project bamboo scholarly practice report* https://googledrive.com/host/0B3zU098zQ8VMc2xfMUJZaWxXNWs/wp-content/uploads/Project-Bamboo-Scholarly-Practices-Report.pdf [last accessed 15. October 2015].

Benardou, Agiatis / Constantopulos, Panos / Dallas, Costis / Gavrilis, Dimitris (2010): "A conceptual model for scholarly research activity", in: Reilly, Maeve (ed.): *iConference Papers 2010* https://www.ideals.illinois.edu/bitstream/handle/2142/14945/benardou.pdf?sequence=2 [last accessed 15. October 2015].

Candela, Leonardo / Castelli, Donatella / Pagano, Pasquale (2013): "Virtual research environments: An overview and a research agenda", in: *Data Science Journal* 12: GRDI75-GRDI81.

CLARIN: *CLARIN*. Common Language Resources and Technology Infrastructure http://www.clarin.eu/ [last accessed 15. October 2015].

DARIAH: *DARIAH-EU*. Digital Research Infrastructure for the Arts and Humanities http://www.dariah.eu/ [last accessed 15. October 2015].

DASISH: *DASISH*. Data Service Infrastructure for the Social Sciences and Humanities http://dasish.eu/about_dasish/ [last accessed 15. October 2015].

DM2E: "Digital Humanities Advisory Board" http://dm2e.eu/dhab/ [last accessed 15. October 2015].

ESFRI: *ESFRI*. European Strategy Forum on Research Infrastructures http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri [last accessed 15. October 2015].

Geertz, Clifford (1973): *The interpretation of cultures*. Selected essays. New York: Basic Books.

Gold, Matthew K. (2012): *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Hennicke, Steffen / Gradmann, Stefan / Dill, Kristin / Tschumpel, Gerold / Thoden, Klaus / Morbidoni, Christian / Pichler, Alois: *D3.4*. Research Report on DH Scholarly Primitives http://dm2e.eu/files/D3.4_2.0_Research_Report_on_DH_Scholarly_Primitives_150402.pdf [last accessed 15. October 2015].

Hughes, Lorna / Constantopoulos, Panos / Dallas, Costis (2016): "Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines", in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A new Companion to Digital Humanities*. New York: John Wiley & Sons 150-170.

McCarty, Willard (2002): "Humanities computing: Essential problems, experimental practice", in: *Literary and Linguistic Computing* 17, 1: 103-125 http://llc.oxfordjournals.org/content/17/1/103.full.pdf+html [last accessed 15. October 2015].

McCarty, Willard (2003): "Knowing true things by what their mockeries be: Modelling in the humanities", in: Woolridge, Russon / McCarty, Willard / Winder, William (eds.): *Computing in the Humanities Working Papers* http://journals.sfu.ca/chwp/index.php/chwp/article/view/A.24/52 [last accessed 15. October 2015].

McCarty, Willard (2004): "Modeling: A study in words and meanings", in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A Companion to Digital Humanities*. Malden: Blackwell 254-272.

McCarty, Willard (2005): *Humanities computing*. Houndmills: Palgrave Macmillan.

McCarty, Willard / Short, Harold (2002): "Mapping the field: Report of ALLC meeting held in Pisa, April 2002" http://www.allc.org/node/188 [last accessed 15. October 2015].

NeDiMAH: *NeDiMAH*. Network for Digital Methods in the Arts and Humanities http://www.nedimah.eu/ [last accessed 15. October 2015].

NeMo: *NeMo*. NeDiMAH Methods Ontology http://nemo.dcu.gr/ [last accessed 15. October 2015].

Net7: *Pundit* http://thepund.it/ [last accessed 15. October 2015].

Oldman, Dominic / Doerr, Martin / Gradmann, Stefan (2016): "Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge", in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A new Companion to Digital Humanities*. New York: John Wiley & Sons 251-273.

Palmer, Carole L. / Teffeau, Lauren C. / Pirmann, Carrie M. (2009): *Scholarly Information Practices in the Online Environment*. Themes from the Literature and Implications for Library Service Development. Dublin, OH: OCLC Research http://www.oclc.org/content/dam/research/publications/library/2009/2009-02.pdf [last accessed 15. October 2015].

Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.) (2004): *A Companion to Digital Humanities*. Malden: Blackwell http://www.digitalhumanities.org/companion/ [last accessed 15. October 2015].

Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.) (2016): *A new Companion to Digital Humanities*. New York: John Wiley & Sons.

Terras, Melissa / Nyhan, Julianne / Vanhoutte, Edward (eds.) (2013): *Defining Digital Humanities*. A Reader. Farnham: Ashgate.

Thaller, Manfred (2013): *Praising Imperfection*. Why editions do not have to be finished http://www.culingtec.uni-leipzig.de/ESU_C_T/node/292 [last accessed 15. Oktober 2015].

Thaller, Manfred (2015a): *Digital Humanities*. Eine Bestandsanalyse https://de.dariah.eu/documents/10180/472723/Thaller_Digital+Humanities+-+eine+Bestandsanalyse.pdf [last accessed 15. October 2015].

Thaller, Manfred (2015b): *Wenn die Quellen überfliessen*. Spitzweg und Big Data https://static.uni-graz.at/fileadmin/veranstaltungen/von-daten-zu-erkenntnissen/thaller_keynote.pdf [last accessed 15. October 2015].

Unsworth, John (2000): "Scholarly Primitives: What methods do humanities researchers have in common, and how might our tools reflect this?" In: *A symposium on Humanities Computing: formal methods, experimental practice sponsored by King's College, London, May 13, 2000* http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html [last accessed 15. October 2015].

Play(s): Crowdbasierte Anreicherung eines literarischen Volltext-Korpus

Göbel, Mathias

mathiasgoebel@web.de
Universität Göttingen, Seminar für Deutsche Philologie

Meiners, Hanna-Lena

h-l.meiners@gmx.de
Universität Göttingen, Seminar für Deutsche Philologie

Digitale Korpora entstehen unter bestimmten Voraussetzungen, werden von verschiedensten Institutionen gefördert und haben unterschiedliche Ziele in Bezug auf Qualität und Quantität. In vielen Fällen müssen Forscherinnen weitere Verbesserungen vornehmen, bestehende Daten erweitern und verbessern, im schlimmsten Fall auch neu erheben. In diesem Paper beschreiben wir eine Vielzahl einfach(st)er Probleme, die es zu bewältigen gilt, will man ein Korpus bestehend aus möglichst vielen genuin deutschsprachigen Dramen computergestützt analysieren. Ausgehend von den Beständen des TextGrid Repository (TextGrid Konsortium 2015) soll mittels einer simplen grafischen Oberfläche – abrufbar in jedem Webbrowser – ein Programm zur Verfügung gestellt werden, das sich spielähnlicher Mittel bedient: die Nutzer auffordert, mehrere Level zu durchlaufen, Punkte zu sammeln und mit jeder Eingabe das Korpus zu verbessern, um schließlich Ausgangsmaterial für eine Vielzahl von Fragestellungen zu bieten.

Crowdsourcing, Social Editing und viele verwandte Begriffe sind Konzepte, die innerhalb der Digital-Humanities-Community in den vergangenen Jahren einen kleinen Hype erfahren haben. An Umsetzungen mangelt es, während man sich noch über die Definitionen streitet. Dabei sind die Lösungsansätze sehr vielversprechend – allen voran die von Zooniverse etablierten Projekte, die sich nun auch geisteswissenschaftlichen Themen widmen. Es werden alte Texte transkribiert und jede Person, die über Computer und Internetanschluss verfügt, kann einen aktiven Beitrag leisten und die Forschung unterstützen. Erfahrungen aus bereits gut etablierten Ansätzen, wie dem Projekt DigitalKoot , entwickelt von der Finnischen Nationalbibliothek, zeigen, dass die Anwendung und Umsetzung spielerischer Verfahren

durchaus einen Mehrwert für die Wissenschaft generieren können. Darüber hinaus können Aufgaben gemeistert werden, die durch einen Einzelnen oder auch eine kleinere Forschungsgruppe niemals würden selbst bewältigt werden können. Voraussetzung dafür ist das Interesse und die Teilnahme vieler Personen an der zu bewältigenden Aufgabe. Diese so zu isolieren und danach zu vereinfachen, dass auch Laien damit umgehen können, stellt die Herausforderung dar. Viel Resonanz bekommen Projekte wie EyeWire oder GalaxyZoo , beides naturwissenschaftliche Citizien-Science-Vorhaben. Im Projekt GalaxyZoo geht es um die Klassifizierung und Beschreibung von Galaxien. Das Projekt war in der Lage, 50 Millionen Klassifizierungen zu sammeln, und das innerhalb seines ersten Betriebsjahres (2007, vgl. Prestopnik 2011: 2).

Methodisch betrachtet bieten Konzepte, die auf der zunehmenden Beteiligung der sogenannten *Crowd* aufbauen, ein großes Potential zur Weiterentwicklung oder Herausbildung neuer Forschungsfragen und -themen. Das bereits erwähnte *Crowdsourcing* kann dabei als Überbegriff für verschiedene Ansätze gesehen werden, die zwar z. T. deutlich voneinander abzugrenzen sind, in manchen Bereichen jedoch deutliche Ähnlichkeiten aufweisen. In einem groben Kategorisierungsversuch kann eventuell eine Zweiteilung vorgenommen werden, um einen besseren Überblick über diese verschiedenen und doch ähnlichen Methoden und Konzepte zu gewinnen. *Serious Games*, *Games with a Purpose* und *Meaningful Play* wollen das Spiel als Medium nutzen um bestimmte Inhalte oder Absichten zu transportieren und dem Nutzer ein bestimmtes Problem näher zu bringen. Ansätze wie *Gamification*, *Gameful Design*, *Social Editing* oder *Human Computation* nutzen gezielt einzelne Elemente aus dem Spieldesign, um einen eigentlich spielfremden Kontext anzureichern und durch die Schaffung einer neuen Atmosphäre attraktiver für potentielle Nutzer_innen zu gestalten.

Während Play(s) sich methodisch eher in der zweiten genannten Kategorie wiederfinden soll, ist wichtig zu betonen, dass der Erfolg solcher Projekte eng mit der Entwicklung und dem Design selbst zusammenhängt. Die Oberfläche sollte beispielsweise ansprechend gestaltet bzw. angemessen bezogen auf die Zielgruppe und einfach zu bedienen sein. Als Spielelemente können Levels, das Sammeln von Punkten in Kombination mit einfachen Spielanweisungen dienen. Neben der tatsächlichen Entwicklung einer neuen Anwendung hängt ein großer Teil des Erfolgs von der zu tätigenden Handlung der Teilnehmer_innen ab. Die Aufgabe, die im Rahmen eines Crowdsourcing oder Citizien-Science-Projektes von den Teilnehmer_innen bearbeitet werden soll, ist im besten Fall einfach zu verstehen und in simple Teilbereiche unterteilt. Gleichzeitig unterliegt die Einbindung von freiwilligen, fachfremden Teilnehmer_innen gewissen eher impliziten und wenig ausgesprochenen Regeln. So sollten die Teilnehmenden generell als Partner oder Mitarbeiter_innen betrachtet werden und nicht als günstige Arbeitskräfte. Zudem sollten sie nicht zur Bewältigung von Aufgaben angehalten werden, die eigentlich einfacher und besser von einem Computer ausgeführt werden könnten. Diese Grundethik sollte bei jeder Umsetzung einer neuen Idee zumindest mitbedacht werden, um künftige Teilnehmer_innen nicht zu verärgern oder zu verschrecken.

Die wenigen bisher gesammelten und verfügbaren Erfahrungen aus Projekten für die Geisteswissenschaft sollen nun ausgewertet werden und in die Umsetzung einer neuen Projektidee eingebracht werden. "Play(s)" ist der Name der Anwendung, die sich damit befassen soll, ein literaturwissenschaftliches Volltext-Korpus anzureichern. Das TextGrid-Repositorium bietet dafür optimale Voraussetzungen: alle Texte sind im TEI-Format erfasst und diese Quelle ist frei zugänglich.

In diesem Projekt knüpfen wir an die von einer Projektgruppe (vgl. Trilcke et al. 2015) bereits herausgefilterten Dramen des Repositoriums an. Dabei wurden bereits in einem manuellen Durchgang allen Sprecherinstanzen im Auswahlkorpus eindeutige Namen (IDs) zugewiesen, um eine Ausgangsbasis für Netzwerkanalysen zu schaffen. In diesen Vorarbeiten wurden die genuin deutschen Texte ausgewählt und dabei aus den insgesamt 666 Dramen auf 465 Werke zurückgegriffen. Um diese Analysen mit einer quantitativ und qualitativ erweiterten Quellenbasis zu vertiefen, bedarf es einer noch genaueren Referenzierung. So sollten zum Beispiel die als Sprecher auftretenden Personengruppen aufgelöst werden und zu diesen die beteiligten Akteure genannt werden. Auch eine Klassifizierung des Geschlechtes, der sozialen Stellung und weitere Features sind denkbar, um differenzierte Analysen tätigen zu können. Hier wird deutlich, dass jeder Text einer bestimmten Aufbereitung bedarf, die aber in vielen kleinen Einzelschritten erfolgen kann, da die Informationen und einzelnen semantischen Anreicherungen in ihren Kategorien unabhängig voneinander sind.

Innerhalb des TextGrid-Korpus beschränken wir uns auf die Betrachtung der Dramen und innerhalb derer sind es die Strukturinformationen, die auf Grundlage des XML-Codes Netzwerkanalysen auf Basis des gemeinsamen Auftretens in einer Szene ermöglichen. Gemeinsames Auftreten heißt in diesem Fall, dass innerhalb einer Szene alle Sprecher_innen in Verbindung gebracht werden. Dazu gilt es die einzelnen Akteure ausfindig zu machen, da das Korpus selbst keine Information, wie man sie im TEI-Attribut who (TEI Consortium 2015) erwarten kann, mitliefert. Betrachtet man als Beispielfall das Drama "Fraw Wendelgard" von Nicodemus Frischlin, finden sich innerhalb der Sprecherbenennung drei verschiedene Schreibweisen, die alle auf die Gräfin Wendelgard verweisen: Wendelgard, Wendelgart und Wendelgardt. In einem anderen Werk taucht in einem Dialog zwischen Faust und Mephistopheles ein einziges Mal der Sprecher "Mephistoph" auf. Häufig beobachtet man Akteure, die

mit einem unbestimmten Artikel eingeführt werden, im Folgenden aber mit bestimmtem oder ohne Artikel angegeben werden, wobei offensichtlich ist, dass es sich um die vorangehende genannte Entität handelt.

Die Ursache kann drucktechnisch bedingt in den Buchausgaben liegen, in denen Sprechernamen abgekürzt werden, um Platz und Papier zu sparen, es können auch schlicht Fehler im Satz auftauchen und eine weitere Fehlerquelle kann der Digitalisierungsprozess sein. In all diesen Fällen ist die Korrekturaufgabe denkbar simpel: man muss jene Sprecher zusammenführen, bei denen es sich offensichtlich um die gleiche Person handelt. Getreu der Buchausgaben handelt es sich dabei nicht um Fehler, das Encoding muss hier schlicht um semantische Information erweitert werden, wie es das Attribute who in den TEI Guidelines vorsieht. Dazu zählen auch Fälle, in denen das Markup innerhalb des TextGrid-Korpus fehlerhaft ist. Das betrifft leere speaker-Elemente, solche, in denen noch Teile der Bühnenanweisung mit einfließen und auch jene, die noch ein leeres Element stellvertretend für zum Beispiel einen Seitenumbruch beinhalten und dadurch als Auswertung des Inhaltes von tei:speaker ein Leerzeichen voran steht.

Man findet außerdem bei gemeinsam sprechenden Personengruppen unterschiedliche Nennungen. In einem Drama Friedrich Kaisers ist eine solche Aggregation mit “HELFER UND ROBERT” benannt, später folgt aber "ROBERT UND HELFER”. Eine bestimmte vom Autor intendierte Hierarchie soll das Datenmodell nicht abdecken und somit gilt es die verschiedenen Zeichenketten als eine Entität zu betrachten. Zudem soll die Tiefenauszeichnung dieser Elemente weiter gehen und jeder Gruppe die einzelnen, sofern bestimmbaren, Akteure zugewiesen werden.

Diesen Beobachtungen folgt die Spielstruktur.

In einem ersten Level gilt es die unterschiedlich benannten aber in der fiktiven Welt gleichen Sprecher zu identifizieren. Dazu werden alle unterschiedlichen Zeichenketten innerhalb der tei:speaker-Elemente eines Dramas zunächst in der Reihenfolge ihres ersten Auftretens gelistet. Mutmaßlich gleiche Namen sind nacheinander auswähl- und abspeicherbar. Ist dies für ein Drama vollständig geschehen, kann dieses Drama als “gelöst” markiert werden.

Weiterhin gilt es Aggregationen ausfindig zu machen (Level 2).

Diese Aggregationen sollen schließlich aufgelöst werden (Level 3). Dazu sind nicht nur die an einer Gruppe beteiligten Akteure zu nennen, sondern auch deren Vollständigkeit zu deklarieren. Es kann zum Beispiel das Volk sprechen und weiterhin einzelne Personen aus dem Volk auftreten. Diese sind Teil des Volkes, die Gruppe selbst ist aber eine weitaus größere und daher unvollständig durch die einzelnen Akteure belegt. Sprechen zwei auch näher bestimmte Einzelpersonen gemeinsam, so kann diese Gruppe vollständig aufgelöst werden.

Die Geschlechter der Akteure sind in Level 4 zu bestimmen. Dabei ist zu wählen aus male, female, both, und unknown. Die letzte Gruppe umfasst dann schließlich auch metaphysische Konstrukte, die personifiziert auftreten.

In Level 5 sollen diese dann genauer spezifiziert werden. Dabei stehen die Kategorien Tier, metaphysisches Wesen (z. B. Gottheit, Hexen und Magier) und Eigenschaft / Gefühl / Moral zur Auswahl.

Schließlich lässt sich noch der soziale Status bestimmen, sofern Berufsbezeichnungen, Adelstitel oder andere Indikatoren ausfindig zu machen sind.

Zwischen den einzelnen Levels gilt es die Eingaben anderer Spieler zu verifizieren oder auch zu falsifizieren. Diese Eingabe wirkt sich auf die eigenen Punkte immer positiv aus, die jeweils anonym bleibende begutachtete Spielerin wird bei Fehleingaben aber Punktabzüge bekommen. Da die Dramen immer zufällig gewählt werden und auch mehrfach erfasst werden, stehen damit verschiedene Qualitätskontrollen zur Auswahl, die auch kontinuierliche nicht sinnvolle Eingaben erkennen lassen. Die betreffenden Spielerinnen können weiterspielen, finden aber nur noch eine persönliche Highscoreliste vor, während sie aus den Highscorelisten anderer getilgt werden und ihre Eingaben auch nicht in den weiteren Forschungsprozess Einzug halten. Zudem stehen die Daten für 465 Dramen im LINA Zwischenformat (vgl. Trilcke et al. 2015) zur Verfügung, die mit den in Level 1 erfassten Eingaben übereinstimmen sollten. Zudem können alle hier getätigten Erhebungen ebenfalls in das Zwischenformat einfließen, womit sie dann für die Netzwerkanalysen des Projektes zur Verfüggun stünden. Bei Bedarf ließen sich die Ergebnisse sogar direkt in die Quelldokumente übernehmen.

Kritisch betrachtet stammen aus der Welt der Computerspiele die Levelstruktur und einzelne Elemente, wie Avatar und Highscoreliste. Tatsächlich ist das Angebot eines, das Social Editing auf einfachste Fragestellungen hin anwendet und jeder Spielerin die Möglichkeit bietet, aktiv an der Tiefenerschließung von literarischen Texten mitzuwirken. Außerdem gibt es einen didaktischen Aspekt, da komplexe Probleme im Hinblick auf Korpuserstellung implizit aufgezeigt werden.

Bibliographie

Prestopnik, Nathan R. (2011): *Citizen Science Case Study*. Galaxy Zoo / Zooniverse http://citsci.syr.edu/system/files/galaxyzoo.pdf [letzter Zugriff 15. Oktober 2015].

TEI Consortium (2015): *TEI P5*. Guidelines for Electronic Text Encoding and Interchange. Version 2.9.0. Updated on 9th October 2015. http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 15. Oktober 2015].

TextGrid Konsortium (2015): *Die Digitale Bibliothek bei TextGrid* https://textgrid.de/digitale-bibliothek [letzter Zugriff 15. Oktober 2015].

Trilcke, Peer / Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario (2015): *Network Analysis of Dramatic Texts* https://dlina.github.io/about/ [letzter Zugriff 15. Oktober 2015].

DH-Projekte Österreichischer Literaturarchive: Ein Problembericht

Hannesschläger, Vanessa
vanessa.hannesschlaeger@oeaw.ac.at
ACDH-OEAW Austrian Center for Digital Humanities, Österreich

Einleitung

In diesem Beitrag werden die Probleme skizziert, die sich aus Praktiken Österreichischer Archive bei der Umsetzung von online-Projekten ergeben. Die Beschränkung auf Österreich ergibt sich, um die Beispiel-Palette überschaubar zu halten; die Problembereiche und Lösungsvorschläge lassen sich allerdings allgemein anwenden. State-of-the-art Projekte von Bibliotheken und Archiven im deutschsprachigen Raum, die den neusten Stand der Forschung umsetzen, werden in einem ersten Schritt beschrieben. Der zweite Abschnitt skizziert Gründe dafür und Konsequenzen daraus, dass diese Standards häufig nicht herangezogen werden. Schließlich werden Lösungsvorschläge präsentiert und eine Agenda vorgeschlagen, die die Situation nachhaltig verbessern könnte. Diese wird im Rahmen des Vortrags auf der DHd 2016 im Zentrum stehen. Dort werden auch die hier allgemein beschriebenen Schwierigkeiten anhand mehrerer Beispielprojekte illustriert.

Forschungsstand / Vorbildprojekte

Vor allem im Bereich der digitalen Edition haben sich im deutschsprachigen Raum auf breiter Ebene Standards und “best practices” entwickelt, die von einer etablierten Community umgesetzt werden. Umfangreiche Bibliotheken publizierter / rechtefreier Werke bieten etwa das *Deutsche Textarchiv* oder der Forschungsverbund *TextGrid* . Diese Plattformen greifen zur Annotation der zur Verfügung gestellten Texte auf die Auszeichnungssprache XML und das Datenformat TEI zurück, die sich in den DH als Standards etabliert haben, und machen auch ihre entsprechenden Tools verfügbar. Die Texte selbst sind creative-commons-lizenziert und können von der Website des Deutschen Textarchivs downgeloadet werden, ebenso wie aus dem *TextGrid Repository*, das darüberhinaus als Langzeitarchiv fungiert.

Im Bereich der Beforschung von Archivbeständen überwiegen im digitalen Raum ebenfalls Editionsprojekte, die die erwähnten Standards zur Anwendung bringen. Beispiele hierfür sind das *Heinrich-Heine-Portal* , das u. a. vom Deutschen Literaturarchiv Marbach unterstützt und vom *Trier Centre for Digital Humanities* umgesetzt wird, oder die *Digitale Edition der Korrespondenz August Wilhelm Schlegels* , an der letztgenanntes Zentrum ebenfalls beteiligt ist. Die Herzog August Bibliothek Wolfenbüttel erarbeitet für ihre in der Reihe *Editiones Electronicae Guelferbytanae* publizierten digitalen Editionen ebenfalls neue Editionstechniken auf Basis der TEI und stellt dazu Dokumentation zur Verfügung. Ein positives Beispiel mit Österreichischer Beteiligung (eine Kooperation zwischen dem Literaturarchiv der ÖNB und dem Institut für Germanistik der Universität Hamburg) ist die Hybridedition des Briefwechsels August Sauer – Bernhard Seuffert (ÖNB 2015), die die Metadaten der online edierten Briefe entsprechend TEI Standards codiert.

Wegweisend speziell für die Arbeit mit literarischen Nachlässen ist auch das Virtual Research Environment (VRE) *SALSAH* (System for Annotation and Linkage of Sources in Arts and Humanities) des *Digital Humanities Lab* der Universität Basel, das ähnlich dem System von Susan Schreibmans *versioning machine* funktioniert.

Im Bereich der Archivierung und Bereitstellung von elektronischen Publikationen, Multimedia-Objekten und anderen digitalen Daten wird in Österreich das Projekt *e-Infrastructures Austria* umgesetzt, das u. a. mit Horizon 2020 verbunden ist und wichtige Impulse für Forschungswebsitearchivierung bringen könnte.

Problemanalyse

Die in Österreichischen Literaturarchiven aufbewahrten Bestände werden im Rahmen von wissenschaftlichen Projekten mit direkt an der Institution angesiedelten Mitarbeitenden erforscht und publiziert. Aufgrund der Vergabepolitik des FWF Forschungsfonds, der in den allermeisten Fällen Geldgeber dieser Unternehmen ist, haben die betreffenden Projekte mittlerweile häufig eine digitale Komponente. Projektleitende und Mitarbeitende sind zumeist literaturwissenschaftlich ausgebildet. Sie konzipieren und entwerfen, wie die digitale Repräsentation ihrer Arbeit strukturiert wird und erarbeiten das wissenschaftliche Konzept, das Inhalt und Funktionalität zugrundeliegt. Für die technische Umsetzung werden meist erst nach Abschluss der konzeptionellen Arbeit externe Auftragnehmende engagiert, oft privatwirtschaftliche IT-Unternehmen, die von den Möglichkeiten, die im Bereich der DH bereits verfügbar wären, nur eingeschränkte Kenntnis haben.

Aus dieser Situation ergeben sich Probleme in mehreren Bereichen:

Langzeitarchivierung von Scans

Die im Rahmen von Projekten erstellten Scans sollten in einer digitalen Langzeitarchivierung der projekttragenden Institution abgelegt werden, was fallweise versäumt wird. Gründe:

- fehlender Speicherplatz
- fehlende Arbeitszeit (sowohl auf Projekt- als auch auf Institutionsseite)
- pragmatische Lösungen: Langzeitarchivierungstaugliche Scans bedeuten einen aufwändigeren Arbeitsprozess
- die Materialien werden häufig nicht in größere Projekte eingespeist (z. B. Europeana)

Datenmodellierung

Die Projektzuständigen haben aufgrund ihrer Ausbildung meist einen editorischen oder von archivarischen Ordnungsprinzipien geprägten Zugang zur Modellierung und Strukturierung der Projektdaten. Gründe und Konsequenzen:

- die gewählten Datenmodelle sind selbst innerhalb einzelner Geisteswissenschaften nicht homogen
- wie konkret die gewählten Daten zu notieren und annotieren sind, wird in jedem Projekt individuell entschieden
- selbst bei Orientierung an vorhandenen Standards ergeben sich unterschiedliche Auslegungen
- keine ausreichenden Kenntnisse von TEI und anderen Standards
- “Selbstverständlichkeiten” werden im Datenmodell oft vergessen (z. B. die Angabe der Verfassenden, wenn ein Projekt sich mit dem Werk einer Einzelperson beschäftigt.)
- keine transdisziplinäre (Wieder-)Verwendung der Daten

Technische Umsetzung, Vernetzung, Visualisierung

Forschungsprojekte werden häufig in Zusammenarbeit mit Firmen umgesetzt, die nicht (primär) mit wissenschaftlicher Klientel arbeiten, deren Wünsche und Methoden daher nicht im Detail verstehen und nicht mit bereits existierenden DH-Tools und Ressourcen vertraut sind. Konsequenzen:

- keine semantische Annotation, Vernetzung mit verfügbaren Datensätzen bzw. LOD
- keine projektinterne Vernetzung der Daten
- keine strukturierte Visualisierung des Datensatzes

Langzeitarchivierung / Verfügbarmachung von Daten

An hostenden Institutionen werden kaum personelle Ressourcen zur Wartung abgeschlossener online-Projekte einkalkuliert. Projekt-Websites sterben daher oft nach wenigen Jahren, mit ihnen die Daten. Auch in Projektfinanzierungsplänen wird dieser Aspekt bislang nicht berücksichtigt. Konsequenzen:

- Daten werden nicht als LOD zur Verfügung gestellt
- Trägerinstitutionen archivieren die Daten nicht sachgemäß

Lösungsansätze

Der skizzierten Situation muss auf allen Ebenen begegnet werden:

Institutionen

Seitens der hostenden Institutionen muss stärker daran gearbeitet werden, für langfristige Datensicherung Möglichkeiten zu entwickeln und anzubieten oder Kooperationen mit Langzeitdatenarchiven einzugehen. Dafür müssen sowohl substanzielle finanzielle als auch personelle Ressourcen explizit dieser Aufgabe zugeordnet werden. Die Postionen, die die *AG Datenzentren im Verband DHd* 2015 formuliert hat, sind dafür wertvolle Impulse und sollten stärker an Institutionen herangetragen werden. Da die betreffenden Institutionen meist Bibliotheken bzw. an solche angeschlossen sind, scheint es auch gewinnbringend, die institutionsinterne Kommunikation zu intensivieren: "digitale Biliothek"-Abteilungen haben für viele der skizzierten Probleme bereits gute Lösungsansätze entwickelt, wie das umfangreiche Vortragsprogramm zu diesem Thema am Österreichischen Bibliothekartag 2015 gezeigt hat. Auch die ÖNB hat mittlerweile eine wichtige Initiative in Angriff genommen: Eine interne Arbeitsgruppe Digital Humanities soll dort Lösungen für die digitale Arbeit entwickeln. Diese Herangehensweise kann für Bibliotheken allgemein als Vorbild dienen, sollte aber auch mit einer Initiative verbunden sein, die im Idealfall kollaborativ erarbeiteten (technischen) Lösungen mit anderen Bibliotheken und Digital Humanists zu teilen.

Fördergebende

Bei der Bewilligung von Projektanträgen sollten Fördergebende die skizzierten Probleme ernsthaft berücksichtigen und Projekte, die keine ausreichenden Ressourcen für die Arbeit an der digitalen Repräsentation der Projektergebnisse vorsehen, ablehnen - anstatt die Praxis, utopische Ziele in Projektanträge einzubauen, zu unterstützen. (Inter)Nationale DH-Plattformen sollten es sich zur Aufgabe machen, ein entsprechendes Empfehlungspapier zur Verfügung zu stellen. Bedenkenswert ist auch die Forderung nach einem "offenen Lebenszyklus" von Forschungsprojekten, die etwa von der Plattform *digital humanities austria* gestellt wird.

Forschende

Das größte Potential zur nachhaltigen Verbesserung der Situation liegt im Bereich der Projektangestellten. Geisteswissenschaftlich Forschende erfahren im Rahmen des Studiums unzureichende Ausbildung zur Arbeit im digitalen Raum und haben in der Folge entsprechende Hemmungen, mit digital humanists in Austausch zu treten. Deshalb werden DH von nicht primär im digitalen Raum arbeitenden Forschenden noch immer als eigene Disziplin wahrgenommen anstatt als Teil und Methode des geisteswissenschaftlichen Forschens an sich. Hier muss Bewusstsein geschaffen und Skepsis abgebaut werden, indem die Lehrpläne grundlegend überarbeitet und gegenwärtigen Standards angepasst werden. Dadurch würden viele der umrissenen Probleme gar nicht erst entstehen. Neben den Forschenden der Zukunft, die man so erreichen kann, müssen kurz- bis mittelfristig auch die Forschenden der Gegenwart stärker animiert werden, sich mit seriösen Methodiken und Frameworks für Forschungsprojekte im digitalen Raum auseinanderzusetzen, indem sie dort, wo sie mit ihrem Wissen stehen, abgeholt werden. Dafür kann etwa die Vorgehensweise des *Austrian Centre for Digital Humanities* der Österreichischen Akademie der Wissenschaften als vorbildlich genannt werden: Hier steht den Forschenden ein digitaler Helpdesk zur Verfügung, der über Möglichkeiten und potentielle Partner für Projekte informiert. Auch das Outreach-Programm, in dessen Rahmen Vorträge, Workshops und eine jährliche DH-Konferenz veranstaltet werden, bietet die Möglichkeit, sich über digitale Methoden zu informieren und mit digital humanists in Kontakt zu treten. Ebenso zu begrüßen sind die Outreach Programme des *Zentrum für Informationsmodellierung – ACDH* der Universität Graz, das vor allem im Bereich der Lehre ein Österreichweites Vorbild sein sollte, und von *e-Infrastructures Austria.*

Die Arbeit in den Bereichen Helpdesk und Outreach zeigt, dass für eine zeitnahe Verbesserung der Situation Adaptionen in der Ausbildung der Forschenden der Zukunft alleine nicht ausreichen; um die Forschenden der Gegenwart zu erreichen, die nicht in digitaler Methodik ausgebildet wurden und sich (noch) nicht damit auseinandergesetzt haben, braucht es "Übersetzende", die die Kommunikation zwischen rein geisteswissenschaftlich und rein digital Denkenden erleichtern und Brücken bauen. Die Wichtigkeit solcher Bindeglieder, die "beide Sprachen sprechen", kann nicht hoch genug eingeschätzt werden, da sie allen Beteiligten Frustration, Zeit, überflüssige Arbeit und letztlich auch Geld ersparen können. Zentren, Institute und Verbände, die in den Digital Humanities arbeiten, sollten ihre Aufmerksamkeit vermehrt auf diesen neuen Arbeitsbereich der digital Übersetzenden richten und ihre Aktivitäten gezielt in diese Richtung lenken.

Conclusio

Die unzureichende Vernetzung geisteswissenschaftlich Forschender mit der DH Community führt zu technischen Unzulänglichkeiten in abseits davon ambitionierten digitalen Projekten, die ihre Nachhaltigkeit gefährden. Gegenseitige Annäherung über Outreach-Programme und die Adaption der Lehrpläne geisteswissenschaftlicher Studienrichtungen, vor allem aber verbesserte interne und externe Kommunikation sind notwendig, um zu nachhaltiger Verbesserung der Situation zu gelangen.

Bibliographie

AG Datenzentren der DHd (2015): "Was sind und was sollen Datenzentren in den Geisteswissenschaften?", in: *Panel der AG Datenzentren im Verband DHd*. Verbandstagung der DHd in Graz, 26.02.2015 https://www.conftool.pro/dhd2015/index.php?page=browseSessions&form_session=37 [letzter Zugriff 09. Februar 2016].

Berlin-Brandenburgische Akademie der Wissenschaften (2007-2015): *Deutsches Textarchiv* www.deutschestextarchiv.de .

Digital Humanities Lab der Universität Basel (2009-2016): *SALSAH*. System for Annotation and Linkage of Sources in Arts and Humanities http://salsah.org/ [letzter Zugriff 09. Februar 2016].

Europeana Foundation (o. J.): *Europeana Collections*. http://www.europeana.eu/portal/ [letzter Zugriff 09. Februar 2016].

Herzog August Bibliothek Wolfenbüttel (o. J.): *Editiones Electronicae Guelferbytanae* http://www.hab.de/de/home/bibliothek/digitale-bibliothek-wdb/digitale-editionen.html [letzter Zugriff 09. Februar 2016].

Karl-Franzens-Universität Graz (2016): *Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities* https://informationsmodellierung.uni-graz.at/ [letzter Zugriff 09 Februar 2016].

Österreichische Akademie der Wissenschaften (2015): *Austrian Centre for Digital Humanities* http://www.oeaw.ac.at/acdh/ [letzter Zugriff 09. Februar 2016].

Österreichische Akademie der Wissenschaften / Universität Wien / Universität Graz (o. J.): *digital humanities austria* http://digital-humanities.at/ [letzter Zugriff 09. Februar 2016].

Österreichische Nationalbibliothek (2011-2015): *Handkeonline*. http://handkeonline.onb.ac.at/ [letzter Zugriff 09. Februar 2016].

ÖNB = Österreichische Nationalbibliothek (2015): *Briefwechsel Sauer - Seuffert*. http://sauer-seuffert.onb.ac.at/ [letzter Zugriff 09. Februar 2016].

Susan Schreibman (2010): *Versioning Machine V4.0* http://v-machine.org/.

TextGrid Konsortium (2006–2015): *TextGrid*. Digitale Bibliothek. Göttingen https://textgrid.de/digitale-bibliothek .

Trier Centre for Digital Humanities (2014): *Digitale Edition der Korrespondenz August Wilhelm Schlegels (beta-Version)* http://august-wilhelm-schlegel.de/briefedigital/ [letzter Zugriff 09. Februar 2016].

Trier Centre for Digital Humanities / Heinrich-Heine-Institut Düsseldorf (2004-2009): *Heinrich-Heine-Portal* http://www.hhp.uni-trier.de/Projekte/HHP/ [letzter Zugriff 09. Februar 2016].

Universität Wien (2014-1016): *e-Infrastructures Austria* http://e-infrastructures.at/ [letzter Zugriff 09. Februar 2016].

Digitale Workflows in Langzeitprojekten am Beispiel einer Infrastruktur zur Dokumentation indigener nordeurpoäischer Sprachen (INEL)

Hedeland, Hanna
hanna.hedeland@uni-hamburg.de
Universität Hamburg, Deutschland

Lehmberg, Timm
timm.lehmberg@uni-hamburg.de
Universität Hamburg, Deutschland

Wagner-Nagy, Beata
beata.wagner-nagy@uni-hamburg.de
Universität Hamburg, Deutschland

Zusammenfassung

Gegenstand des Beitrages sind die Arbeiten zu digitalen Workflows und infrastruktureller Einbindung im Rahmen des Langzeitprojektes *INEL* (Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages), das am Institut für Finnougristik und Uralistik (IFUU) und dem Hamburger Zentrum für Sprachkorpora (HZSK) der Universität Hamburg verortet ist. Ziel des Projektes ist es, über den Zeitraum von 18 Jahren die dringend erforderliche Erschließung der sprachlichen Ressourcen des genealogisch diversen nordeurasischen Sprachraums (s. Abbildung 1) zu leisten. Durch den Einsatz von State-of-the-Art-Methoden und -Werkzeugen der linguistischen Datenaufbereitung, die bisher nur für gut erforschten Sprachen und Varietäten zum Einsatz kamen, wird eine Lücke in diesen für die empirische Sprachwissenschaft bisher schlecht zugänglichen Arealen der Welt nachhaltig geschlossen.

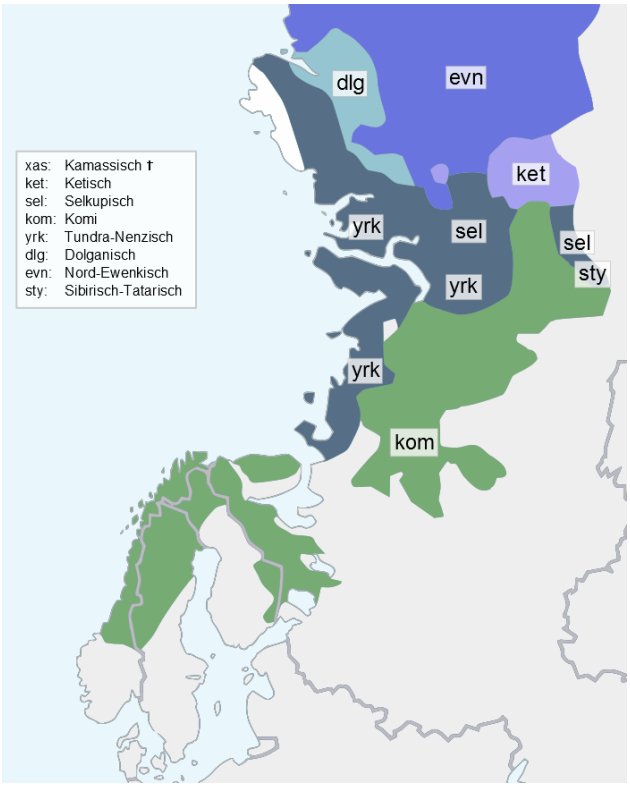


Abb. 1: Der geographische Skopus des Projekts.

Dieses ehrgeizige Ziel stellt hohe Anforderungen an die Organisation der Projektworkflows und erfordert zudem die Schaffung einer eigenen nachhaltigen und international vernetzten digitalen Forschungsinfrastruktur. Das Projekt leistet somit jenseits seiner linguistischen Ausrichtung einen wichtigen Beitrag für die Digital Humanities.

Anforderungslage

Aufgrund des drohenden Verfalls der zum großen Teil auf obsoleten analogen Originalträgern (Wachswalzen, Schellackplatten, Mikrofiche etc.) vorhandenen Audio-Aufnahmen, Niederschriften und Beschreibungen, schließt sich in absehbarer Zeit das Fenster für einen Erhalt dieser Daten. Gleichzeitig gehen die Sprecherzahlen vieler Sprachen und Varietäten stetig zurück. Indem existierende Materialien zu digitalen Korpora aufbereitet und der bisherige Gesamtbestand um neue Ressourcen ergänzt wird, kann dieses Erbe als wertvolle empirische Basis für vielfältige Forschungsvorhaben erhalten werden.

Vielmehr als nur ein digitales Archiv entsteht im Rahmen von *INEL* jedoch eine umfassende virtuelle Forschungsumgebung, die durch die Integration in supranationale Forschungsinfrastrukturen der wissenschaftlichen Öffentlichkeit dauerhaft zugänglich gemacht wird. Ein primäres Ziel des Projektes besteht zunächst darin, existierende Beschreibungen einzelner nordeurasischer Sprachen und Varietäten, die aufgrund der bisher begrenzten Auswahl von verfügbaren Sprechern und Genres eher partikuläre Idiolekte dokumentieren, zusammenzutragen und mit ergänzenden Korpora als umfangreiche digitale Ressource zugänglich zu machen. Durch die so geschaffene, der Vielfalt der Sprache angemessenen, Datenbasis werden für zukünftige Generationen von Forschenden erstmalig varietätenübergreifende Analysen möglich, etwa die Erforschung kontaktinduzierter Sprachveränderungen, Anwendungen aus dem Bereich der Dialektometrie oder sprachsoziologische Untersuchungen. Die unterschiedlichen Erhebungszeiten der Sprachdaten erlauben zudem erstmalig datengestützte Untersuchungen von diachronem Sprachwandel sowie Grammatikalisierungsprozessen. Ebenso bedeutend sind die Art des Zugangs zu den Sprachdaten und die damit verbundenen Analysemöglichkeiten. Die Sprachdaten können in der entstehenden Forschungsumgebung kollaborativ und dezentral um beliebige weitere Beschreibungsebenen angereichert werden, die dann für verschiedene Auswertungsszenarien zur Verfügung stehen. Auf diese Weise wird die virtuelle Forschungsumgebung modular aufgebaut und in vielen Fällen so generisch sein, dass auch die Resultate technologischer und methodologischer Entwicklungen der akademischen Öffentlichkeit als Best Practices und als konkrete Grundlage für vergleichbare Vorhaben zur Verfügung stehen werden.

Modularisierung und Workflows

Die oben beschriebene Ausgangslage determiniert zwei Dimensionen der Entwicklung der entstehenden Ressourcen, denen durch entsprechende Modularisierung von Workflows begegnet werden muss.

1. Entwicklung hinsichtlich der arealen Abdeckung durch Erschließung der Einzelsprachen.

2. Entwicklung hinsichtlich der Komplexität der Daten infolge von hinzuzufügenden Glossierungen und Mehrebenenannotationen, die sowohl innerhalb des Erfassungsprozesses jeder Einzelsprache als auch über den gesamte Laufzeit erfolgen.

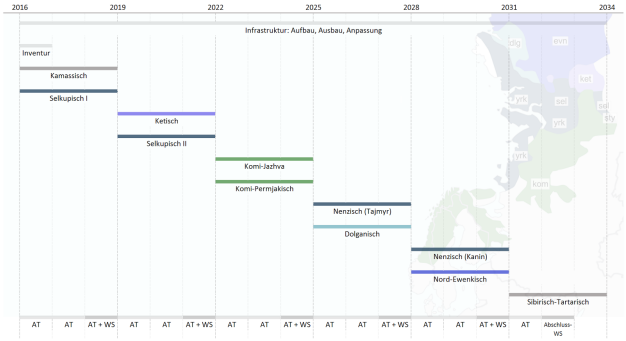


Abb. 2: Teilprojekte.

Das Projekt gliedert sich dem entsprechend in insgesamt zwölf Teilprojekte (s. Abbildung 2), von denen elf jeweils die Erschließung einer der zu erfassenden Sprachvarietäten zum Gegenstand haben. Das zwölfte technisch-infrastrukturelle Teilprojekt läuft im Gegensatz zu den elf jeweils auf drei Jahre angelegten Erschließungsprojekten durchgängig über die gesamte Projektlaufzeit und schafft somit die notwendige Kontinuität für die Entwicklung, Anpassung und Vermittlung der Funktionalitäten der technischen Infrastruktur. Das Arbeitsprogramm wird für die jeweiligen Teilprojekte aber auch teilprojektübergreifend in Form von methodischen Arbeitspaketen umgesetzt:

Arbeitspaket 1: Korpusaufbau

In diesem Arbeitspaket werden existierende Ressourcen erschlossen, kuratiert und aufbereitet sowie ggf. um neu zu akquirierende und zu erhebende Daten ergänzt. Die dabei erforderlichen Verarbeitungsschritte sind, wie in Abbildung 3 dargestellt, modularisiert. Die Module Digitalisierung, linguistische Modellierung, Annotation / Glossierung und Finalisierung/Integration entsprechen jeweils Aufbereitungsschritten, die abhängig vom Ausgangszustand der einzelnen Ressource für die Integration in den Gesamtbestand erforderlich sind.

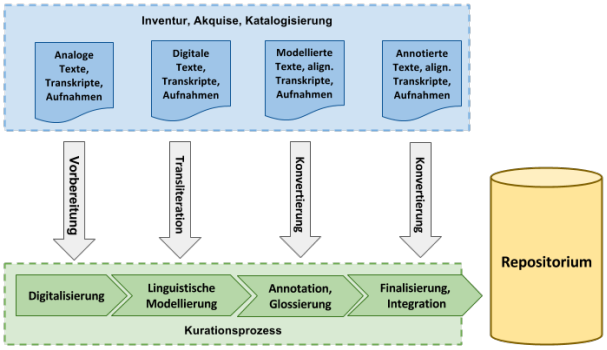


Abb. 3: Modularisierung und Anpassung der Verarbeitungsschritte.

In der Praxis handelt es sich jedoch keineswegs um einen linearen Prozess, den einzelne Texte einmalig durchlaufen und an dessen Ende die Speicherung und Zugänglichkeit einer in sich geschlossenen Ressource in einem digitalen Repositorium steht. Vielmehr müssen die zu planenden Workflows der Aufbereitung und Speicherung den tatsächlichen Gegebenheiten der Datenaufbereitung in Projekten dieser Art Rechnung tragen (s. Abbildung 4):

- Es ist in vielen Fällen wünschenswert, Versionen von Ressourcen bereits in einem frühen Stadium der Aufbereitung (beispielsweise noch vor Abschluss der Annotation und Glossierung) der wissenschaftlichen Öffentlichkeit zugänglich zu machen.
- Bei dem Arbeitsschritt der Annotation und Glossierung handelt es sich wiederum um iterative Prozesse, in deren Rahmen mehrere Ebenen der Auszeichnung, möglichwerweise sogar zeitlich überlappend, zu den Primärdaten hinzugefügt werden, was hohe Anforderungen an Koordination und Qualitätskontrolle stellt.
- Insbesondere bei Langzeitvorhaben entstehen oft neue Versionen von Korpora aus bereits bestehenden Ressourcen, indem diese mit zusätzlichen Annotationsebenen ausgezeichnet werden.

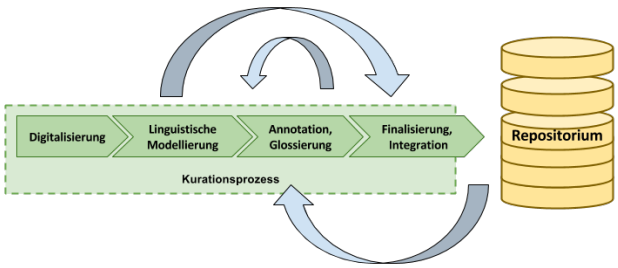


Abb. 4: Nicht-lineare Abfolge der Verarbeitungsschritte.

Diesen Gegebenheiten kann in der Praxis durch ein ausdifferenziertes Versionierungskonzept begegnet werden. Im Rahmen des Vortrages werden einige konkrete Workflows aus dem Projekt vorgestellt und ihre Implementierung unter Verwendung eines Git -basierten Repositoriums ausführlich erläutert.

Arbeitspaket 2: Infrastruktur und Best Practices

Die zu errichtende Infrastruktur basiert in vielerlei Hinsicht auf den Vorarbeiten des HZSK sowie generell auf vorangegangenen Erkenntnissen aus dem Aufbau digitaler Forschungsumgebungen. Große Teile der gewünschten Funktionalitäten wurden bisher in Form von Standalone-Werkzeugen (bspw. EXMARaLDA) oder Webapplikationen (bspw. als Teil der CLARIN-D-Infrastruktur) am HZSK entwickelt und eingesetzt. Einen integralen Bestandteil der Infrastruktur bildet ein Repositorium, mit dessen Hilfe die nachhaltige Datenvorhaltung gewährleistet werden kann. Weitere Komponenten, die unmittelbar daran anknüpfen, bilden zusätzliche relevante Aspekte der gewünschten Funktionalität der Infrastruktur ab, wie etwa die kollaborative Bearbeitung und Aufbereitung von Daten in der Arbeitsumgebung, die Auslieferung von Metadaten an Kataloge und Archive, welche die Auffindbarkeit der Ressourcen für andere Forscher ermöglicht, sowie Schnittstellen für die Exploration und Analyse der vorgehaltenen Ressourcen. Auch die fortlaufende Anbindung an bzw. Vernetzung mit weiteren bestehenden Forschungsinfrastrukturen wird als essentielles Merkmal des Projektes betrachtet.

Arbeitspaket 3: Evaluation und Dissemination

Neben der organisatorischen und der technisch-infrastrukturellen Organisation und Vernetzung, die mit den Arbeitspaketen 1 und 2 abgedeckt ist, erfordert ein Langzeitprojekt wie *INEL* zudem umfassende Arbeiten im Bereich der Dissemination und den Austauschs im supranationalen interdisziplinären Kontexten mit anderen Forschenden. Um neue Arbeitsinstrumente zu entwickeln und zu diskutieren, Arbeitspläne zu koordinieren, aktuelle Forschungsfragen und die mit der Projektarbeit zusammenhängenden praktischen Fragestellungen zu diskutieren sowie zu Zwecken der Fortbildung werden im Rahmen von *INEL* jährliche Workshops abgehalten, an denen ausgewählte Konsultanten sowie Kooperationspartner aus dem Ausland und die Projektmitarbeiter selbst teilnehmen.

Ausblick

Der Beitrag basiert auf den Planungen zu der *INEL*-Langzeitprojekt, dessen 18-jährige Laufzeit im Januar 2016 beginnen wird. Die Vorarbeiten zu dem Projekt lieferten wichtige Erkenntnisse hinsichtlich der Modularisierung von Projektabläufen sowie der Priorisierung von Verarbeitungsschritten der Datenaufbereitung. So wird beispielsweise der technische Fokus nicht auf der Neuentwicklung von weiteren Werkzeugen und Standards der Datenaufbereitung, sondern der Entwicklung von modularisierten Infrastrukturen und Workflows liegen, die auf die Interoperabilität, Interaktion und Integration existierende Komponenten abzielen. Nur so kann ein flexibler Betrieb der *INEL*-Infrastruktur in einer sich permanent wandelnden Ressourcen- und Infrastrukturlandschaft gewährleistet werden.

Von der Entwicklung und Erprobung kontrollierter und modularisierter Workflows der Datenaufbereitung, die beispielsweise eine transparente Dokumentation und Publikation entstehender Versionen von Forschungsdatensammlungen erlauben, sind zudem wichtige Beiträge auf dem Feld des Forschungsdatenmanagements von Projekten in den Digital Humanities zu erhoffen.

Bibliographie

Git-Hub (o.J.): *Git*. Local branching on the cheap http://git-scm.com/ [letzter Zugriff 16. Februar 2016].

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (o.J.): *EXMARaLDA*. Werkzeuge für mündliche Korpora http://www.exmaralda.org/ [letzter Zugriff 16. Februar 2016].

HZSK (o.J.): *Hamburger Zentrum für Sprachkorpora* https://corpora.uni-hamburg.de/drupal/ [letzter Zugriff 16. Februar 2016].

Universität Hamburg (o.J.): *Institut für Finnougristik / Uralistik* https://www.slm.uni-hamburg.de/ifuu [letzter Zugriff 16. Februar 2016].

Sprachwandel im Sanskrit? Eine Corpusstudie zum Einfluss P##inis auf die Lexik des Sanskrit

Hellwig, Oliver
hellwig7@gmx.de
Universität Düsseldorf, Deutschland

Petersen, Wiebke
petersen@phil.uni-duesseldorf.de
Universität Düsseldorf, Deutschland

Problemstellung und Vorarbeiten

Die Sanskrit-Grammatik A##ādhyāyī („[Buch mit] acht Kapitel[n]“) des Linguisten Pā#ini, der wahrscheinlich gegen 350 v. u. Z. in Nordwest-Indien lebte, ist eine der ältesten wissenschaftlichen Grammatiken einer Sprache (Cardona 1976). Die A##ādhyāyī bildet die Grundlage für eine bis heute kontinuierlich fortgeführte Tradition wissenschaftlicher Sanskrit-Linguistik in Indien, und ihr Einfluss auf die indische, aber auch auf die europäische Geistes- und Wissenschaftsgeschichte lässt sich kaum hoch genug ansetzen. Die A##ādhyāyī systematisiert linguistische Phänomene des – seinerzeit wohl noch nicht verschriftlichten – spätvedischen Sanskrit und legt so die Grundlage für das klassische Sanskrit, die *lingua franca* von Wissenschaft, Philosophie und Literatur, die eines der größten vormodernen Textcorpora hervorgebracht hat. Daneben wirken Inhalt und Beschreibungsmethoden der A##ādhyāyī auf beinahe jeden Bereich der altindischen Geistesgeschichte ein. Aus Sicht der westlichen Sprachwissenschaft entwickelt Pā#ini in der A##ādhyāyī formale Methoden wie eine Metasprache oder Ersetzungsregeln, die heute zu den zentralen Elementen moderner linguistischer Theorien gehören (Kiparsky 2009).

Während der Inhalt der A##ādhyāyī und die auf ihr gründende altindische Sanskrit-Linguistik in der Indologie intensiv erforscht wurden, wurde ihr Einfluss auf die Textproduktion und den Sprachwandel in Sanskrit weniger systematisch untersucht. Zum Sprachwandel ist anzumerken, dass das klassische Sanskrit nahezu ausschließlich für Literatur, Wissenschaft und Religion, aber nicht für die alltägliche Kommunikation eingesetzt wurde. Zudem hat die grammatikalische Tradition das Sanskrit als unveränderliche Sprache interpretiert (Deshpande 1993). Daher ist zu erwarten, dass der Sprachwandel im Sanskrit weniger stark ausgeprägt ist als bei gesprochenen Sprachen. Eine kursorische Lektüre von Sanskrit-Texten zeigt, dass das Pā#inäische Regelsystem auf den Ebenen von Phonetik und Morphologie fast uneingeschränkt verwendet wurde. Die A##ādhyāyī hat sich hier von einer ursprünglich deskriptiven zu einer präskriptiven Grammatik gewandelt und setzt einen „Goldstandard“, von dem sich Sprachvarianten wie z. B. das epische Sanskrit (Oberlies 2003) unterscheiden lassen.

Kaum erforscht ist bisher die Frage, inwieweit die A##ādhyāyī nicht nur grammatische Phänomene wie z. B. Wortbildung oder Phonetik, sondern auch die Lexik des Sanskrit beeinflusst hat. Sind Wörter, die in der A##ādhyāyī vorkommen, durch die über Jahrhunderte anhaltende Rezitation der Grammatik vor Sprachwandel geschützt? Hier schließt sich die Frage an, ob Autoren die A##ādhyāyī selbst oder eine vereinfachende Darstellung als Referenzwerk verwendet

haben. Aufgrund der Komplexität der A##ādhyāyī formieren sich v.a. seit dem 11. Jahrhundert u. Z. neue grammatikalische Schulen, die auf der A##ādhyāyī aufbauen, ihr Regelsystem aber vereinfachen und teilweise auch erweitern (Coward / Raja 1990: 19-20). Diese Schulen werden erst im 16. Jahrhundert durch Bhaṭṭojī Dīk#itas Grammatik Siddhāntakaumudī verdrängt, die die Pā#inäische Tradition wiederherstellt, obwohl auch hier die Regeln neu angeordnet werden. Dazu passt Houbens Beobachtung, wonach sich die aktive Leserschaft der A##ādhyāyī auf einen kleinen Zirkel von Spezialisten beschränkte (Houben 2008: 566), obwohl ihr die grammatikalische Tradition, wie Deshpande bemerkt, im Lauf der Jahrhunderte eine wachsende Wertschätzung entgegenbringt (Deshpande 1998).

Die vorliegende Studie schätzt den Einfluss des Vokabulars der A##ādhyāyī auf die spätere Textproduktion mit einem corpusbasierten Ansatz ab. Dabei wird die zeitliche Verteilung Pā#inäischer Beispielnomina (s. Abschnitt) in der Literatur des klassischen Sanskrit untersucht. Die Studie geht von Vorarbeiten in (Hellwig / Petersen 2015) aus, wo gezeigt wurde, dass das Beispielvokabular der A##ādhyāyī im Lauf der Zeit immer seltener auftritt. Dieses Ergebnis widerspricht der These, dass die Verwendung von Worten in der A##ādhyāyī diese vor dem Aussterben schützt. Allerdings ließ sich mit der in (Hellwig / Petersen 2015) verwendeten Datengrundlage nicht ermitteln, ob frühe Sanskrit-Nomina aus anderen Textklassen eine ähnliche diachrone Verteilung zeigen. Die vorliegende Studie erweitert den Referenzrahmen daher um eine zweite Gruppe von Nomina aus der frühen religiösen Literatur (*śruti*, „das Anhören [‘heiliger’ Texte]“), deren Texte (Brāhma#as, Upani#ads) in der Sanskrit-Tradition ebenso hoch geschätzt werden wie die A##ādhyāyī. Die Verwendung von „religiösen Nomina“ wird mit derjenigen von Beispielnomina aus der A##ādhyāyī kontrastiert.

Daten

Als Corpus dient das Digital Corpus of Sanskrit (DCS) mit rund 3.570.000 Tokens, dessen Texte automatisch tokenisiert und morphologisch-lexikalisch analysiert und danach manuell korrigiert wurden (Hellwig 2015). Alle Texte des DCS sind einer von fünf Hauptperioden der Sanskrit-Literatur zugeordnet. Die Struktur dieser Zeitachse stellt angesichts der Tatsache, dass zahlreiche wichtige Sanskrit-Texte anonym und ohne Datierung überliefert sind, gerade in den frühen Zeitschichten nur eine grobe Näherung dar (Hellwig 2010). Daneben sind alle Texte mit einem inhaltlichen Label versehen, das aus einem traditionellen Klassifikationsschema abgeleitet wurde.

Der Einfluss der A##ādhyāyī auf die spätere Literatur wird anhand der Verteilung von 1341 Nominaltypen untersucht, an denen Pā#ini grammatikalische Phänomene exemplifiziert. Diese Beispielnomina werden in Pā#ini’s Metasprache von Nomina unterschieden, die auf ein externes Objekt referieren (sūtra 1.1.68 der A##ādhyāyī, vgl. Séaghdha 2004: 19ff. für eine Übersicht über die Forschung). Während z. B. das Nomen *nāsikā* in *sūtra* 1.1.8 der A##ādhyāyī auf die Nase referiert, [1] beziehen sich die Nomina *deva* und *brahman* in *sūtra* 1.2.38, das das vedische Akzente diskutiert, nicht auf einen Gott und einen Brahmanen, sondern werden in ihrer „lautlichen Erscheinung“ verwendet: “Bei [den Nomina] *deva* und *brahman* tritt der *anudātta* und *svarita* an die Stelle des *svarita*.” (Übersetzung nach Böhtlingk 1886; unsere Ergänzungen in eckigen Klammern). In der Datenbank A##ādhyāyī 2.0 , die den Text der A##ādhyāyī mit zahlreichen grammatischen und semantischen Annotationen unterlegt, sind solche nicht-referierenden Beispielnomina explizit markiert (Petersen / Hellwig 2015). Dadurch lässt sich das Pā#inäische Beispielvokabular problemlos für corpuslinguistische Studien erschließen.

Das Vergleichsvokabular der frühen religiösen Literatur entstammt wie die A##ādhyāyī der frühesten Zeitschicht des DCS, die von der weiteren Auswertung ausgeschlossen wird. Da für die *śruti* keine äquivalenten wortsemantischen Annotationen vorliegen, werden aus dem *śruti*-Untercorpus die 1268 Nominaltypen ausgewählt, die mehr als einmal auftreten. Diese „ *śruti*-Nomina“ enthalten Begriffe wie *loka* („(Menschen- / Götter-)Welt“), *ātman* („Selbst“) und die Namen spätvedischer Gottheiten und spiegeln damit die zentralen Inhalte der *śruti*-Literatur wider.

Da das DCS weder thematisch noch zeitlich ausgewogen ist, beruhen die im folgenden dargestellten Verteilungen von Lexemen auf wiederholten Stichproben (*samplings*) fester Größe. Dazu wird das Corpus zuerst anhand von Zeitperioden und inhaltlichen Labels in Faktorstufen aufgeteilt. Anschließend werden aus jeder Faktorstufe 100 zufällige Stichproben von 500 Nomina gezogen. Für jede Stichprobe wird ausgezählt, wie viele der Referenzwörter vorkommen, und die resultierenden prozentualen Anteile werden für jede Faktorstufe gemittelt. Diese Mittelwerte bilden den Ausgangspunkt für die Auswertung im nächsten Abschnitt.

Auswertung

Abbildung 1 zeigt die gemittelten relativen Häufigkeiten, mit denen Beispielwörter aus der A##ādhyāyī im späteren Sanskrit auftreten, als*heatmap*, in der dunkle Farbtöne eine höhere durchschnittliche Häufigkeit in einer Faktorstufe anzeigen. Obwohl gerade seltene Beispielwörter in einigen Domänen wie der Sanskrit-Lexikographie (Zelle „Lex.“, Zeitstufe 3) gehäuft auftreten, nimmt die Verwendung des Pā#inäischen Vokabulars insgesamt mit der Zeit ab. Dieses Ergebnis deutet darauf hin, dass der direkte

Einfluss der A##ādhyāyī auf das Sanskrit-Vokabular im Lauf der Zeit geringer wird.

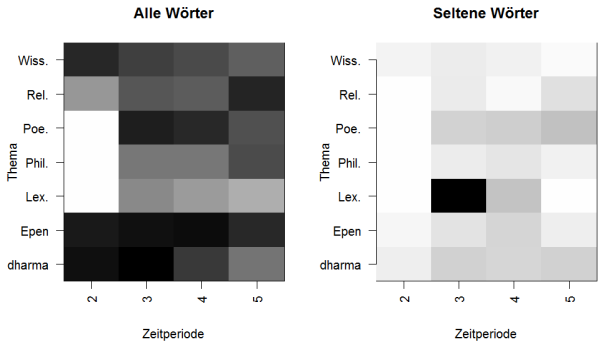


Abb. 1: Verteilung von Beispielnomina aus der A##ādhyāyī

Abbildung 1 klärt allerdings nicht, ob die Verwendungshäufigkeit bei Lexemen aus der A##ādhyāyī weniger stark abnimmt als bei Lexemen aus anderen frühen Texten. In Abbildung 2 wird deshalb die zeitliche Verteilung Pā#inäischer Beispielnomina derjenigen der *śruti*-Nomina in fünf Konfigurationen gegenübergestellt. Der linke Teilbereich zeigt, dass sich die Nomina aus beiden Gruppen grundsätzlich ähnlich über die Sanskrit-Literatur verteilen. Während in der epischen, wissenschaftlichen und „Rechtsliteratur" eine allgemeine Abnahme früher Nomina zu beobachten ist, erleben sie in der religiösen und philosophischen Literatur eine Art *revival*. Die Analyse der daran beteiligten Begriffe zeigt, dass hierfür v.a. Begriffe wie *yoga*, *prā#a* („Atem"), *ātman* („Selbst") oder *jñāna* („Wissen") verantwortlich sind, die z. B. in Yoga-Texten oder bei der damit verbundenen Mikro-Makrokosmos- Spekulation eine zentrale Rolle spielen.

Für die Plots auf der rechten Seite von Abbildung 2 wurde die Vereinigungsmenge des Pā#inäischen Beispielvokabulars (P) und der *śruti*-Nomina (S) in die drei Teilmengen *P*∩ *S* (Schnittmenge), *P*# *S* (alle aus P, die nicht in S sind) und *S*# *P* (alle aus S, die nicht in P sind) aufgeteilt. Der allgemeine Trend dieser Detailverteilungen unterscheidet sich auf den ersten Blick nicht grundsätzlich von der linken Seite des Plots. Bemerkenswert ist allerdings der vergleichsweise hohe Anteil, den die Schnittmenge *P*∩ *S* am später verwendeten frühen Nominalvokabular einnimmt. Dieser Befund deutet aus unserer Sicht weniger auf eine verstärkte Rezeption Pā#ini's hin als auf die Tatsache, dass die A##ādhyāyī populäre Beispielwörter verwendet.

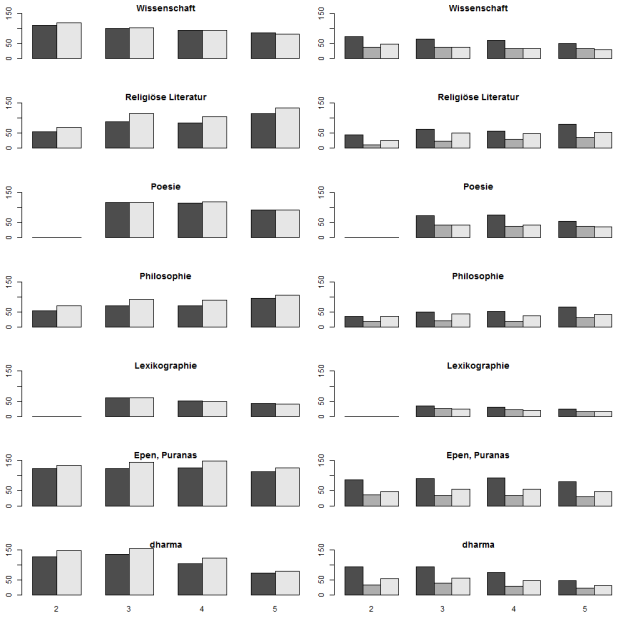


Abb. 2: Links: Verteilung von Nomina, die in der A##ādhyāyī (dunkelgrau) und in der *śruti* (hellgrau) vorkommen. Rechts: Verteilung von Nomina, die in *śruti* und A##ādhyāyī (dunkelgrau), nur in der A##ādhyāyī (mittelgrau) und nur in der *śruti* (hellgrau) vorkommen.

Abschließend werden die Häufigkeiten für alle fünf Konfigurationen pro Zeitperiode über alle Themengebiete gemittelt. Die drei oberen Plots in Abbildung 3 zeigen, dass der Kernbestand des spätvedischen Vokabulars im klassischen Sanskrit zunehmend seltener verwendet wird. Weniger eindeutig ist die Verteilung in den beiden unteren Plots, die eine geringe, aber recht konstante Verwendung der Nomina aus *P*# *S* erkennen lassen.

Diese geringere Abnahme des Pā#inäischen Kernvokabulars scheint die These zu stützen, dass die A##ādhyāyī dauerhaft rezipiert wurde und ihre Wörter durch diese regelmäßige Verwendung weniger stark dem allgemeinen Wandel der Lexik unterliegen. Allerdings muss diese Schlussfolgerung in zwei Punkten deutlich eingeschränkt werden. Erstens ist die Vergleichsgruppe der religiösen Texte nur deshalb bis heute überliefert, weil sie ebenfalls als wichtig angesehen und deshalb regelmäßig rezipiert wurden. Hier fehlt ein Corpus von Texten aus der Zeit Pā#ini's, die danach keine Rolle mehr spielten. Zweitens sollten die Vergleichswörter aus der *śruti* genauer semantisch differenziert werden, da z. B. durch den hohen Anteil von Eigennamen für Gottheiten, die auch in späteren Texten eine wichtige Rolle spielten, die lexikalische Aussterberate in den *śruti*-Texten unterschätzt werden könnte. Dieser zweite Punkt wird in einer Folgestudie untersucht werden.

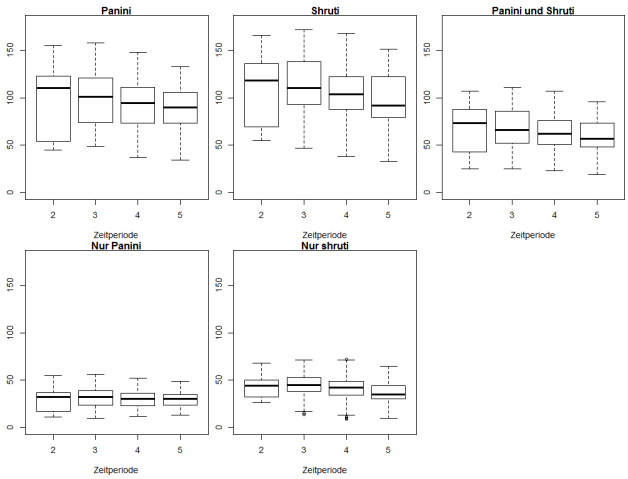


Abb. 3: Über Themengebiete gemittelte Häufigkeiten für die fünf Vergleichskonfigurationen

Notes

1. *mukhanāsikāvacano 'nunāsika#*, „Ein mit Mund und Nase ausgesprochener Laut heisst *anunāsika*(nasal)".

Bibliographie

A##ādhyāyī: *A##ādhyāyī 2.0* http://panini.phil.hhu.de/panini/panini/ [letzter Zugriff 10. Februar 2016].

Böhtlingk, Otto (1886): *Pā#ini's Grammatik.* Leipzig: Verlag von H. Haessel.

Cardona, George (1976): *Pā#ini*. A Survey of Research. The Hague / Paris: Mouton.

Coward, Harold G. / Raja, K. Kunjunni (1990): *The Encyclopedia of Indian Philosophies*. 5: The Philosophy of the Grammarians. Princeton: Princeton University Press.

Deshpande, Madhav M. (1993): "Historical Change and the Theology of Eternal (nitya) Sanskrit", in Deshpande, Madhav M.: *Sanskrit and Prakrit*. Sociolinguistic Issues. Delhi: Motilal Banarsidass Publishers 53-74.

Deshpande, Madhav M. (1998): "Evolution of the Notion of Authority (Prāmā#ya) in the Pā#inian Tradition", in:*Histoire Épistémologie Langage* 20, 1: 5-28.

Hellwig, Oliver (2010): "Etymological Trends in the Sanskrit Vocabulary", in: *Literary and Linguistic Computing* 25, 1: 105-118.

Hellwig, Oliver (2015): "Morphological Disambiguation of Classical Sanskrit", in: Mahlow, Cerstin / Piotrowski, Michael (eds.): *Systems and Frameworks for Computational Morphology*. Fourth International Workshop, SFCM 2015, Stuttgart, Germany, September 17-18, 2015. Proceedings (= Communications in Computer and Information Science 537). Cham: Springer 41-59.

Hellwig, Oliver / Petersen, Wiebke (2015): What's Pā#ini got to do with it? The use of ga#a-headers from the A##ādhyāyī in Sanskrit literature from the perspective of Corpus Linguistics. *Proceedings of the WSC 2015* (forthcoming).

Houben, Jan (2008): "Bhaṭṭoji Dīk#ita's "Small Step" for a Grammarian and "Giant Leap" for Sanskrit Grammar", in: *Journal of Indian Philosophy* 36: 563-574.

Kiparsky, Paul (2009): "On the Architecture of Pā#ini's Grammar", in: Huet, Gérard / Kulkarni, Amba / Scharf, Peter (eds.):*Sanskrit Computational Linguistics*. Berlin / Heidelberg: Springer 33-94.

Oberlies, Thomas (2003): *A Grammar of Epic Sanskrit.* Berlin: De Gruyter.

Petersen, Wiebke / Hellwig, Oliver (im Druck): "Annotating and Analyzing the Astādhyāyī", in: Alonso Almeida, Francisco / Ortega Barrera, Ivalla / Quintana Toledo, Elena / Sánchez Cuervo, Margarita (eds.): *Input a Word*. Analyse the World. Selected Approaches to Corpus Linguistics. Cambridge: Cambridge Scholars Publishing.

Petersen, Wiebke / Soubusta, Simone (2013): "Structure and implementation of a digital edition of the A##ādhyāyī", in: Kulkarni, Malhar (ed.): *Recent Researches in Sanskrit Computational Linguistics*. Delhi: D.K. Printworld 84-103.

Séaghdha, Diarmuid (2004): *Object-Language and Metalanguage in Sanskrit Grammatical Texts*. Diplomarbeit. Cambridge: University of Cambridge.

Aufbau und Annotation des Kafka/Referenzkorpus

Herrmann, J. Berenike
bherrma1@gwdg.de
Universität Göttingen, Deutschland

Lauer, Gerhard
gehard.lauer@phil.uni-goettingen.de
Universität Göttingen, Deutschland

Der vorgeschlagene Beitrag dokumentiert das Ineinandergreifen philologischer und informatischer Fragestellungen und Entscheidungen bei Aufbau und Aufbereitung eines digitalen Korpus für vergleichende quantitative Stilanalysen von Franz Kafkas Prosa.

In den letzten Jahren haben digitale Ressourcen wie TextGrid, das Deutsche Textarchiv [DTA], und Gutenberg-DE reichhaltige digitale Korpora von historischen Texten (literarischer und nichtliterarischer Art) zur Verfügung gestellt. Kafkas Werk selbst ist zudem fast vollständig digitalisiert. Dennoch liegen derzeit weder ein vollständiges Kafka-Kernkorpus noch ein „Kafka-

Referenzkorpus“ vor, das eine sinnvolle quantitative Analyse seines Sprachgebrauchs durch den Vergleich mit ausreichend großen Stichproben anderer Texte zulässt. Unser Projekt möchte diese Lücke füllen und ein Kafka/Referenzkorpus vorstellen, das sowohl philologisch als auch informatisch solide aufbereitetet ist, und eine hypothesengetriebene aber auch explorative quantitative Stilistik ermöglicht.

Bei der Konzeption des Kafka/Referenzkorpus verfolgen wir einen autororientierten Ansatz der digitalen Stilistik. Ausgehend von der Hypothese, dass der Stil eines Autors durch von ihm rezipierte Texte beeinflussbar ist, und dass Stil quantitativ beschreibbar ist (vgl. Herrmann / van Dalen-Oskam / Schöch 2015), gehen wir zunächst vom faktischen textuellen Input Kafkas aus und ergänzen diesen durch Stichproben kanonischer und populärer zeitgenössischer Texte. Der Aufbau des Kafka/Referenzkorpus wird von drei Kriterien geleitet:

(a) Vollständigkeit von Kafkas Schriften in der Originalfassung (=Kafka-Kernkorpus);

(b) Abbildung von Texten, die Kafkas Schreibprozess beeinflusst haben könnten / Abbildung von Texten, die eine näherungsweise Repräsentativität der Epoche der klassischen Moderne herstellen (=Kafka-Referenzkorpus);

(c) eine hohe Akkuratheit bzw. Konsistenz bei informatischer Vorverarbeitung wie Normalisierung, linguistischer Annotation (*Part of Speech*), Metadaten und Textmarkup (XML-TEI) in einem *Stand-Off* Korpus, das einen hohen Grad an Forschungsflexibilität ermöglicht.

Das Kafka-Kernkorpus (je nach Zählart ca. 120 Texte) wurde dabei intern in die Dimensionen Kafka_Publikation (zu Lebzeiten/posthum) und Kafka_Genre (Literarisch, Brief, Tagebuch, Amtliche Schriften) unterteilt. Das Referenzkorpus (ca. 8.000 Texte) wurde hauptsächlich aus TextGrid, DTA, Gutenberg-DE und Gutenberg.org extrahiert, und beinhaltet Metadaten zu Autor (Name, Gender), Publikationsdatum und -Ort, sowie Gattung. Es umfasst literarische Texte der Kategorien „kanonisch“ und „populär“ ebenso wie Gebrauchstexte. Neben Kinder- und Jugendliteratur die Kafka rezipierte sind hier auch Sach- und Fachliteratur von Interesse, nicht zuletzt weil Kafkas Stil durch Elemente der Fachsprache, aber auch ein hochsprachliches „Prager Deutsch“ ohne sozio- oder dialektale Einflüsse geprägt sein soll (vgl. Nekula 2003). Zur Korpuszusammensetzung wurden Aufzeichnungen zu Kafkas Lesegewohnheiten untersucht, wobei Zeugnisse über seine Bibliothek, biographische Berichte, aber auch Dokumente zur zeitgenössischen Rezeption sowie Autor- und Werk-Indices in Literaturgeschichten konsultiert wurden (Blank 2001; Born 1990; Born / Koch 1983; Born / Mühlfeit 1979). Das Ergebnis dieses Forschungsschrittes ist eine Liste von 765 Titeln, die das Metadatum „in Kafkas Bibliothek“ tragen, und einen Schwerpunkt zu Kafkas Lebzeiten setzen, aber eben auch Werke von älteren Autoren wie Goethe und Kleist, sowie Flaubert und Dostojewski (in Übersetzung) beinhalten. Dass die von uns einbezogenen Online-Repositorien hinsichtlich der editionsphilologischen Textqualität variieren ist ein hinzunehmendes Übel, dem wir zum einen pragmatisch (Wahl der bestmöglichen verfügbaren Ausgabe; Ziel, die Fehlermarge unter 2% zu halten), zum anderen unter Hinweis auf die flexible Struktur des Korpus (Austausch durch eine qualitativ hochwertigere Version ist möglich) begegnen. Durch die nahtlose Dokumentation des Korpus wird zudem die nötige Transparenz gewährleistet um auch Nachnutzern flexible Kontrolle der Daten zu ermöglichen.

Die Hauptaufgabe der informatischen Dimension des Projektes besteht neben der Einbettung in einen praktikablen und anschlussfähigen Workflow (eXist Datenbank) und der Homogenisierung und informatischen Aufbereitung der Ausgangsdaten (Tokenisierung, Normalisierung, Lemmatisierung) in einer reliablen linguistischen Annotation auf POS (STTS Tagset). Wortarten gelten als verlässliche Indikatoren für Register und Genrevariation (vgl. z. B. Biber / Conrad 2009), und sind im Vergleich mit anderen Variationsmarkern durch eine relativ akkurate automatische Annotation besonders praktikabel. Obwohl bei der POS-Annotation gute Accuracy für das gegenwärtige Standarddeutsch mithilfe von *Hidden-Markov-Modellen* und *Markov-Modellen* erzielt wird (RF-tagger, Tree-Tagger), wurden diese Tagger auf Zeitungstexten trainiert und erfordern deshalb in unserem Korpus manuelles Fehlermanagement: Ein Ausschnitt des Gesamtkorpus (repräsentativer Querschnitt auf Satzebene, randomisiertes Sampling) wird manuell auf POS getaggt und mit dem Output der Tagger verglichen. Liegt die Übereinstimmung größer oder gleich der Standard-Accuracy (ca. 97%), ist eine umfangreiche Fehleranalyse nicht notwendig. Sollte die Accuracy niedriger sein, wird in der Folge über ein manuell kodiertes Sample von ca. 40.000 Wörtern durch *Machine Learning* ein Algorithmus trainiert, der bessere Werte erreicht. Hierbei ist auch der Gebrauch von Taggern aus dem *Conditional Random Field* (CRF) Framework wie MarMoT vorgesehen, die eine größere Input-Spanne berücksichtigen. Der Workflow beinhaltet einen automatischen Vergleich des Tagger-Outputs durch eine eXist-Datenbank mit Annotationsinterface. Der Output wird in einem Stand-Off Format (TCF) gespeichert, wie es auch das DTA benutzt. Geplant ist zusätzlich eine Qualitätskontrolle des TEI-Markups, der Metadaten und der POS-Annotation durch einen bereits entwickelten Ansatz der *Gamification* (s. https://personae.gcdh.de/index.html). Das Kafka / Referenzkorpus soll im Rahmen der TextGrid Infrastruktur in SADE veröffentlicht und so der Forschungsgemeinschaft zur Verfügung gestellt werden. Gleichzeitig planen wir eine detaillierte Dokumentation der Arbeitsschritte zu veröffentlichen, die ähnlichen Projekten als Leitfaden zur Verfügung zu stehen soll.

Unser Projekt dokumentiert in seinem gegenwärtigen Status Entscheidungen auf verschiedenen konzeptionellen, analytischen und prozeduralen Ebenen. Es zeigt, dass der Aufbau eines digitalen Autor-Korpus, das

den quantitativen Vergleich mit synchronen und diachronen Daten erlauben soll, bei Weitem keine triviale Aufgabe darstellt. So wird zum Beispiel deutlich, wie Forschungsfragen beziehungsweise Hypothesen zur Konstitution von Schreibweisen und Autorschaft die Korpuskompilation steuern – und deshalb auf einer möglichst präzisen Modellierung der zugrundeliegenden textwissenschaftlichen Theorien fußen sollten. Gleichzeitig sind Metadaten (u. a. Autor, Titel, Publikationsdatum, Publikationsort, Genre) und linguistische Parameter (wie POS) gerade die Ansatzpunkte, an denen philologische Fragestellungen in präzise und praktikable Kategorien umgewandelt werden können. Nicht zuletzt deshalb sollten literarische Daten in flexiblen Architekturen gespeichert werden, die zusätzliche Annotationsebenen zulassen – denn hermeneutische Erkenntnisprozesse stellen eine erwachsene Stärke der Geisteswissenschaften dar, die auch im digitalen Zeitalter einen explizit modellierten Platz einnehmen muss.

Bibliographie

Biber, Douglas / Conrad, Susan (2009): *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Blank, Herbert (2001): *In Kafkas Bibliothek*. Werke der Weltliteratur und Geschichte in der Edition, wie sie Kafka besaß oder kannte; kommentiert mit Zitaten aus seinen Briefen und Tagebüchern. Stuttgart: Blank.

Born, Jürgen (1990): *Kafkas Bibliothek*. Ein beschreibendes Verzeichnis; mit einem Index aller in Kafkas Schriften erwähnten Bücher, Zeitschriften und Zeitschriftenbeiträge. Frankfurt am Main: S. Fischer.

Born, Jürgen / Koch, Elke (eds.) (1983): *Franz Kafka: Kritik und Rezeption, 1924-1938*. Frankfurt am Main: S. Fischer.

Born, Jürgen / Mühlfeit, Herbert (eds.) (1979): *Franz Kafka: Kritik und Rezeption zu seinen Lebzeiten, 1912-1924*. Frankfurt am Main: S. Fischer.

Herrmann, J. Berenike / van Dalen-Oskam, Karina / Schöch, Christof (2015): "Revisiting Style, a Key Concept in Literary Studies", in: *Journal of Literary Theory* 9, 1: 25-52.

Nekula, Marek (2003): "Franz Kafkas Deutsch", in: *Linguistik online* 13, 1 https://bop.unibe.ch/linguistik-online/article/view/879/1533 [letzter Zugriff 29. Dezember 2015].

Classification of Literary Subgenres

Hettinger, Lena
lena.hettinger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reger, Isabella
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Hotho, Andreas
hotho@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Introduction

Literary scholars and common readers use labels like educational novel, crime novel or adventure novel to organize the large domain of fiction. In both discourses the use of these categories is well-established even though they are evolving and tend to be inconsistent. The classification of genres is one of the standard tasks in document classification and has been researched intensively (cf. Biber 1989; Santini 2004; Freund et al 2006; Sharoff et al. 2010). Some results seem impressive, for example distinguishing clear-cut genres like poetry from fiction (Underwood 2014), but most texts on literary genre classification emphasize, as the literature on genre classification in general, the variability of genre signals (Allison et al. 2011: 19; Underwood et al. 2013; Underwood 2014). The scores for genre classification over all categories are therefore often not very high. Jockers for example reports an accuracy of 67% (Jockers 2013: 81). Genre classification in general works best with most frequent words, all words or character tetragrams (Freund et al. 2006; Sharoff et al. 2010) and most of the reported experiments for literary genre classification also use all words or only the *n* most frequent word (sometimes including punctuation) as features. In a series of experiments we examine whether it is possible to enhance these results for the classification of subgenres of novels. Our research is motivated by an understanding of novel genres as concepts which are differentiated by style, settings, character constellations and plots. We use most frequent words as an indicator for style and network characteristics as an indicator for character constellations. Setting is partially covered by topic models which also represent information on typical ways of telling a story, narrative *topoi*. We have to omit plot, as we don't have a reliable way to represent plot by any indicators yet.

Setting

In the following we will describe the corpus and the features we use for the task of subgenre classification.

Corpus

Our corpus consists of 628 German novels mainly from the 19th century (roughly 1745 to 1935) obtained from sources like the or the German Projekt Gutenberg . The novels have been manually labeled according to their subgenre after research in literary lexica and handbooks. The corpus contains 221 adventure novels, 88 social novels and 86 educational novels; the rest are novels from different subgenres.

Features

As mentioned in section 1 we use three types of features (stylometric, topic based and network) that are described in more detail in Hettinger et al. (2015). Features are extracted and normalized to a range of [0,1] based on the whole corpus consisting of 628 novels.

Stylometric features

We use word frequencies as well as character tetragrams to represent stylometric features. We tested different amounts of most frequent words and decided to work with the top 3000 (mfw3000). Additionally we use the top 1000 character tetragrams (4gram).

Topic features

We use Latent Dirichlet Allocation (LDA) by Blei et al. (2003) to extract topics from our data. In literary texts topics sometimes represent themes, but more often they represent topoi, often used ways of telling a story or parts of it. For each novel we derive a topic distribution, i.e. we calculate how strongly each topic is associated with each novel. We try different preprocessing approaches and topic numbers and build ten models for each setting to reduce the influence of randomness in LDA models. In every setting we first remove a set of predefined stop words from the novels and then use LDA on our corpus of 628 novels. The different forms of preprocessing we use are:

- no additional preprocessing
- removal of Named Entities (Jannidis et al. 2015)
- word stemming
- word lemmatization
- word lemmatization + removal of Named Entities

Network features

We use the character recognition system described in Jannidis et al. (2015) to identify the characters of each novel. Although the NER tool may be employed with co-reference resolution we do not make use of this option here. We extract proper names to build a network where each node is a character and the number of co-occurrences of two characters in the same paragraph is the weight of the edge between these two. The network of each novel is reduced to the most central characters and the most frequent interactions in order to bring out their basic shape. The network feature set consists of the total number of characters in a novel and six network measures: maximum degree centrality, global efficiency, transitivity, average clustering coefficient, central point dominance and density.

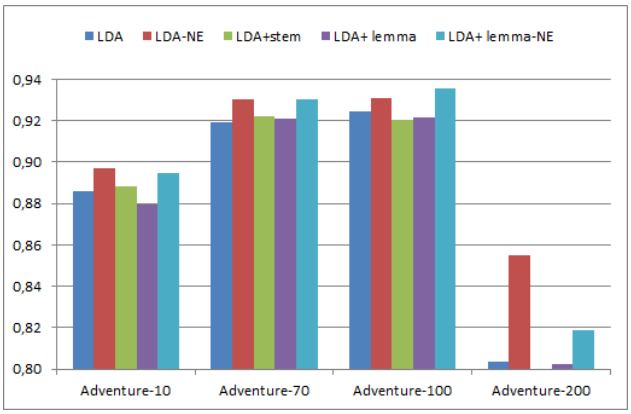
Evaluation

Classification is done by means of a linear Support Vector Machine (SVM) as we have already shown in Hettinger et al. (2015) that it works best in this setting (see also Yu 2008). In each experiment we apply 100 iterations of 10-fold cross validations to account for the small data sets. The depicted results are the average over 1000 classification accuracy values. We want to investigate the following subgenre constellations:

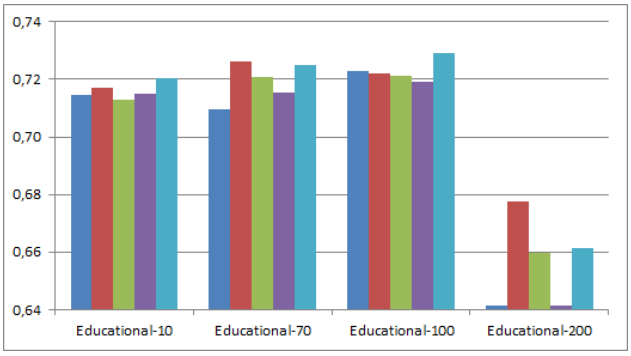
- Adventure versus non-Adventure novels
- Educational versus non-Educational novels
- Social versus non-Social novels
- Adventure versus Educational novels and
- Educational versus Social novels

Depending on the setting the label distribution is often imbalanced. To make results comparable we use undersampling where in each of the 100 iterations a new sample is drawn from the larger class while all instances of the smaller class are used. This accounts for a majority vote (MV) baseline that always yields an accuracy score of 0.50 as both classes have equal size.

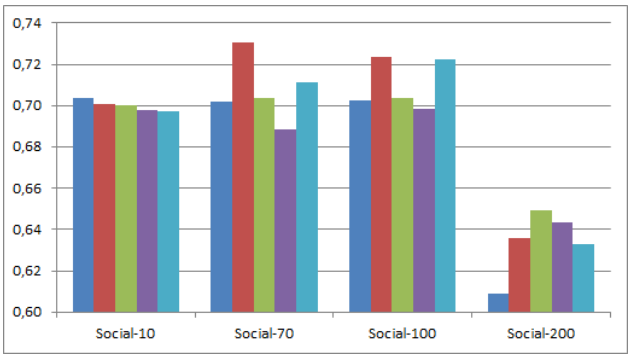
To determine the influence of the LDA topic parameter t and different preprocessing procedures we report accuracy for t =10, 70, 100 and 200 (see figure 1). Differences between different preprocessing categories are minimal, but removal of named entities seems to improve results overall. We observe comparable result for t =10, 70 and 100 and a drop in performance for t = 200. Therefore, we will use LDA with lemmatized words and named entities removed for t = 100 in the following experiments.



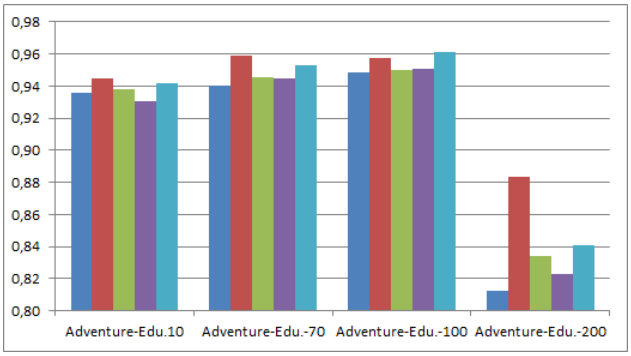
a) adventure/non-adventure



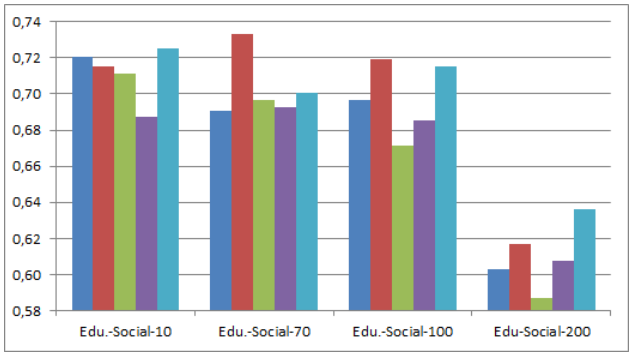
b) educational/non-educational



c) social/non-social



d) adventure/educational



e) educational/social

Fig. 1: Classification results for different subgenre settings in terms of accuracy using LDA with topic size t = 10, 70, 100, 200 and five different preprocessing procedures on 628 German novels

When comparing different feature sets across our subgenre constellations we can see that semantic based features (mfw, 4grams, lda) all perform quite good while network features perform rather poorly (see figure 2). With an accuracy score of more than 90% adventure novels seem to be fairly easy to differentiate from other genres. In contrast, the other genres don't show such a distinct signal using surface features.

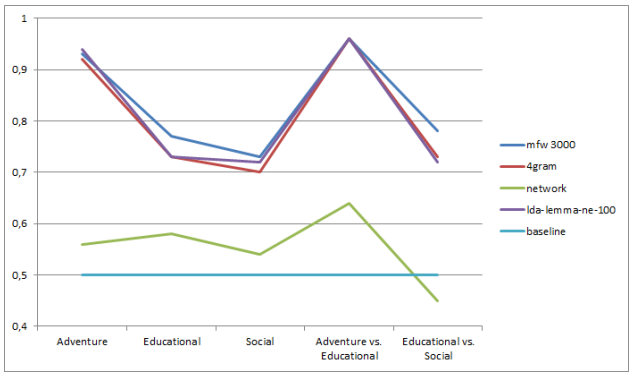


Fig. 2: Accuracy for different scenarios and feature sets including the majority vote baseline.

As the classification performance of adventure/ educational is quite impressive we take a closer look at the discriminating words of these genres (Figure 3). Some of the most typical words of adventure novels include *captain, shore, (on) board, help, danger, immediately*. On the other hand words like *soul, love, world, heart, (at) home, mother, joy, often, father, emotion, human, to love* are characteristic for educational novels.

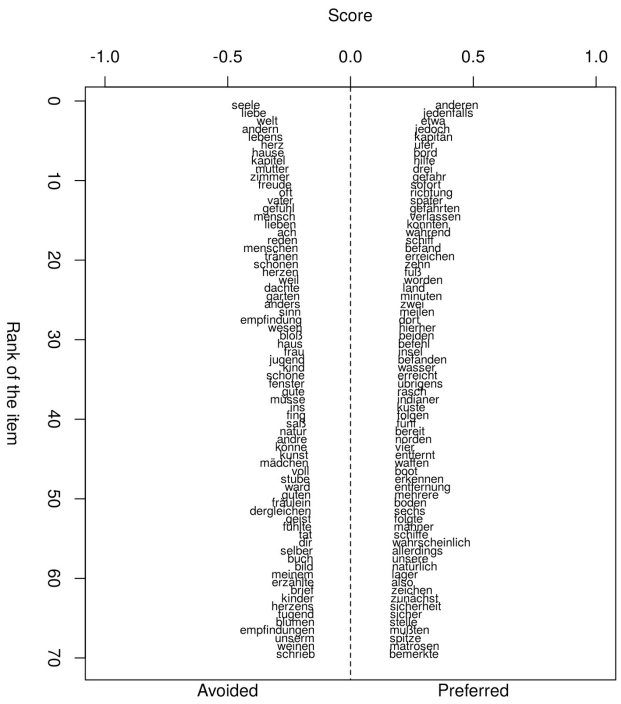


Fig. 3: Discriminating words for adventure (preferred) and educational (avoided) novels (Craig's Zeta)

To test whether authorship of novels influences our results we removed the author signal by allowing only one document per author. In this way, we construct a new dataset called 'uni'. The sampling is done once so that the same novels are used in each setting. As shown in figure 4, we observe a much lower quality after removing the authorship information indicating an overemphasized focus of features and models on the hidden authorship signal. This varies for different settings as adventure/educational shows a loss of 0.09 (blue lines) and educational/social loses 0.23 (red lines). The relatively small loss in the first setting is remarkable as it contains 8 novels per author on average. One would expect the opposite given the weaker author signal of two novels per author for the other categories.

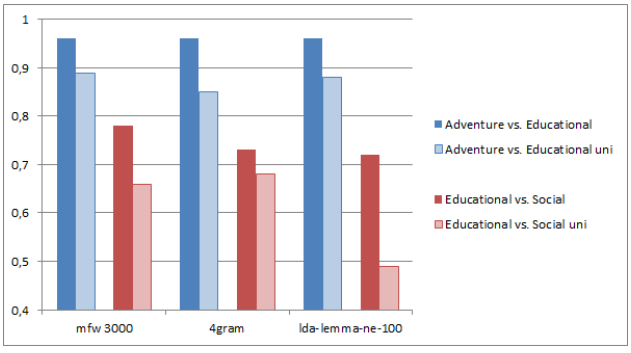


Fig. 4: The influence of authorship

In the next experiment we test whether the combination of feature sets changes our classification results. To balance the size of the feature sets we use Principal Component Analysis (PCA) and construct 100 features from the 3000 mfw and 1000 4gram features each. As shown in figure 5 some feature sets improve when combined (e.g. 4gram100 and lda100) and for others (e.g. lda100 and network) performance decreases. But classification results vary greatly in this setting as signaled by standard deviation bars so these differences should not be overrated.

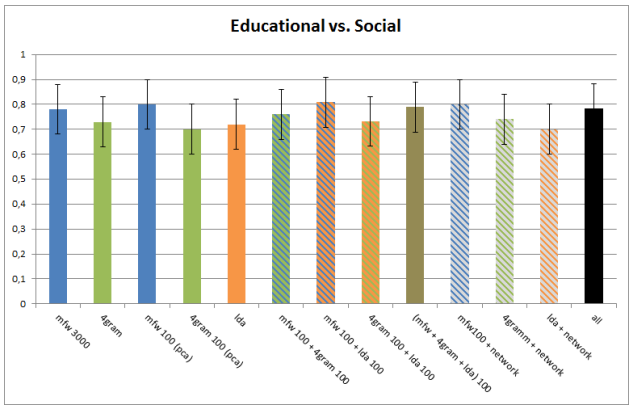


Fig. 5: Classification results for combinations of different feature sets including bars for standard deviation

Conclusion

In this work we classified subgenres of German novels using different feature sets (mfw 3000, 4gram, lda etc.). Some subgenres, like adventure novels, are much easier to classify than others. Most of the applied feature sets showed a varying but comparable performance except the network features. The weak performance of network features might be caused by the weak link between the novel genre and character constellation. The variability of the subgenre signal could not be countered by using higher level features like topics and network characteristics. Interestingly, the author signal has a strong influence on the classification quality. The strength of influence seems to depend on category but is visible in all experiments. In the future, we would like to extend our work by using different network features, work on advanced topic models and find a reliable indicator for plot. Another challenge we have not faced yet is the development of subgenres over time.

Bibliographie

Allison, Sarah / Heuser, Ryan / Jockers, Matthew / Moretti, Franco / Witmore, Michael (2011): *Quantitative Formalism*. An Experiment (= Stanford Literary Lab Pamphlet 1) http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf [letzter Zugriff 09. Februar 2016].

Biber, Douglas (1989): "A typology of English texts", in: *Linguistics* 27: 3-43.

Blei, David / Ng, Andrew / Jordan, Michael (2003): "Latent Dirichlet allocation", in: *The Journal of Machine Learning Research* 3: 993-1022.

Finn, Aidan / Kushmerick, Nicholas (2006): "Learning to classify documents according to genre", in: *Journal of the American Society for Information Science and Technology (JASIST)*. Special Issue on Computational Analysis of Style 57, 11: 1506-1518.

Freund, Luanne / Clarke, Charles L. A. / Toms, Elaine G. (2006): "Towards genre classification for IR in the workplace", in: *Proceedings of the 1st international conference on Information interaction in context* (IIiX). New York, NY: ACM 30-36. http://dx.doi.org/10.1145/1164820.1164829 [letzter Zugriff 09. Februar 2016].

Hettinger, Lena / Becker, Martin / Reger, Isabella / Jannidis, Fotis / Hotho, Andreas (2015): "Genre classification on German novels", in: *Proceedings of the 12th International Workshop on Text-based Information Retrieval*.

Jannidis, Fotis / Krug, Markus / Reger, Isabella / Toepfer, Martin / Weimer, Lukas / Puppe, Frank (2015): "Automatische Erkennung von Figuren in deutschsprachigen Romanen", in: *DHd-Tagung 2015.*. Von Daten zu Erkenntnissen, 23. bis 27. Februar 2015, Graz.

Jockers, Matthew L. (2013): *Macroanalysis*. Digital Methods and Literary History. Champaign: University of Illinois Press.

Krug, Markus (2015): *NERDetection* https://github.com/MarkusKrug/NERDetection [letzter Zugriff 09. Februar 2016].

Petrenz, Philipp / Webber, Bonnie (2011): "Stable classification of text genres", in: *Computational Linguistics* 37, 2: 385-393.

Porter, Martin / Boulton, Richard (2001-2014): *Snowball* http://snowball.tartarus.org/ [letzter Zugriff 09. Februar 2016].

Projekt Gutenberg-DE (1994-): *Projekt Gutenberg-DE* http://gutenberg.spiegel.de/ [letzter Zugriff 09. Februar 2016].

Santini, Marina (2004): "State-of-the-art on Automatic Genre Identification", in: *Technical Report ITRI-04-03, ITRI, University of Brighton (UK)*.

Schmid, Helmut(1994-): *TreeTagger*. A language independent part-of-speech tagger http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [letzter Zugriff 09. Februar 2016].

Sharoff, Serge / Wu, Zhili / Markert, Katja (2010): "The Web Library of Babel: evaluating genre collections", in: *Proceedings of the conference on Language Resources and Evaluation (LREC), Malta* 3063-3070.

TextGrid (2006-2015): *Die Digitale Bibliothek bei TextGrid* [letzter Zugriff 09.Februar 2016].

Underwood, Ted (2014): *Understanding Genre in a Collection of a Million Volumes*. White Paper Report 29.12.2014 http://files.figshare.com/1857045/UnderstandingGenreInterimReport.pdf [letzter Zugriff 09. Februar 2016].

Underwood, Ted / Black, Michael L. / Auvil, Loretta / Capitanu, Boris (2013): "Mapping Mutable Genres in Structurally Complex Volumes", in: *2013 IEEE International Conference on Big Data* http://arxiv.org/abs/1309.3323v2 [letzter Zugriff 09. Februar 2016].

Yu, Bei (2008): "An Evaluation of Text Classification Methods for Literary Study", in: *Literary and Linguistic Computing* 23: 327-343 http://dx.doi.org/10.1093/llc/fqn015 [letzter Zugriff 09. Februar 2016].

Modellierung: eine Begriffsbestimmung

Heßbrüggen-Walter, Stefan

shessbru@hse.ru
National Research University - Higher School of Economics, Russland

Meine Präsentation ist der Frage gewidmet, inwiefern der Modellbegriff zur umfassenden Charakterisierung von Forschungsleistungen innerhalb der *digital humanities* tauglich ist. Dabei gehe ich aus von der sehr allgemeinen Definition William McCartys, der ein Modell als "von Natur aus vereinfachte und deshalb fiktionale oder idealisierte Repräsentation" eines gegebenen Gegenstandes oder Gegenstandsbereiches auffasst und daraus die Behauptung ableitet, die *digital humanities* seien wesentlich mit der heuristischen Exploration solcher digital erfassbarer Modelle von Texten oder anderen Daten befasst (McCarty 2004: par. 3). Was als übergreifendes Erkenntnisziel solcher heuristischer Bemühungen in Frage kommen soll, bleibt bei ihm jedoch im Dunkeln. Zu finden ist lediglich die Bestimmung, dass digitale Modelle "für eine bestimmte Zeit andauernde Zustände in einem Prozess, in dem man zu Wissen gelangt, anstatt fixe Strukturen von Wissen" (McCarty 2004: par. 9). Ich halte diese These für falsch, jedenfalls was den fest etablierten Gebrauch des Modellbegriffs innerhalb der Naturwissenschaften betrifft.

Geisteswissenschaftlerinnen steht es frei, den Begriff des Modells nach eigenem Ermessen unter Berücksichtigung ihrer eigenen methodischen Vorfestlegungen neu zu definieren. Deswegen ist kurz ein zweites, möglicherweise eher einschlägiges Verständnis des 'Modells' in den Geisteswissenschaften zu diskutieren, das in Diskursen des *software engineering* entstanden sein dürfte. In jenen Bereichen, in denen die Methodologie der *digital humanities* sich mit der Generierung von Computercode überschneidet, kann der Modellbegriff legitim gebraucht werden. Jedoch ist Modellierung in diesem Verständnis ein Werkzeug der digitalen Geisteswissenschaften und darf nicht als Selbstzweck

missverstanden werden. Insofern ist Modellierung für die *digital humanities* zwar von Interesse, aber kein Begriff, der unsere Praxis insgesamt erfassen kann.

Modell und Naturwissenschaft

Um die Diskussion übersichtlicher zu gestalten, werde ich mich nicht mit ikonischen Modellen, also etwa dem Billardballmodell von Gasen oder der virtuellen Rekonstruktion einer Predigt von John Donne auseinandersetzen, da hier spezifische Probleme jenseits des allgemein etablierten Modellbegriffs zu erörtern wären (Virtual Paul‘s Cross s. a.). Vielmehr bin ich an dem interessiert, was McCarty als 'komputationale Nachverfolgbarkeit' (*computational tractability*) bezeichnet, da hier vielleicht ein gemeinsames Charakteristikum von Modellen in den digitalen Geisteswissenschaften und mathematischen Modellen in den Naturwissenschaften festzustellen wäre. Hier ist die wesentliche Bezogenheit mathematischer Modelle auf Theorien von besonderem Belang (Portides 2008: 386-387). Mathematische Modelle können in theoriegetriebene Modelle und phänomenologische Modelle unterschieden werden. Theoriegetriebene Modelle werden aus einer bereits etablierten Theorie unter Benutzung von lokalen Hypothesen und entsprechender Randbedigungen abgeleitet. Phänomenologische Modelle umfassen lediglich empirische Daten und ad-hoc-Hypothesen, wenn eine Theorie des Gegenstandsbereiches noch nicht zur Verfügung steht. Theoriegetriebene Modelle machen implizite Folgerungen aus bereits existierenden Theorien explizit. Phänomenologische Modelle sollen uns dabei helfen, überhaupt erst einmal eine Theorie zu formulieren.

Aber beide Arten von Modellen dienen dazu, Vorhersagen über das erwartbare Verhalten des jeweils untersuchten Systems zu ermöglichen (Forster 2008: passim). Solche Vorhersagen sind deswegen möglich, weil Theorien über die Natur Erklärungen von Naturereignissen oder -phänomenen enthalten und damit Wissen über allgemeine kausale Relationen zwischen allgemeinen Klassen von Ereignissen: "Kein Modell steht sozusagen für sich selbst; es greift zurück auf riesige Mengen theoretischen und empirischen Wissens, […] Hintergrundtheorien und empirische Daten verleihen den verschiedenen in einem Modell verwendeten Symbolen Bedeutung." (Barberousse / Ludwig 2009: 61).

Geisteswissenschaftler sind nur selten an Naturgesetzen oder ähnlichen Verallgemeinerungen interessiert. Im Mittelpunkt unseres Interesses steht das einzelne Objekt, selbst wenn es sich etwa um den europäischen Roman zwischen 1800 und 1900 handelt (Moretti 1998).

Dieser Gegenstand ist abstrakt, kann nur durch die Analyse komplexer Daten erforscht werden kann, aber er ist doch ein partikularer Gegenstand. Zwar sind auch Geisteswissenschaftlerinnen an Kausalaussagen interessiert, aber für Historiker sind eher einzelne Kausalaussagen als eventuelle 'Gesetze der Geschichte' von Belang: wichtig ist, warum Caesar den Rubikon überquerte, nicht welche allgemeinen Gesetze für die Überquerung eines Flusses durch Diktatoren gelten.

Fragen, die innerhalb der Wissenschaftstheorie in Bezug auf Modelle erörtert werden, können innerhalb der digitalen Geisteswissenschaften nicht sinnvoll gestellt werden. Ob Modelle als nützliche Fiktionen gelten oder uns zu Wahrheiten über die Welt führen können, ist innerhalb der digitalen Geisteswissenschaften von nachrangiger Bedeutung: unsere Ansprüche auf Wahrheit sind bestenfalls prekär und provisorisch. Ob Modelle Voraussagen über das Verhalten von Systemen ermöglichen ist nicht von Belang, denn wir sind nicht an der Vorhersage der Zukunft interessiert. Wenn Modelle in den Naturwissenschaften immer in Bezug zu einer bereits existierenden oder zukünftigen Theorie des jeweiligen Gegenstandsbereichs stehen, fehlt dieser Bezug, weil die digitalen Geisteswissenschaften keine Theorie in diesem Sinne entwickeln.

Denkbar wäre jedoch, dass der Gebrauch des Modellbegriffs innerhalb der Statistik als Brücke zwischen naturwissenschaftlichen und geisteswissenschaftlichen Verwendungen dienen könnte. Jedoch sind Modelle auch innerhalb der Statistik eng verschränkt mit dem Begriff der Vorhersage. Hier geht es darum, begründen zu können, ob Daten die Behauptung der Wahrheit einer Hypothese rechtfertigen können oder nicht: "[…] alle statistischen Verfahren beruhen auf der Annahme eines statistischen Modells, hier verstanden als jegliche beschränkte Menge statistischer Hypothesen. Außerdem zielen beide [sc. statistische Verfahren und Modelle] auf die Beurteilung dieser Hypothesen." (Romeijn 2014) Wenn eine statistische Hypothese durch ein Modell und ein entsprechendes Verfahren als wahr erwiesen wird, wird sie dann Teil einer wissenschaftlichen Theorie.

Diese 'kanonisierte' Anwendung statistischer Verfahren ist von ihrem Einsatz in den digitalen Geisteswissenschaften deutlich unterschieden. Hier werden statistische Werkzeuge zur Erkennung von Mustern eingesetzt, ohne dass dem eine zuvor definierte Hypothese zugrundeliegen muss, also ohne ein statistisches Modell zur Anwendung zu bringen. Wir suchen gerade Muster, die wir nicht kennen. Und diese Muster sind wieder einzelne (abstrakte) Gegenstände in diesem Text, in diesem Corpus und nicht Allgemeinbegriffe, die sich auf natürliche Arten beziehen, wie sie in den Naturwissenschaften vorausgesetzt werden.

Außerdem sind nicht alle Muster gleich oder gleich wichtig. Ob ein Muster von Interesse ist, hängt nicht von vorhergehenden Hypothesen über den jeweiligen Gegenstandsbereich ab, sondern vielmehr von unserer Kenntnis anderer Einzelgegenstände, ihren Korrelationen, ihrer historischen Entwicklung und der Erfahrung und Urteilskraft des Forschenden, der sich mit diesem

Gegenstandsbereich häufig über lange Zeit vertraut gemacht hat.

Modell und Software

Es existieren also tiefgreifende Unterschiede zwischen dem etablierten Gebrauch des Modellbegriffs in Naturwissenschaften und Statistik und einem denkbaren Gebrauch in den Geisteswissenschaften. Allerdings ist zu beachten, dass der Gebrauch des Ausdrucks "Modell" auch in der Informationstechnologie üblich ist. Vielleicht also leitet sich der Gebrauch des Modellbegriffs in den digitalen Geisteswissenschaften gar nicht aus den Naturwissenschaften ab, sondern aus Theorie und Praxis des Software Engineering. Leider ist der Modellbegriff in diesen Kontexten ähnlich uneindeutig wie in der naturwissenschaftlichen Praxis (Ludewig 2003: 9). Ihm scheint indes die basale Intuition zugrundezuliegen, dass Software als "Modell der Welt" aufzufassen ist (Ludewig 2003: 9-10). In diesem Zusammenhang ist auf zweierlei hinzuweisen: Software mag als Modell der Welt fungieren, dies ist aber nicht ihre einzige Funktion. Software interagiert auch mit der Welt. Damit unterliegt der Erfolg eines Softwareprodukts Einschränkungen, die außerhalb seiner selbst liegen. Wenn in einer Maschine aufgrund eines Denkfehlers im der Steuersoftware zugrundeliegenden Modell eine Fehlfunktion auftritt, kann sich der Entwickler nicht darauf berufen, dass das der Software zugrundeliegende Modell nur als arbiträre Konstruktion eines bestimmten Wirklichkeitsausschnitts gedacht war. Dies gilt auch für Software, die jenen Ausschnitt der Welt repräsentiert, der aus menschlichen Artefakten besteht und von Geisteswissenschaftlerinnen untersucht wird. Führt ein Bug in einem Tokenizer zu falschen Wortzahlen, würden wir die Dinge nicht so lassen, wie sie sind, weil Modelle sowieso nur Approximationen an die Wirklichkeit darstellen.

Wichtiger noch erscheint der Hinweis, dass Repräsentationen der menschlichen Kultur in Softwarewerkzeugen, die in den digitalen Geisteswissenschaften zum Einsatz kommen, nie ein Selbstzweck sind. Selbst wenn man zugesteht, dass Software diese repräsentationale Funktion hat, so soll sie am Ende doch als Werkzeug geisteswissenschaftlicher Forschung funktionieren. Lesen, Denken und Schreiben sind essentielle Bestandteile des analogen geisteswissenschaftlichen Arbeitsprozesses. Aber niemand nähme an, dass diese Aktivitäten intrinsischen Wert haben, wenn sie nicht zum Fortschritt der jeweiligen Disziplin beitragen würden. Auf gleiche Weise sind auch Software und ihre Modelle Teil unseres digitalen Arbeitsprozesses. Solche Repräsentationen haben aber primär instrumentellen Wert und sollten nicht als Selbstzweck missverstanden werden.

Bibliographie

Barberousse, Anouk / Ludwig, Pascal (2009): "Models as Fictions" in: Suárez, Mauricio (ed.): *Fictions in Science*. Philosophical Essays in Modeling and Idealizations. London / New York: Routledge 56-73.

Forster, Malcolm (2008): "Prediction" in: Psillos, Statis / Curd, Marin (eds.): *The Routledge Companion to Philosophy of Science*. London / New York: Routledge 405-413.

Ludewig, Jochen (2003): "Models in software engineering – an introduction" in: *Software and Systems Modeling* 2: 5-14.

McCarty, Willard (2004): "Modeling: A Study in Words and Meanings", in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A Companion to Digital Humanities*. Oxford: Blackwell http://www.digitalhumanities.org/companion/ [letzter Zugriff 14. Oktober 2015].

Moretti, Franco (1998): *Atlas of the European Novel, 1800-1900.* London: Verso.

Portides, Demetris (2008): "Models", in: Psillos, Statis / Curd, Marin (eds.): *The Routledge Companion to Philosophy of Science*. London / New York: Routledge 385-395.

Romeijn, Jan Willem (2014): "Philosophy of Statistics", in: Zalta, Edward (ed.): *The Stanford Encyclopedia of Philosophy (Winter 2014 Edition)* http://plato.stanford.edu/archives/win2014/entries/statistics/ [letzter Zugriff 14. Oktober 2015].

Virtual Paul's Cross Project: *Virtual Paul's Cross*. A Digital Recreation of John Donne's Gunpowder Sermon. Teaching and Visualization Lab, NC State's James B. Hunt Library http://vpcp.chass.ncsu.edu/ [letzter Zugriff: 14. Oktober 2015].

Korpusanalyse in der computergestützten Komparatistik

Ivanovic, Christine
christine_ivanovicl@hotmail.com
Universität Wien, Österreich

Frank, Andrew U.
frank@geoinfo.tuwien.ac.at
Technische Universität Wien

Aufgaben der Vergleichenden Literaturwissenschaft (Komparatistik)

Komparatistische Forschung zielt weniger auf hermeneutische Auslegung von Einzeltexten, als darauf, (a) generalisierende Aussagen über (literarische) Texte, ihre Fomen und Funktionen zu machen, (b) deren historische Entwicklung innerhalb oder (c) im Austausch von kulturellen Systemen zu rekonstruieren, und (d) literarische Repräsentationen von 'Welterfahrung' mit anderen Repräsentationssystemen zu vergleichen.

Zu diesem Zweck vergleicht die Komparatistik eine Vielzahl von Texten (resp. von Texten und anderen künstlerischen Repräsentationsformen) in unterschiedlichen Sprachen. Vergleiche erfordern die Annahme einer Anzahl von Eigenschaften der verglichenen Gegenstände als gleichwertig, während andere Eigenschaften desselben Gegenstands variieren. Die Bestimmung der Eigenschaften eines Textes ist demnach eine der unabdinglichen Voraussetzungen für den komparatistischen Vergleich.

Korpusbasierte Komparatistik

Um allgemeine Aussagen machen zu können, muss die komparatistische Forschung andererseits eine größere Anzahl von Texten untersuchen: sie muss Textkorpora bilden und evaluieren. Die Anzahl und Auswahl der in Betracht gezogenen Texte sowie der evaluierten Texteigenschaften bestimmen maßgeblich die Ergebnisse einer komparatistischen Untersuchung.

Eine computergestützte korpusbasierte komparatistische Untersuchung unterscheidet sich von den bisher praktizierten Ansätzen nicht in der – die Disziplin charakterisierenden – Operation des Vergleichens, wohl aber in der Art und Weise, wie Auswahl und Anzahl der verglichenen Texte resp. Textkorpora begründet und dokumentiert werden.

Computergestützte korpusbasierte Komparatistik

Potentiell sind alle jemals verfassten und mündlich oder schriftlich tradierten Texte aller Sprachen und aller Zeiten Gegenstand der Komparatistik. Ein umfassender systematischer Zugriff auf alle diese Texte ist (bisher) jedoch nicht möglich. Die Möglichkeiten der Evaluierung sind durch verschiedene Faktoren limitiert: nicht alle Texte sind faktisch (mehr) zugänglich, und die jeweilige Forscherperspektive beschränkt grundsätzlich die Erfassung der zum Vergleich herangezogenen Texte.

Bisher sind die Kriterien für die Textauswahl in wesentlichem Maße abhängig gewesen von (a) der Subjektivität und (b) der natürlicherweise begrenzten Kapazitätder Forscher, die nur die ihnen bekannten Texte berücksichtigen können und die unter dem Credo arbeiten, nur Texte zu erforschen, die ihnen in der Originalfassung zugänglich sind. Dies führt dazu, dass die Komparatistik bisher mehrheitlich Texte aus den dominanten Sprachen (Englisch, Französisch,...) bearbeitet und Texte in 'kleinen' Sprachen (Finnisch, Urdu,....) oder Textvergleiche zwischen kaum verwandten Sprachen (Chinesisch gegen Arabisch) eher selten vorkommen.

Ein weiteres Problem der Textauswahl stellt (c) die Gefahr des logischen Zirkelschlusses dar: Bei der Evaluation beispielsweise “des” europäischen Romans werden aus der Lesepraxis resp. -tradition herrührende Vorannahmen in die Auswahl einbezogen, wenn es darum geht, dieses Genre anhand verschiedener Beispiele zu bestimmen; sie haben unweigerlich Einfluss auf das erzielte Ergebnis. Schließlich beruhen, und auch dies bedeutet eine wesentliche Einschränkung, (d) generalisierende Aussagen wie über “den europäischen Roman” immer auf einer im Vergleich zur Gesamtmenge der je produzierten Texte verschwindend kleinen Auswahl.

Die Auswahl der evaluierten Texte kann bei einer computergestützten korpusorientierten komparatistischen Untersuchung zumindest statistisch anders begründet werden:

* durch einen definitiv bestimmten Korpus, der so angelegt ist, dass er Repräsentativität beanspruchen kann
* durch einen Korpus, der in seinem Umfang weit über das Lesevermögen des Einzelnen hinausreicht und der große Textmengen in einer Vielfalt von Sprachen umfaßt, die kein Einzelleser je bewältigen könnte;
* durch die Möglichkeit der Überprüfung der erzielten Ergebnisse in Wiederholungs- und Vergleichsstudien sowie mittles Vergleichskorpora;
* durch die Trennung der Auswahlkriterien für die Erstellung des Korpus von den fokussierten Untersuchungsergebnissen;
* durch die Möglichkeit von Negativabfragen (z. B. eine bestimmte Eigenschaft ist in einigen der Texte des Korpus nachweisbar, während andere Eigenschaften in keinem davon nachweisbar sind).

Anforderungen an die computergestützte korpusbasierte Komparatistik

Computergestützt korpusbasiert arbeitende Komparatistik sieht sich mit folgenden Aufgaben konfrontiert:

Erarbeitung einer effizienten Infrastruktur

Für den Aufbau und die Pflege großer Textkorpora bedarf es entsprechend bearbeiteter Texte: alle in den Korpus aufgenommene Texte müssen bibliographisch genau erfasst und mit Markups (Taggern) versehen sein.

Markups können gesetzt werden u. a. zur Auszeichnung der Sprachform (insbesondere bei mehrspachigen Texten), der Textstruktur, nicht-literarischer Elemente (z. B. Abbildungen im Text) etc. Bevorzugt werden treebank getaggte Versionen mit verzweiger Struktur (parse tree), die Koreferenzen, Personen- und Orstnamen u. a. erkennen lassen. Im Textmarkup werden einzelne Elemente der Textstruktur identifiziert: Worte oder Wortbestandteile, Sätze und deren Teile, Abschnitte, Kapitel und andere Texteinteilungen bis hin zum Buchlayout. Es erscheint wichtig, auch die Elemente zu erfassen, die nicht unmittelbar textimmantent sind, die aber zur Identifizierung und Charakterisierung des Textes gehören wie Seitenangaben, Verfassername und weitere Angaben, die im Rahmen einer Buchpublikation vorkommen. Der Text sollte in UTF-8 codiert sein, um auch Texte in nicht-alphabetischen Sprachen wie Chinesisch, Arabisch u.w.m. einbeziehen zu können. Unserer Konzeption nach sollen die Markierungen in den Text hineingesetzt werden, so dass der mit den Annotationen versehene Text mit der Originalstruktur verbunden bleibt.

Der mit Markups versehene Text und die POS-Annotationen werden in ein einziges Format zusammengefasst. Wir bevorzugen derzeit RDF (Manola / Miller 2004), das für die von uns avisierte Datenmenge auszureichen scheint. Unseren bisherigen Beobachtungen nach erhalten wir bei einem Text mit reichhaltiger linguistischer Auszeichnung für jedes Wort etwa 10 RDF Triples; bei einem literarischen Text von 100.000 Wörtern würde das eine Million Triples ergeben, bei einem Korpus von 10.000 Büchern wäre man mit 10 Milliarden Triples noch bei weitem innerhalb des Rahmens dessen, was das heutige RDF Depot erlaubt; Untersuchungen zur derzeitigen Kapazitätsgrenze haben für 1 Billiarde Triples eine Hochladezeit von wenigen Stunden ergeben (Boncz / Pham 2013). Die Antwortzeiten nehmen bei unterschiedlichen Abfragen nicht ab und die Anforderungen der Hardware bleiben im Rahmen des für ein Projekt Möglichen (die Anschaffungskosten der Versuchskonfiguration von 2012 beliefen sich auf 70.000 Euro).

Entwicklung brauchbarer Methoden zur Abfrage und digitalen Analyse von Texten

Es müssen Methoden sein, die Abfragen und Analysen von Texten ermöglichen unabhängig von der Sprache, in der sie verfasst sind, d. h. wir arbeiten an (statistischen) Methoden, die Texteigenschaften in (mathematische) Werte 'übersetzen', um eine Vergleichbarkeit von Textstrukturen über die Sprachgrenzen hinweg zu ermöglichen. Die einschlägige Fachliteratur kennt bereits eine große Anzahl derartiger Methoden; auf viele von ihnen kann man über das web zugreifen. Die computergestützt korpusbasiert arbeitende Literaturwissenschaft erlaubt z. B. die Identifizierung von Zitaten und intertextuellen Bezugnahmen (Ganascia et al. 2014). Sie dient dem Autornachweis oder der Evaluierung von gender-bedingten Spezifika (Stanczyk 2011), wie auch der Feststellung literarischer Moden und Bewegungen (Amancioa et al. 2012). Sie unterstützt die Analyse von Emotionen (Dichiu et al. 2010), oder die Rekonstruktion von Wissenstransfer (Cappelli et al. 2001).

Einführung von Clusteranalyse in die Komparatistik

Die Anwendung aller Methoden auf alle Texte generiert eine Matrix evaluierbarer Werte; jeder Text läßt sich durch einen Vektor aus diesen Werten darstellen. Diese Darstellungsweise ermöglicht Textvergleiche mittels Clusteranalyse wie sie in der konventionellen Komparatistik aufgrund der o.g. Beschränkungen bisher nicht zugänglich waren.

Korpusaufbau und Abfragemethoden müssen so gestaltet sein, dass Texte umstandslos dem Korpus hinzugefügt werden und die Methoden problemlos appliziert werden können. Dies setzt kontinuierliche Pflege und Aktualisierung des bestehenden Korpus resp. der Abfrageergebnisse voraus: wenn Texte hinzugefügt werden, müssen alle bisher angewandten Methoden automatisch darauf angewandt werden können; wenn Methoden hinzugefügt werden, müssen automatisch alle Texte einer entsprechenden Evaluierung unterzogen werden.

Zusammenfassung

In unserem Beitrag treten wir für die Etablierung eines computergestützten korpusbasierten Forschungsansatzes in der literaturwissenschaftlichen Komparatistik ein. Dazu wollen wir darstellen, (a) welche Vorteile die Erstellung umfangreicher Korpora literarischer Texte aus verschiedenen Sprachen für die komparatistische Analyse bietet, (b) wie sie konstruiert und gepflegt werden können, und (c) welche Abfragemöglichkeiten auf dem gegenwärtigen Stand der Technik sie bieten. In Betracht gezogen werden dafür sowohl bereits vorhandene und online zugängliche Korpora wie auch von einzelnen Forschergruppen erarbeitete, intern genutzte Korpora wie das Austrian Academy Corpus am ICLTT der ÖAW. Des weiteren wollen wir einen kursorischen Überblick über die bisher erprobten Ansätze computergestützter literaturwissenschaftlicher Analyse geben, um das gegenwärtige Spektrum der Methoden der Textevaluierung darstellen und zukünftige Desiderata aufzeigen zu können.

Bibliographie

Amancio, Diego R. /Oliveira Jr., Osvaldo N. / F. Costa, Luciano da (2012): "Identification of literary movements using complex networks to represent texts", in: *New Journal of Physics* 14 http://arxiv.org/abs/1302.4099 [letzter Zugriff 09. Januar 2016].

Boncz, Peter / Pham, Minh-Duc (2013): *BSBM V3.1 Results (April 2013)*. Centrum Wiskunde & Informatica, Amsterdam http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V7/ [letzter Zugriff 09. Januar 2016].

Cappelli, Amedeo / Catarsi, Maria Novella / Michelassi, Patrizia / Moretti, Lorenzo / Baglioni, Miriam / Turini, Franco / Tavoni, Mirko (2002): "Knowledge mining and discovery for searching in literary texts", in: *Proceedings of the Third International Conference on Language Resources and Evaluation*, LREC 2001, May 29-31, 2002, Las Palmas, Canary Islands, Spain.

Dichiu, Daniel /Pais, Ana Lucia / Moga, Sunita Andrea / Buiu, Catalin (2010): "A cognitive system for detecting emotions in literary texts and transposing them into drawings", in: *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. Istanbul, Turkey, 10-13 October 2010: 1958-1965.

Ganascia, Jean-Gabriel / Glaudes, Pierre / Del Lungo, Andrea (2014): "Automatic detection of reuses and citations in literary texts", in: *Literary and Linguistic Computing* 29, 3: 412-421.

Manola, Frank / Miller, Eric (eds.) (2004): *RDF primer 1.0*. W3C recommendation, 10 (1-107): 6 https://www.w3.org/TR/2004/REC-rdf-primer-20040210/ [letzter Zugriff 09. Januar 2016].

Moretti, Franco (1999): *Atlas of the European novel*. 1800-1900. London, New York: Verso.

Stanczyk, Ursula (2011): "Recognition of author gender for literary texts", in: *Man-Machine Interactions* 2. Proceedings of the 2nd International Conference on Man-Machine Interactions, ICMMI 2011, The Beskids, Poland.

Takeda, Masayuki / Fukuda, Tomoko / Nanri, Ichiro (2002): "Mining from literary texts: Pattern discovery and similarity computation", in: *Progress in Discovery Science*. Final Reports of the Japanese Discovery Science Project. Berlin / Heidelberg: Springer 518-531.

Kollaboratives Annotieren literarischer Texte Eine Anleitung

Jacke, Janina
janina.jacke@uni-hamburg.de
Universität Hamburg, Deutschland

Gius, Evelyn
evelyn.gius@uni-hamburg.de
Universität Hamburg, Deutschland

Kollaboratives Annotieren ist eine gute Möglichkeit, um mehr als bloß eine subjektive Perspektive auf den Untersuchungsgegenstand – beispielsweise einen Text – abzubilden: Sobald mehr als ein Individuum MarkUp an einem digitalen Objekt anbringt, kann deutlich werden, welche unterschiedlichen Aspekte des Objekts im Zentrum des individuellen Interesses stehen oder welche verschiedenen Sichtweisen in Bezug auf denselben Aspekt möglich sind. Soll die kollaborative Annotation jedoch nicht bloß die Pluralität von Perspektiven und Meinungen aufzeigen, sondern einem spezifischeren Erkenntnisinteresse dienen, so sollte die Annotationspraxis mithilfe von Guidelines strukturiert und reguliert werden. Für die Annotation linguistischer Phänomene (bspw. in Gebrauchstexten) werden solche bereits entwickelt (vgl. bspw. Pyysalo / Ginter 2014; Mamoouri et al. 2008); dagegen existieren für die kollaborative Annotation semantischer Phänomene in literarischen Texten kaum *best practice*-Vorschläge. Eine direkte Übertragung linguistischer Annotationsanleitungen auf den literaturwissenschaftlichen Bereich scheint dabei aus mindestens drei Gründen nicht möglich:

Literarische Texte sind in der Regel polyvalent, d. h. mehrdeutig: Während kollaboratives Annotieren im Bereich der Linguistik letztlich der Vermeidung individueller Annotations*fehler* dient und eine autoritative Version zum Ziel hat, muss im Falle literaturwissenschaftlicher Annotation immer bedacht werden, dass auch unterschiedliche bzw. widersprüchliche Annotationen gleichermaßen legitime Lesarten ausdrücken können – ohne dabei jedoch in Beliebigkeit abzugleiten.

Die Analysekategorien, mithilfe derer literarische Texte untersucht bzw. annotiert werden, sind häufig unterdefiniert. Grund hierfür scheint zum einen die Tatsache zu sein, dass Textanalyse und -interpretation in der traditionell-literaturwissenschaftlichen Praxis häufig weniger textnah ausgeführt werden als im digitalen Kontext, was dazu führt, dass vage Definitionen keine Anwendungsprobleme mit sich bringen. Zum anderen zeichnet sich die traditionelle Literaturwissenschaft durch individuelles Arbeiten und die Erschließung neuer Lesarten von Texten aus. Dadurch bleibt häufig unbemerkt, dass die Analysekategorien nicht konsistent verwendet werden. Im Kontext der kollaborativen Annotation literarischer Texte muss also die Frage berücksichtigt werden, wie spezifisch und klar die Definition

literaturwissenschaftlicher Annotationskategorien ausfallen kann und sollte.

Literaturwissenschaftliche Analysekategorien stehen häufig in bisher untertheoretisierten Abhängigkeitsverhältnissen zueinander. Das kann dazu führen, dass Annotationen deshalb unterschiedlich ausfallen, weil die Annotatoren im Kontext impliziter Analyseschritte, die der Annotation logisch vorgeordnet sind, zu unterschiedlichen Ergebnissen gekommen sind. Auch für den Umgang mit solchen impliziten Abhängigkeitsverhältnissen muss im Kontext literarischer Annotation eine Regelung gefunden werden.

Im Folgenden möchten wir einen *best practice* Vorschlag für das kollaborative Annotieren literarischer Texte vorstellen, der die oben genannten Schwierigkeiten berücksichtigt. Da *best practices* – besonders im literaturwissenschaftlichen Bereich – allerdings nicht vollständig unabhängig vom zugrundeliegenden Erkenntnisinteresse sind, stellen wir unserem *How to* eine kurze Beschreibung des Projektkontextes voran, im Rahmen dessen die Guidelines entstanden sind. Mit kleineren, am jeweiligen Erkenntnisinteresse orientierten Modifikationen sollte diese Anleitung allerdings problemlos auf anders ausgerichtete Projekte übertragbar sein.

Die Annotationsguidelines sind im Kontext des Projekts heureCLÉA entstanden (vgl. Bögel et al. im Erscheinen). Ziel des Projekts ist die Entwicklung einer digitalen Heuristik, d. h. eines Funktionsmoduls, das automatisch semantische Phänomene in literarischen Texten – hier: narratologische Phänomene der Zeitgestaltung [1] – annotiert. Für die Generierung dieses Tools wurde zunächst ein Korpus literarischer Texte kollaborativ in Bezug auf die zu automatisierenden Phänomene annotiert. Basierend auf diesen Annotationen soll die Funktionalität dann mithilfe regelbasierter Verfahren und Machine Learning-Methoden entwickelt werden.

Nach einigen Anläufen hat sich folgende Annotationspraxis als *best practice* herausgestellt (vgl. Abbildung 1):

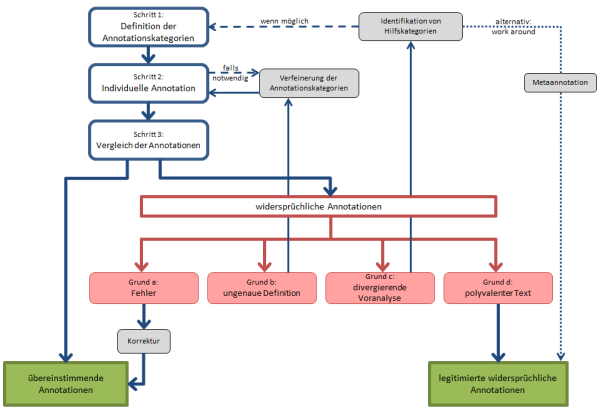


Abb. 1: Ablaufschema kollaborative literaturwissenschaftliche Annotation

Schritt 1: *Möglichst genaue Definition der Annotationskategorien.* Es ist sinnvoll, die zugrundeliegenden Annotationskategorien bereits vor der Annotation so spezifisch wie möglich zu definieren. Geschieht dies nicht, so ist es letztlich schwieriger festzustellen, ob sich Annotationen tatsächlich auf Basis der Polyvalenz des literarischen Textes unterscheiden oder aufgrund vager Kategoriendefinitionen. Wenn die Definition einer traditionellen literaturwissenschaftlichen Kategorie auf zwei unterschiedliche Arten verstanden werden kann und beide Varianten interessante Textphänomene beschreiben, dann ist es problemlos möglich, beide Varianten zu operationalisieren – allerdings als sorgfältig getrennte Konzepte. Die Forderung der möglichst genauen Definition geht also nicht notwendigerweise mit einer Beschränkung des Erkenntnisinteresses einher. Zusätzlich zur Definition der Annotationskategorie sollten auch die Textoberflächenindikatoren festgehalten werden, die auf das jeweilige Phänomen hinweisen, sowie die Granularität der annotierten Zeichenkette (Fragment, Wort, Teilsatz, Satz etc.). Die Festlegungen werden in Form von Annotationsguidelines dokumentiert, die allen Annotatoren als Annotationsbasis dienen.

Schritt 2: *Individuelle Annotation.* Jeder Annotator annotiert auf Basis der Guidelines die Korpustexte individuell. Dadurch soll gewährleistet werden, dass nicht gleich zu Beginn eine gegenseitige Beeinflussung der Anntotatoren besteht, sondern unterschiedliche Lesarten auch tatsächlich in den Annotationen abgebildet werden.

falls notwendig: Verfeinerung der Annotationskategorien. Eventuell stellt sich bereits in der individuellen Annotationsrunde heraus, dass einige Annotationskategorien noch nicht spezifisch genug definiert sind, um eine regelgeleitete Anwendung zu gewährleisten. Ist dies der Fall, so müssen die Kategorien spezifiziert werden, damit die individuelle Annotationsrunde sinnvoll beendet werden kann.

Schritt 3: *Vergleich der Annotationen*. Nach Abschluss der individuellen Annotation vergleichen die Annotatoren ihre Arbeit. Im Fall diskrepanter Annotationen werden die jeweiligen Gründe der Annotationsentscheidung diskutiert. Generell können vier Arten von Gründen für widersprüchliche Annotationen auftreten, die unterschiedliche Maßnahmen erfordern:

falsche Annotation: Eine der Annotationen basiert auf einem eindeutig falschen Verständnis der fraglichen Textstelle oder der Kategoriedefinition. In diesem Fall muss die falsche Annotation korrigiert werden.

ungenaue Definition: Es ist möglich, dass der Vergleich der Annotationen weitere Defizite der Kategoriedefinition offenbart, die im Kontext vorheriger Stadien nicht deutlich geworden sind. Ist dies der Fall, muss die Definition (ein weiteres Mal) überarbeitet werden. Es folgt eine weitere individuelle Annotationsphase, gefolgt von einem Vergleich der Annotationen.

divergierende Voranalyse: Ein Vergleich der diskrepanten Annotationen kann ergeben, dass die Anwendung bestimmter Annotationskategorien vorbereitende Analyseschritte notwendig macht. Diese Schritte sind von den Annotatoren jeweils implizit ausgeführt worden. Wenn diese Analysen zu unterschiedlichen Ergebnissen führen, dann können auch die darauf aufbauenden Annotationen variieren. Ist dies der Fall, so müssen die für die vorbereitenden Analyseschritte notwendigen Hilfskategorien identifiziert und definiert werden. Wenn der Projektrahmen es zulässt, dann sollte das gesamte Textkorpus auch unter Rückgriff auf die Hilfskategorien nach dem hier dargestellten Ablaufschema annotiert werden. Die Hilfsannotationen können dann ebenfalls auf ihre Richtigkeit hin überprüft werden, wodurch die Hauptannotationen in der Regel deutlich einheitlicher ausfallen. Ist es aus arbeitsökonomischen Gründen nicht möglich, eine Voranalyse des Korpus anhand von Hilfskategorien durchzuführen, so ist ein *work around* möglich: Anstatt den Versuch zu unternehmen, die vorbereitenden Analysen zu vereinheitlichen, können die Annotatoren ihre Hauptannotationen mit Metaannotationen versehen, die die Ergebnisse der Voranalysen festhalten. Auf diese Weise wird nicht die Frage nach der Korrektheit der Voranalysen gestellt, wohl aber der Grund für die divergierenden Hauptannotationen explizit gemacht. [2]

polyvalenter Text: Wird als Grund für eine widersprüchliche Annotation textuelle Mehrdeutigkeit herausgestellt, so müssen die Annotationen nicht überarbeitet werden, sondern die widersprüchlichen Annotationen werden als legitimiert verstanden.

Wie deutlich geworden ist, werden die eingangs genannten drei Probleme literarischer Annotation im Kontext dieses Ablaufschemas berücksichtigt: Die Möglichkeit, einen literarischen Text unterschiedlich zu deuten, wird zum einen durch die individuelle Annotationsphase (vgl. Schritt 2) gewährleistet, zum anderen dadurch, dass widersprüchliche Annotationen erlaubt sind, sofern sie durch die Polyvalenz des Textes bedingt sind (vgl. Grund d). Dass die Berücksichtigung der Polyvalenz nicht in eine Beliebigkeit von Annotationsentscheidungen ableitet, wird dadurch erreicht, dass andere Gründe (d. h. mindestens Gründe a und b) für widersprüchliche Annotationen ausgeschlossen werden. Die Spezifikation literaturwissenschaftlicher Annotationskategorien wird schrittweise optimiert, um aussagekräftige Annotationsergebnisse zu gewährleisten (vgl. Schritt 1 sowie ggf. weitere Optimierungsschritte ausgehend von Schritt 2 und Grund b). Da klare Definiertheit nicht notwendig mit einer Einschränkung der Perspektive auf Texte einhergeht (vgl. Schritt 1), ist sie mit der literaturwissenschaftlichen Praxis bzw. mit der Pluralität möglicher Erkenntnisinteressen kompatibel. Die Abhängigkeit literaturwissenschaftlicher Kategorien untereinander wird schließlich, je nach verfügbaren Ressourcen, entweder in einem reduzierten oder in einem ausführlichen Ansatz explizit gemacht (vgl. Grund c).

Notes

1. Die zentralen zeitlichen Phänomene in heureCLÉA beziehen sich auf die temporale Relation zwischen erzählter Geschichte und ihrer Repräsentation. Diese Relation kann in dreierlei Hinsicht untersucht werden: *Ordnung* bzw. Reihenfolge (Wann findet ein Ereignis statt? – Wann wird es erzählt?), *Frequenz* bzw. Häufigkeit (Wie oft findet es statt? – Wie oft wird es erzählt?) und *Dauer* bzw. Geschwindigkeit (Wie lange dauert es? – Wie lange dauert es, davon zu erzählen?) (vgl. Genette 1994).

2. Ein Beispiel für eine solche Hilfskategorie in heureCLÉA ist die der *Erzählebenen* (Finden innerhalb einer Erzählung weitere eingebettete Erzählungen statt?, vgl. Ryan 1991). Die Identifikation von Erzählebenen beeinflusst die Analyse zeitlicher Phänomene wie *Dauer*: Wo immer eine eingebettete Erzählung auftaucht, kann die Erzähldauer entweder auf der Ebene der Haupterzählung oder auf der der eingebetteten Erzählung analysiert werden. In heureCLÉA wurden deswegen narrative Ebenen vor der finalen Annotation von Dauer annotiert. Ein *work around* hätte folgendermaßen aussehen können: Wann immer die Diskussion widersprüchlicher Annotationen ergibt, dass einer der Anntatoren eine eingebettete Erzählung identifiziert hat und der andere nicht bzw. dass beide Annotatoren eine eingebettete Erzählungen identifiziert haben, sich ihre Daueranntotationen sich jedoch auf unterschiedliche Ebenen beziehen, wird dies in Form einer Metaannotation festgehalten. – Das Ergebnis der oben genannten Arbeitsschritte in heureCLÉA – die heureCLÉA Annotationsguidelines Version 1.0 – sind unter www.heureCLÉA.de/guidelines einsehbar.

Bibliographie

Bögel, Thomas / Gertz, Michael / Gius, Evelyn / Jacke, Janina / Meister, Jan Christoph / Petris, Marco / Strötgen, Jannik (im Erscheinen): "Collaborative Text Annotation Meets Machine Learning.

heureCLÉA, a Digital Heuristic of Narrative", in: *DHCommons* 1.

Genette, Gérard (1994): *Die Erzählung*. München: Fink.

Gius, Evelyn / Jacke, Janina (2015): *Zur Annotation narratologischer Kategorien der Zeit.* Guidelines zur Nutzung des CATMA-Tagsets (Version 1.0) www.heureclea.de/guidelines [letzter Zugriff 11. Oktober 2015].

Maamouri, Mohamed / Bies, Ann / Kulick, Seth (2008): "Enhancing the Arabic Treebank. A Collaborative Effort toward New Annotation Guidelines", in: *7th International Conference on Language Resources and Evaluation*. Marrakech (LREC 2008) https://catalog.ldc.upenn.edu/docs/LDC2010T13/Enhancing_Arabic_Treebank.pdf [letzter Zugriff 11.Oktober 2015].

Pyysalo, Sampo / Ginter, Filip (2014): "Collaborative development of annotation guidelines with application to Universal Dependencies", in: *Fifth Swedish Language Technology Conference*. Uppsala (SLTC 2014) http://www2.lingfil.uu.se/SLTC2014/abstracts/sltc2014_submission_32.pdf [letzter Zugriff 11.Oktober 2015].

Ryan, Marie-Laure (1991): *Possible Worlds, Artificial Intelligence, and Narrative Theory*. Bloomington, Ind.: Indiana University Press.

DiaCollo: diachronen Kollokationen auf der Spur

Jurish, Bryan
jurish@bbaw.de
Berlin-Brandenburg Akademie der Wissenschaften, Deutschland

Geyken, Alexander
geyken@bbaw.de
Berlin-Brandenburg Akademie der Wissenschaften, Deutschland

Werneke, Thomas
werneke@zzf-pdm.de
Zentrum für Zeithistorische Forschung, Deutschland

Abstract

Wir präsentieren *DiaCollo*, ein Softwarewerkzeug zur effizienten Extraktion, zum Vergleich und zur interaktiven Visualisierung von Kollokationen aus einem diachronen Textkorpus. Im Gegensatz zu konventioneller Kollokationssoftware eignet sich *DiaCollo* für die Verarbeitung diachroner Kollokationsdaten: Kollokationspaare, deren Assoziationsstärke vom Zeitpunkt ihres Auftretens abhängt. Durch das Aufspüren von Veränderungen in den charakteristischen Kollokaten eines Worts im zeitlichen Verlauf kann *DiaCollo* dazu beitragen, ein klareres Bild von Veränderungsprozessen der Wortsemantik zu zeichnen.

Einführung

In den letzten Jahren sind immer mehr große diachrone Textkorpora zu Forschungszwecken verfügbar gemacht geworden (z. B. Geyken et al. 2011; Davies 2012). Die durch diese Korpora abgedeckten großen Zeitspannen stellen diverse Herausforderungen an konventionelle Techniken der maschinellen Verarbeitung natürlicher Sprache, die ihrerseits oft auf impliziten Annahmen der Korpushomogenität basieren – insbesondere der zeitliche Achse betreffend. Tatsächlich haben sogar vermeintlich synchrone Zeitungskorpora eine nicht triviale temporale Extension und können bei entsprechender Behandlung zeitabhängige Phänomene zeigen (Scharloth et al. 2013). In dieser Arbeit gehen wir auf das Problem der automatischen Erstellung von Kollokationensprofilen (Church / Hanks 1990 und Evert 2005) in diachronen Korpora ein, indem wir ein neues, explizit für diesen Zweck entwickeltes Softwarewerkzeug *DiaCollo* vorstellen, das dem Benutzer ermöglicht, für jede Abfrage selber die Granularität der diachronen Achse frei zu wählen. Im Gegensatz zu konventionellen Kollokationswerkzeugen wie dem *DWDS Wortprofil* (Didakowski / Geyken 2014) oder dem *Sketch Engine* (Kilgarriff / Tugwell 2002) eignet sich *DiaCollo* zur Extraktion und Analyse diachroner Kollokationsdaten: Kollokationspaare, deren Assoziationsstärke von dem Zeitpunkt ihres Auftretens abhängt. Durch das Aufspüren von Veränderungen in den typischen Kollokaten eines Worts im zeitlichen Verlauf und Anwendung von J. R. Firths berühmtem Prinzip “ *you shall know a word by the company it keeps*”, kann *DiaCollo* helfen, ein klareres Bild diachroner Veränderungen im Wortgebrauch zu liefern.

Implementierung

DiaCollo ist als modulare Perl Bibliothek implementiert, einschließlich wiederverwendbaren Klassen zum Umgang mit nativen Binärindexstrukturen. *DiaCollo* Indizes sind für Hochlastumgebungen geeignet, da kein persistenter Server-Prozess benötigt wird und jeglicher Laufzeitzugriff auf native Indexstrukturen über direkten Dateisystem I/O stattfindet. Über die programmatische API der Perl-Module hinaus bietet *DiaCollo* sowohl eine Befehlszeilenschnittstelle als auch einen öffentlich zugänglichen RESTful Webservice mit einer formularbasierten Benutzerschnittstelle zur Auswertung von Datenbankanfragen und einer

interaktiven Visualisierung der Anfrageergebnisse. Ein öffentlich zugängliches Web-Frontend für das Korpus des Deutschen Textarchivs ist unter http://kaskade.dwds.de/dstar/dta/diacollo zu finden; der vollständige Quellcode ist via CPAN erhältlich.

Anfragen & Parameter

DiaCollo ist ein anfrageorientierter Dienst. Er behandelt eine Benutzeranfrage als eine Menge von (*Parameter=Wert*)-Paaren und liefert ein korrespondierendes Profil für den/die angefragten Term(e) zurück. Die Parameter werden wie bei einem üblichen Web-Formular an den Service RESTfully via HTTP GET oder POST Anfrage überreicht. Jede Anfrage muss einen *query* Parameter enthalten, der den oder die zu profilierenden Zielterm(e) spezifiziert. Der *date* Parameter selegiert die gewünschte Zeitspanne, während die Granularität der zurückgelieferten Profildaten mithilfe des *slice* Parameters durch Angabe der Grösse einer einzelnen Profilepoche festgelegt werden kann. Kollokatkandidaten können über den *groupby* Parameter gefiltert werden, und die Bereinigung ("pruning") der zurückzuliefernden 'besten' Kollokaten wird von den Parametern *score*, *kbest* und *global* gesteuert.

Profile & Diffs

Das Ergebnis einer einfachen *DiaCollo* Anfrage wird als tabellarisches Profil der *k*-best Kollokate für den/die angefragte(n) Term(e) in jedem der angefragten Zeit-Subintervalle ausgegeben ("Epochen" oder "Slices", e.g. Dekaden), die mit den Parameter *date* und *slice* spezifiziert wurden. Als Alternative kann der Benutzer auch ein Vergleichs- bzw. "Diff"-Profil anfordern, um die salientesten Unterschiede zwischen zwei unabhängigen Anfragen hervorzuheben; z. B. zwischen zwei verschiedenen Worten oder zwischen den Vorkommen eines Wortes in verschiedenen Zeitintervallen, Teilkorpora oder lexikalischen Umgebungen.

Indizes, Attribute & Aggregierung

DiaCollo benutzt eine interne "native" Indexstruktur über alle Inhaltswörter des Eingabekorpus, um Kollokationsprofile zu berechnen. Jedes indizierte Wort wird als *n*-Tupel linguistisch relevanten Token- oder Dokumentattribute behandelt, zusätzlich zum Dokumentdatum. Die Attribute *Lemma (l)* und *Pos (p)* ("part-of-speech") werden per Default indiziert. Die Anfrageparameter *query* und *groupby* werden als logische Konjunktionen von Suchkriterien bezüglich dieser Attribute interpretiert, um die genauen zu profilierenden Token-Tupel zu selegieren. Um eine feinkörnigere Auswahl von Profilzielen zu ermöglichen, unterstützt DiaCollo den gesamten Umfang der DDC-Abfragesprache (Sokirko 2003; Jurish et al. 2014), wenn die DiaCollo-Instanz mit einem zugrundeliegenden DDC Server assoziiert ist.

Scoring & Pruning

DiaCollo weist jedem Kollokat w_2 eines unären Profils für einen Zielterm w_1 mittels einer benutzerspezifizierten *Scorefunktion* einen reellwertigen Assoziationswert ("score") zu. Zu den unterstützten Scorefunktionen zählen absolute und logarithmische Frequenzen (*f, lf*), normierte absolute und logarithmische Frequenzen pro Mio. Token (*fm, lfm*), das *pointwise mutual information* × log Frequenz Produkt (*mi*), und der von Rychlý (2008) vorgeschlagene skalierte log-Dice Koeffizient (*ld*). Kollokatkandidaten werden nach Assoziationswert absteigend geordnet und die *k*-besten Kandidaten jeder Epoche ausgewählt und zurückgegeben. Für "diff" Anfragen werden unabhängige Profile p_a und p_b jeweils für die *query* und *bquery* Parameter berechnet. Nach der Sortierung anhand der selegierten Scorefunktion wird ein Vergleichsprofil p_{a-b} berechnet als $p_{a-b} : w_2 \rightarrow p_a(w_2) - p_b(w_2)$ für jeden der bis zu $2k$ Kollokate w_2 # k-best(p_a) # k-best(p_b), wonach die *k*-besten von diesen Kandidaten mit den größten absoluten Unterschieden $|p_{a-b}(w_2)|$ selegiert und zurückgegeben werden.

Ausgabe & Visualisierung

DiaCollo unterstützt verschiedene Ausgabeformate für die zurückgelieferte Profildaten, darunter TAB-getrennten Text, natives JSON für die weitere automatische Verarbeitung sowie einfaches tabellarisches HTML. Zusätzlich zu den statischen tabellarischen Formaten bietet der Webservice-Plugin auch mehrere interaktive Online-Visualisierungen für diachrone Profildaten, unter anderem zweidimensionale Zeitreihen mithilfe der Highcharts JavaScript Bibliothek, Flash-basierten Motion Charts mithilfe der Google Motion Charts Bibliothek und dynamische Bubble- und Tag-Cloud Visualisierungen mithilfe der D3.js Bibliothek. Das HTML sowie die D3-basierten Formate bieten eine intuitive farbkodierte Repräsentation der Assoziationsscores (bzw. Score-Unterschiede bei "diff"-Profilen) für jedes Kollokationspaar sowie Hyperlinks zu den zugrundeliegenden Korpus-Treffer ("KWIC-links") für jeden abgebildeten Datenpunkt. Beispiele für die Zeitreihen-, Tag-Cloud- und Bubble-Visualisierungen sind in den Abbildungen 1–3 zu finden.

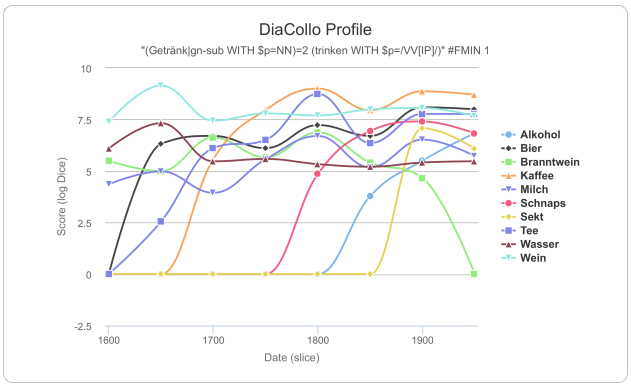


Abb. 1: *DiaCollo* Zeitreihe für die zehn global besten nominale Hyponyme des GermaNet (Hamp / Feldweg 1997; Henrich / Hinrichs 2010) SynSets *Getränk* unmittelbar links vom Verb *trinken* in 50-Jahres Epochen über den Gesamtbestand des Deutschen Textarchivs und DWDS-Kernkorpus.

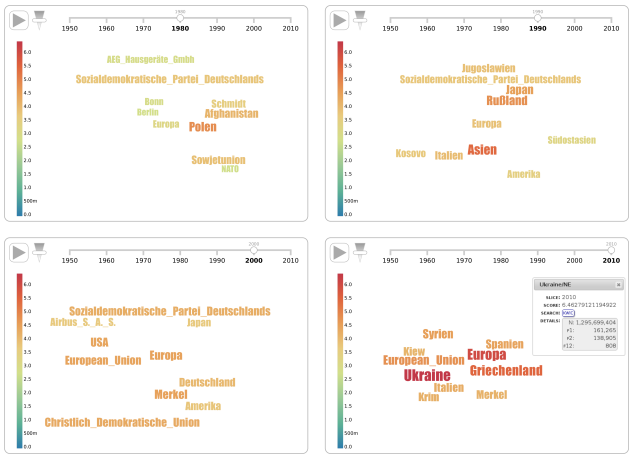


Abb. 2: *DiaCollo* dynamische Tag-Cloud Visualisierung der zehn besten Eigennamenkollokaten für "Krise" in der Wochenzeitung *DIE ZEIT* für die Epochen 1980-1989 (oben links), 1990–1999 (oben rechts), 2000–2009 (unten links) und 2010–2014 (unten rechts).

Fallstudien

Zeitgeschichte

Ein Anwendungsgebiet für *DiaCollo* stellt die (zeit-)historische Forschung dar. Insbesondere politische Prozesse können durch die zeitscheibenbasierte Analyse von DiaCollo auf neue Weise beschrieben werden. Das Potenzial von *DiaCollo* soll hier kurz am Beispiel des Begriffs "Krise" in der Wochenzeitung DIE ZEIT skizziert werden (Abbildung 2). Da der Begriff "Krise" eine inhärent instabile Situation bezeichnet, kann davon ausgegangen werden, dass die damit assoziierten Diskursumgebungen im zeitlichen Verlauf stark variieren. Dies sollte sich auch in den Kollokationsprofilen von *DiaCollo* niederschlagen. Mittels *DiaCollo* lassen sich Eigennamen (Personen, Orte, Institutionen) extrahieren, die als Kollokationspartner von "Krise" in der ZEIT auftreten. Eine Analyse der Zeitreihe dieser Kollokate zeigt, dass mittels *DiaCollo* korrekt politische Krisen (etwa in den 2000ern "CDU"), ökonomische Krisen (in den 1980ern "AEG", in den 1990ern die Finanzkrise in Südostasien), aber auch konflikthafte Krisen (z. B. Jugoslawien in den 1990ern) ermittelt werden können. Je nach Granularität der Abfrage werden auch die Ergebnisse komplexer und damit auch für den Experten interessanter. Ihre Interpretation bedarf dann in der Regel weiterer manueller Aufarbeitung, z.B. mithilfe der von DiaCollo bereitgestellten Verknüpfung mit der zugrundeliegenden Textbasis.

Lexikographie

Ein weiteres Anwendungsgebiet von *DiaCollo* ist die Lexikographie. Da Kollokationen und die Beschreibung des Bedeutungsspektrums (Lesarten) eines Wortes eng miteinander zusammenhängen, lassen sich aus zeitlichen Verläufen von Kollokationen wichtige lexikographische Befunde ableiten: die Verlagerung der Gewichtung von Lesarten untereinander oder das Verschwinden einer Lesart zugunsten einer anderen können dadurch ebenso nachverfolgt werden wie das Auftauchen von neuen Lesarten (Neosemanteme). Bekannte Beispiele hierfür sind Wörter wie 'Maus' (als Computermaus) oder 'Ampel' (in der politischen Bedeutung), die seit den späten 1980er bzw. den frühen 1990er Jahren im öffentlichen Sprachgebrauch sind. Ein komplexeres Beispiel stellt das Adjektiv 'autofrei' dar. Dieses ist im Duden definiert als "keinen Autoverkehr aufweisend". Eine genauere Sicht auf die Korpusbelege ergibt, dass das Wort zwei Unterbedeutungen aufweist: erstens die "(per Verordnung) auferlegte Autofreiheit", die durch die Öllkrise in den 1970er Jahren in der Kollokation 'autofreier Sonntag' erstmals auftrat und später in Verbindungen wie 'autofreie Innenstädte' ein Bedeutungserweiterung erfuhr. In den 1990er Jahren bildet sich die zweite Bedeutung des Wortes heraus, bei der der Verzicht auf das Auto auf Selbstverpflichtung beruht (vgl. http://zwei.dwds.de/wb/autofrei). Diese Lesart ist durch Kollokationen wie 'autofreie Wohnanlage' oder 'autofreie Siedlung' gekennzeichnet. Mit *DiaCollo* lassen sich beide Bedeutungen nicht nur unterscheiden, sondern auch in ihrem zeitlichen Verlauf nachverfolgen (Abbildung 3).

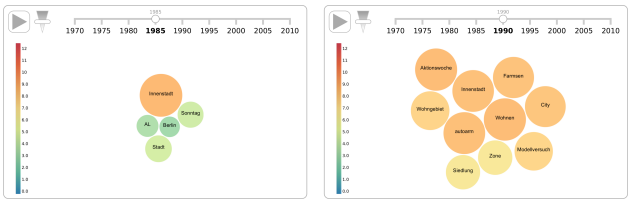


Abb. 3: *DiaCollo* dynamische Bubble-Chart Visualisierung der zehn besten Kollokaten des Ajdketivs autofrei im aggregierten DWDS Zeitungskorpus für die Epochen 1985–1989 (links) und 1990–1994 (rechts).

Zusammenfassung

Wir haben hier *DiaCollo* vorgestellt, ein neues Softwarewerkzeug für die effiziente Extraktion, den Vergleich und die interaktive Visualisierung von Kollokationen, speziell zugeschnitten auf die besonderen Anforderungen diachroner Textkorpora. Darüber hinaus haben wir anhand von zwei Fallstudien skizziert, wie *DiaCollo* als modularer Webservice-Plugin Forscher in den Geistes- und Sozialwissenschaften dabei unterstützen kann, ein klareres Bild der diachronen Variation in der Verwendung eines Wortes zu erhalten.

Bibliographie

Church, Kenneth Ward / Hanks, Patrick (1990): "Word association norms, mutual information, and lexicography", in: *Computational Linguistics* 16, 1: 22–29.

Davies, Mark (2012): "Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English", in: *Corpora* 7, 2: 121–157 http://davies-linguistics.byu.edu/ling450/davies_corpora_2011.pdf [letzter Zugriff 08. Februar 2016].

Didakowski, Jörg / Geyken, Alexander (2014): "From DWDS corpora to a German word profile – methodological problems and solutions", in Abel, Andrea / Lemnitzer, Lothar (eds.): *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern* (= OPAL 2014, 2). Mannheim: Institut für Deutsche Sprache 39-47 http://www.dwds.de/static/website/publications/pdf/didakowski_geyken_internetlexikografie_2012_final.pdf [letzter Zugriff 08. Februar 2016].

Evert, Stefan (2005): *The Statistics of Word Cooccurrences*. Word Pairs and Collocations. PhD, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf [letzter Zugriff 08. Februar 2016].

Geyken, Alexander / Haaf, Susanne / Jurish, Bryan / Schulz, Matthias / Steinmann, Jakob / Thomas, Christian / Wiegand, Frank (2011): "Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv", in Schomburg, Silke / Leggewie, Claus / Lobin, Henning / Puschmann, Cornelius (eds.): *Digitale Wissenschaft*. Stand und Entwicklung digital vernetzter Forschung in Deutschland. Köln: hbz 157– 161 http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wiss enschaft.pdf#page=159 [letzter Zugriff 08. Februar 2016].

Hamp, Birgit / Feldweg, Helmut (1997): "GermaNet – a lexical-semantic net for German", in: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* 9-15.

Henrich, Verena / Hinrichs, Erhard (2010): "GernEdiT – the GermaNet editing tool", in: *Proceedings of the ACL 2010 System Demonstrations* 19-24 http://www.aclweb.org/anthology/P10-4004 [letzter Zugriff 08. Februar 2016].

Jurish, Bryan / Thomas, Christian / Wiegand, Frank (2014): "Querying the Deutsches Textarchiv", in: *Proceedings of the Workshop Beyond Single-Shot Text Queries*. Bridging the Gap(s) between Research Communities (MindTheGap 2014) 25–30 http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf [letzter Zugriff 08. Februar 2016].

Kilgarriff, Adam / Tugwell, David (2002): "Sketching words", in: Corréard, Marie.-Hélène (ed.): *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. United Kingdom: EURALEX 125–137 http://www.kilgarriff.co.uk/Publications/ 2002-KilgTugwell-AtkinsFest.pdf [letzter Zugriff 08. Februar 2016].

Rychlý , Pavel (2008): "A lexicographer-friendly association score", in: *Proceedings of Recent Advances in Slavonic Natural Language Processing*. RASLAN, 6–9 http://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf [letzter Zugriff 08. Februar 2016].

Scharloth, Joachim / Eugster, David / Bubenhofer, Noah (2013): "Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn", in: Busse, Dietrich / Teubert, Wolfgang (eds.): *Linguistische Diskursanalyse*. Neue Perspektiven. Wiesbaden: VS Verlag 345–380 http://www.scharloth.com/files/Rhizom_Zeit.pdf [letzter Zugriff 08. Februar 2016].

Sokirko, Alexey (2003): "A technical overview of DWDS/Dialing Concordance", Vortrag beim Workshop *Computational linguistics and intellectual technologies*. Protvino. http://www.aot.ru/docs/OverviewOfConcordance.htm [letzter Zugriff 08. Februar 2016].

Bearbeitung und Annotation historischer Texte mittels Graph-Datenbanken am Beispiel der Chronik des Matthias von Edessa

Kaufmann, Sascha
sascha.kaufmann@dh.unibe.ch
Universität Bern, Schweiz

Andrews, Tara Lee
tara.andrews@kps.unibe.ch
Universität Bern, Schweiz

Die Chronik des Matthias von Edessa ist den meisten Wissenschaftlern des mittelalterlichen Nahen Ostens und des ersten Kreuzzuges bekannt für ihren Reichtum an Informationen, sowie der (mutmaßlichen) Ignoranz und Naivität ihres Autors. Matthias von Edessa war ein Armenischer Priester, der in der Kreuzfahrer-Grafschaft Edessa lebte und die Chronik zwischen den Jahren 1110 und 1132 verfasste. Gleichwohl der Text oft von Historikern verwendet wird, liegt er bis heute in keiner kritischen Edition vor und wurde zuletzt 1898 (Matthias von Edessa 1898) veröffentlicht. Die Chronik umfasst 35 (kopierte) Manuskripte, deren ältestes auf mindestens 450 Jahre nach dem Tod des Autors datiert werden kann. Diese werden zur Zeit für eine digitale Gesamtedition vorbereitet.

Die Herausforderungen, die sich dabei stellen, beschränken sich nicht nur auf die Bearbeitung aus philologischer Sicht, sondern auch auf die Annotation und Präsentation als historisches Werk, mit dem Ziel, sie auch als Plattform für Historiker zur Verfügung zu stellen (z. B. mit Zeit- und Ortsangaben, etc.). Hierzu greifen wir zum Einen auf eine Reihe aktueller Methoden und Werkzeuge der digitalen Philologie zurück, wie zum Beispiel die Transkription aller Manuskripte, palaeographisches Markup unter der Benutzung des TEI-Vokabulars (TEI Consortium 2015), automatische Manuskript-Kollation mit CollateX (Dekker et al. 2014), stemmatische Analyse mit Hilfe der Werkzeuge von Stemmaweb (Andrews / Macé 2013) und der Publikation aller Transkriptionen, sowie einer editorischen Rekonstruktion des Textes. Zum Anderen werden auch Textkommentare von der digitalen Plattform profitieren, da sie mit weiteren Informationen angereichert werden können. So können unter Anderem Ortsnamen nicht nur ausgezeichnet („getagged“) werden, sondern ihre mögliche Lokalisierung auch angegeben und soweit möglich geographisch angezeigt werden. Personen und ethnographische Bezeichnungen werden nicht nur in einem Index erfasst, sie werden, soweit möglich, mit prosopographischen Datenbanken oder relevanten Seiten auf Wikipedia verlinkt.

Das derzeitige Projekt *The Chronicle of Matthew of Edessa Online* wird bis 2018 von dem Schweizer National Fond (SNF) finanziert und baut auf verschiedenen Projekten und gesammelten Erfahrungen der vergangenen fünf Jahre auf.

Ziel ist es, dem Forscher oder der Forscherin ein Werkzeug in die Hand zu geben, das es ihm / ihr erlaubt Bewegungen von Individuen und Gruppen über Raum und Zeit zu verfolgen, indem wir die Textedition selbst zu einer Plattform für mittelalterliche Geschichte machen.

Im Folgenden werden wir zwei, aus technischer Sicht wesentliche Aspekte des Projektes vorstellen (Speichern und Wiedergabe von Informationen) und erste Ergebnisse präsentieren.

Effizientes Speichern und Bearbeiten von Manuskripten und Annotationen

Aufbauend auf den Erkenntnissen, die wir aus dem Stemmaweb-Projekt gewonnen haben, werden die Manuskripte in Form von Graphen gespeichert [1] . Daher ist für uns der naheliegenste Schritt, die gesamte Arbeit des Benutzers im Graph zu repräsentieren. Als mögliche Darstellungsformen für den Benutzer kommen hierfür Online-Präsentationen oder ein Export, zum Beispiel nach TEI, in Betracht.

Während die Speicherung in Stemmaweb als Perl-Objekte in einer Relationalen-Datenbank (MySQL (Oracle Cooperation 2015)) erfolgt, haben wir uns für dieses Projekt dafür entschieden, Neo4J (Neo Technology Inc 2015), eine Graph-Datenbank zu verwenden.
Zum einen kommt dies unserer intern verwendeten Datenstruktur entgegen, die auf einer Modellierung auf Graphen basiert. Damit verbunden hat dies auch den Vorteil, dass viele, für uns wichtige Funktionen (z. B. Depth First Search, Breadth First Search, etc.), nativ vom Datenbanksystem zur Verfügung gestellt werden und speziell benötigte Operationen, wie z. B. Plausibilitätsprüfungen, einfacher implementiert und effizienter ausgeführt werden können.

Des Weiteren migrieren wir derzeit große Teile, des aus Stemmaweb vorliegenden Quellcodes, von Perl nach Java. Gleichzeitig überarbeiten wir die API, mit dem Ziel, den Service auch über eine Web-API nutzen zu können. Die Hauptgründe hierfür sind, die bestehende Stemmaweb Backend-Engine zu modernisieren und effizienter zu gestalten. Dieser Schritt soll bis Ende 2015 abgeschlossen sein.

Parallel zu der Migration, haben wir bereits einen ersten Prototypen des Editions-Interfaces erstellt. Dieser Prototyp basiert bereits auf Neo4J und implementiert ein

Textmodel das interoperabel mit Stemmaweb ist und wir zu Testzwecken bereits mit zusätzlichen Annotationen versehen haben.

In Abbildung 1 ist ein typischer Ausschnitt unserer Neo4J Datenbankstruktur zu sehen. Man erkennt, dass sich eine „TRADITION“ aus ein oder mehreren „SECTION“-Knoten zusammensetzt, die wiederum mittels gerichteter „NEXT“-Kanten miteinander verbunden sind und somit ihre Reihenfolge gewährleistet bleibt [2] . Jeder „SECTION“-Knoten verweist wiederum auf eine Folge von untereinander mit „LEMMA_TEXT“ verbundenen „READING“-Knoten. Darüber hinaus existieren noch sogenannte „TRANSLATION“-Knoten, die die jeweilige Übersetzung in einer Sequenz aus READING-Knoten speichern. Hierbei bleibt es dem Bearbeiter freigestellt, die Granularität der Übersetzung festzulegen. So sind Einheiten, wie zum Beispiel Sigle-Wörter, Sätze oder ganze Paragraphen, etc. vorstellbar.

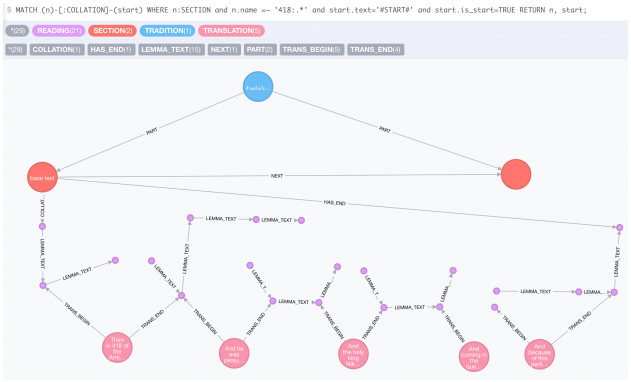


Abb. 1 : Beispiel für die derzeitige Datenbankstruktur in Neo4J. Gut zu Erkennen ist die Gliederung in Tradition (blau), Sections (orange), Readings (pink) und Translations (rosa).

Eine adäquate Wiedergabe der gespeicherten Informationen

Wie bereits erwähnt, liegt ein weiterer Schwerpunkt des Projektes auf einer angemessenen Wiedergabe der im System vorhandenen Informationen. Aus diesem Grund haben wir, parallel zu den Arbeiten am Backend, bereits damit begonnen, den Prototypen einer Webseite zur Darstellung der Informationen zu implementieren.

Dieser Prototyp stellt zur Zeit vier Abschnitte (Sections) aus der Chronik bereit, zwischen denen der Benutzer auswählen kann (Abbildung 2). Zu einem ausgewählten Abschnitt werden dessen Transkription, sowie (englische) Übersetzung angezeigt. Eine Karte, in Form einer eingebetteten Google-Map (Google Inc. 2012), zeigt zusätzlich alle im Abschnitt vorkommenden Örtlichkeiten an, sofern sie lokalisierbar sind. In der Transkription werden Textteile, zu denen Zusatzinformationen vorliegen, dem Benutzer farblich kodiert angezeigt (Ortsangaben (blau), Personenangaben (rot) und Zeitangaben (gelb)).

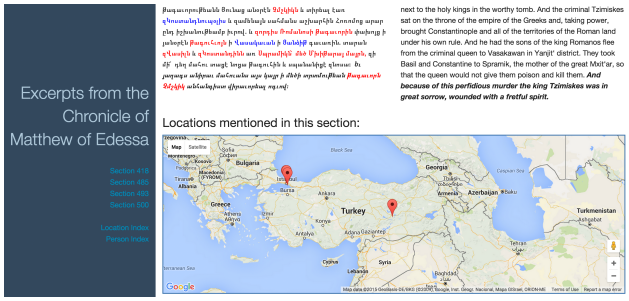


Abb. 2 : Screenshot der Webseite. Neben der Transkription (linke Spalte), befinden sich die englische Übersetzung (rechte Spalte), sowie ein Karte mit Positionsmarkierungen der im Text vorkommenden Ortsbeschreibungen.

Fährt nun der Benutzer mit dem Mauszeiger über einen solchen farblich hervorgehobenen Textteil, werden zu diesem in einem Popup-Fenster weitere Informationen oder Verlinkungen angezeigt. So ist zum Beispiel für Ortsangaben, wie in Abbildung 3 zu sehen, dies üblicherweise der Ortsname, sowie ein vergrößerter Ausschnitt der Karte mit einer Markierung der genauen Position, soweit diese bekannt ist.

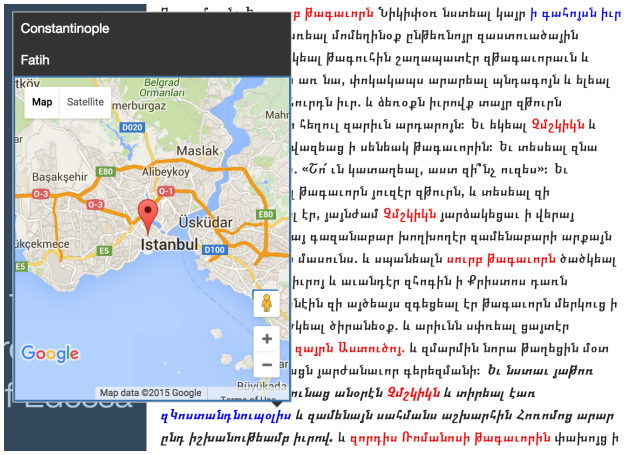


Abb. 3: Beispiel für Zusatzinformationen bezüglich Constantinople (Faith).

Bei Personenangaben, siehe Abbildung 4, wird die Übersetzung des Namens, der volle Name und weitere Informationen, sowie Links, die zu Einträgen bezüglich der Person auf externen Seiten verweisen, hier Links zu *Wikipedia* (Wikimedia Foundation 2015) und *Prosopography of the Byzantine World* (Jeffreys et al. 2011), in einem Popup-Fenster angezeigt.

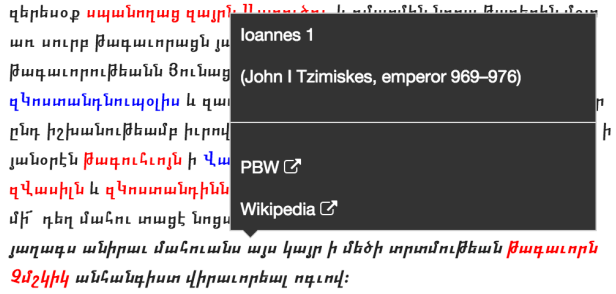


Abb. 4 : Beispiel für Zusatzinformationen bezüglich Ioannes 1 (John I Tzimiskes).

Darüber hinaus arbeiten wir daran, sowohl individuelle Textzeugen als auch verschiedene Variationen eines Textes anzeigen zu können.

Hervorzuheben ist, wie oben bereits angedeutet, das die Plattform auch die Möglichkeit bietet, die Edition zu exportieren. Hierzu werden relevante Standards wie zum Beispiel TEI oder CIDOC-CRM (International Council of Museums 2014) angeboten.

Zusammenfassung und Ausblick

Wir haben einen kurzen Einblick in das Projekt The Chronicle of Matthew of Edessa Online gegeben. Dieses hat als Ziel eine digitale Plattform zur Verfügung zu stellen, welche das Untersuchen und Bearbeiten von Manuskripten aus philologischer, als auch historischer Sichtweise unterstützt. Hierzu haben wir einen Einblick in technische Aspekte und dem derzeitigen Stand der Implementierung gegeben. Zusammendfassend kann man sagen, dass wir schon eine gute Strecke zurückgelegt haben und dass das Projekt noch einige interessante Aufgaben für uns bereit hält [3] .

Notes

1. Stemmaweb bietet seinen Benutzern die Möglichkeit mit Text-Kollationen und Varianten zur stemmatischen Analyse zu arbeiten und diese zu modifizieren. Des Weiteren können Texte basierend auf Kollationen und Stemmata rekonstruiert und erstellt werden. Hierfür hat es sich bewährt, Texte in einem Graph abzubilden.
2. Dieser Ansatz wurde gewählt, da eine automatische Kollation für längere Texte am besten mit einer Einteilung in diskrete Abschnitte (Sections) gelingt. Desweiteren können wir somit Textzeugen, deren Abschnitte abweichend angeordnet sind, ohne Schwierigkeiten darstellen
3. Interessante Aufgaben wären zum Beispiel die Gestaltung der Web-Oberfläche, die Interaktion mit dem Benutzer, Optimierung der Graph-Datenbank, Web-API oder die Bereitstellung verschiedener Exportformate, um nur einige zu nennen.

Bibliographie

Andrews, Tara L. / Macé, Caroline (2013): „Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata“, in: *Literary and Linguistic Computing* 28, 4: 504-21 http://dx.doi.org/10.1093/llc/fqu072 [letzter Zugriff 09. Februar 2016].

Dekker, Ronald Haentjens / Hulle, Dirk van / Middell, Gregor / Neyt, Vincent / Zundert, Joris van (2014): „Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project“, in *Literary and Linguistic Computing* 25: 1-19 http://dx.doi.org/10.1093/llc/fqu007 [letzter Zugriff 09. Februar 2016].

Google Inc (2012): *Google Maps* http://www.google.com/maps [letzter Zugriff 15. Oktober 2015].

International Council of Museums (2014): *The CIDOC Conceptual Reference Model* http://www.cidoc-crm.org [letzter Zugriff 15. Oktober 2015].

Jeffreys, Michael (2011): *Prosopography of the Byzantine World* www.pbw.kcl.ac.uk [letzter Zugriff 15. Oktober 2015].

Matthias von Edessa (Matt#e#os Ur#hayec#i) (1898): *Die Croniken des Matthias von Edessa Žamanakagrut#iwn*. Vałaršapat.

Neo Technology Inc. (2015): *Neo4j* http://www.neo4j.com [letzter Zugriff 15. Oktober 2015].

Oracle Cooperation (2015): *MySQL* https://www.mysql.com [letzter Zugriff 15. Oktober 2015].

TEI Consortium (2015): *Guidelines for Electronic Text Encoding and Interchange. Version 2.8.0* http://www.tei-c.org/Vault/P5/2.8.0/doc/tei-p5-doc/en/html/ [letzter Zugriff 15. Oktober 2015].

Wikimedia Foundation (2015): *Wikipedia* http://www.wikipedia.org [letzter Zugriff 15. Oktober 2015].

Knowledge-Based Support for Scholarly Editing and Text Processing

Kittelmann, Jana
info@janakittelmann.de
MLU Halle-Wittenberg, Deutschland

Wernhard, Christoph
info@christophwernhard.com
TU Dresden, Deutschland

Introduction

Background: Large Knowledge Bases

A large portion of the material on which scholarly editing is based today is available electronically in large knowledge bases. Some of these emerge from the archive, library and museum communities, for example *Kalliope*. Such efforts require the use of standardized vocabularies and databases of entities such as persons and locations. *Kalliope* thus links to *Gemeinsame Normdatei (GND)*, which provides more than 120 million facts about approximately 11 million entities. The prevailing technique to realize such linked knowledge bases is the Semantic Web, as advocated by the W3C, characterized by the use of ontologies to express standardized vocabularies, global identifiers (URIs) and the possibility to express knowledge in a machine understandable way as subject-predicate-object statements with RDF. Further large knowledge bases, such as *Yago* (Hoffart et al. 2013) and *DBpedia* (Lehmann et al. 2015), developed mainly in computer science with Semantic Web techniques, gather and combine machine processable knowledge from "crowd-maintained" sources like *Wikipedia* and centrally maintained sources like *GND* or *GeoNames*.

Beyond *TEI*

The seemingly best developed machine support for scholarly editing today is provided with the *Text Encoding Initiative (TEI)* format, based on document markup. URIs as attribute values of markup elements can provide links to knowledge bases. Envisaged applications include in particular the rendering for different media and extraction of metadata. Some of the recent developments are actually orthogonal to the OCHCO text model and its representation through XML, core characteristics of the original *TEI*. Connecting *TEI* with Semantic Web techniques, data modeling and ontologies is, for example, an ongoing topic of discussion (e.g. Eide 2015). Recent versions of *TEI* provide support for *names, dates, people, and places* as well as *linking, segmentation, and alignment* (The TEI Consortium 2015: Chapters 13 and 16). In a broad long-term perspective, important aspects that further go into these directions become apparent:

- Incorporation of advanced semantics related techniques such as named entity recognition or statistics-based text analysis.
- Relationships to external knowledge bases and to formal semantics.
- Obtaining high-quality presentations without requiring expensive development of dedicated XML transformations and stylesheets.
- Loose coupling of object text and markup: Alternate markup by different authors or for different purposes should be supported. Markup generated by automated methods should not clutter up the document. Queries and transformations should remain applicable also after changes of the markup. Sustainability must not be compromised by dependency on short-lived technology and specifications.

Addressing these issues, we approach the requirements of today's scholarly editing here from the view of computational logic: What can logics – as machine processable symbolic languages with formally specified semantics – contribute? A starting point is that with Semantic Web technology the large knowledge bases can already be considered as large sets of logic facts. Logic languages have various further potential roles in machine supported scholarly editing, such as specifying properties and values associated with texts, specifying pieces of text, specifying knowledge sources and their combination, and specifying inferences involved in automated computation of information associated with texts.

Knowledge-Based Support for Scholarly Editing

High-Quality Support at all Phases

Three main phases of machine assisted scholarly editing can be identified, which all should be supported: (1) Creating the enhanced object text; (2) Generating intermediate representations for inspection by humans or machines; (3) Generating consumable presentations. Support for all three phases should be of high quality – for example entity recognition should precisely identify persons, or the print layout of a finally rendered document should be professional.

Issues of Integrating Different Types of Knowledge

High-quality support is not possible without inclusion of specialized techniques and the combination of automated techniques with information and adjustments provided by humans. The adequate support of this combination is an important aspect where the considered scenario differs from conventional programming or query languages. Relevant techniques include non-monotonic reasoning, semantics-based knowledge partitioning (Wernhard 2004, Ghilardi et al. 2006, Cuenca Grau et al. 2008, Kontchakov et al. 2010) and the use of explanations for inferred information, as exemplified by proofs in mathematical knowledge bases (Urban et al. 2013). A further important integration requirement concerns the combination of statistics-based techniques, which are

essential for natural language processing operations such as named entity recognition or keyphrase extraction, with a symbolic logic-based framework.

External Annotations

The availability of powerful techniques to identify places in text – based on syntactic as well as semantic properties – suggests to prefer external annotations to in-place markup. Annotations are then maintained separated from the object text in annotation documents. An automated processor creates an annotated document by merging annotations and object text.

Representation of Epistemic Status

Scholarly editing requires to associate various forms of epistemic status with facts, which is interesting to model formally from the viewpoint of artificial intelligence. Consider for example a creation date associated with written communication: it can be given by its author or can be inferred – by the editor or by a machine, it can be only partially specified by the author, it can be specified with different precision, considered as a point or range in time, etc. The current version of *TEI* offers some related elements to indicate certainty, precision and responsibility (The TEI Consortium 2015: Chapter 21), but these are not based on any formal semantic treatment and it is seems hardly possible to express the sketched date examples with them.

Utilizing Inferred Access Patterns

Efficient access to large knowledge bases requires caching and preprocessing, which ideally should be performed automatically on the basis of the queries performed by the knowledge processing engine. Relevant techniques come from optimization in databases (Toman / Weddell 2011) and in first-order model computation systems (Pelzer / Wernhard 2007). It seems that recent techniques for view-based query processing (Calvanese et al. 2007) based on variants of Craig's interpolation and second-order quantifier elimination (Toman / Weddell 2011; Bárány et al. 2013; Wernhard 2014) where access patterns can be specifically considered in an abstract way (Bárány et al. 2013) are particularly useful. Logic-based languages for programming as well as data access facilitate the application of such abstract techniques. For an overview on alternate ways to associate computational meaning with logics see (Kowalski 2014).

The Role of Ontologies

Ontologies are an important ingredient for the Semantic Web because they provide agreed vocabularies. However, to evaluate queries arising in the text processing tasks of scholarly editing, ontology reasoning alone is not sufficient. Also, the basic ontologies relevant in the context of scholarly editing are – in contrast to the biomedical area (Horrocks 2013) – rather small and trivial.

A Prototype: The *KBSET* System

Important issues of complex computer systems often become apparent only with applications. Thus, the authors developed the *KBSET* system, an experimental platform to clarify the precise requirements of machine support for scholarly editing and to experiment with advanced techniques. It follows the outlined approach, but, so far, only realizes some of the discussed aspects. A draft version of an edition of *Max Stirner: Geschichte der Reaction, Band 1. Berlin, 1852* accompanies it as comprehensive example. The system is free software and available from http://cs.christophwernhard.com/kbset/.

In a typical setting, the system takes as inputs:

A source text file, possibly in *LaTeX* format. The system can parse *LaTeX*, where the set of recognized commands is configurable, including user defined commands as well as commands that establish some "ordered hierarchy of content objects". In this way plain or structured text is available within the system to modules that operate on such text models.

Annotation documents, that is, text files with annotations, possibly in *LaTeX* format. The associated places in the source text to which they are referring are specified abstractly.

Large fact bases, currently in particular *GND* and *GeoNames,* as well as extracts from *YAGO2* and *DBpedia*.

A so-called *assistance document*, that is, a configuration file, where, among other things, the fact bases are specified and information is given to bias or override automated inferencing such that fully correct results are obtained.

A user interface is provided that integrates the system into the *Emacs* editor, which is free software. The system includes a facility for named entity recognition, which – essentially based on *GND* and *GeoNames* as gazetteers – identifies persons, locations and dates. The system produces a variety of outputs, supporting all the phases of scholarly editing mentioned above:

- *LaTeX* documents where annotations and inferred information are merged in. By passing unrestricted *LaTeX* access to the user, high-quality layouts can be achieved.

- Support during development by possibilities to highlight and inspect entities recognized by the system.
- An export possibility to visualize detected locations mentioned in the source text with the *Dariah* geobrowser.

A typical application would be the development of an annotated essay or book, where the source text is edited in *LaTeX* and the configuration evolves step-by-step until the inferred information is fully correct.

Acknowledgments

This work was supported by *Alexander von Humboldt-Professur für neuzeitliche Schriftkultur und europäischen Wissenstransfer* and by *DFG grant WE 5641/1-1* .

Bibliographie

Bárány, Vince / Benedikt, Michael / ten Cate, Balder (2013): "Rewriting guarded negation queries", in: *Mathematical Foundations of Computer Science 2013 (MFCS 2013)*, volume 8087 of LNCS. Berlin / Heidelberg / New York: Springer 89-110.

Calvanese, Diego / De Giacomo, Giuseppe / Lenzerini, Maurizio / Vardi, Moshe Y. (2007): "View-based query processing: On the relationship between rewriting, answering and losslessness", in: *Theoretical Computer Science* 371, 3: 169-182.

Cuenca Grau, Bernardo / Horrocks, Ian / Kazakov, Yevgeny / Sattler, Ulrike (2008): "Modular reuse of ontologies: Theory and practice", in: *Journal of Artificial Intelligence Research* 31: 273-318.

Eide, Øyvind (2015): "Ontologies, data modeling, and TEI", in: *Journal of the Text Encoding Initiative* 8.

Ghilardi, Silvio / Lutz, Carsten / Wolter, Frank (2006): "Did I damage my ontology? A case for conservative extensions in description logics", in: Doherty, Patrick / Mylopoulos, John / Welty, Christopher A. (eds.): *Proc. 10th Int. Conf. on Principles of Knowledge Representation (KR'06)*. Cambridge, MA: AAAI Press 187-197.

Hoffart, Johannes / Suchanek, Fabian M. / Berberich, Klaus / Weikum, Gerhard (2013): "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia", in: *Artificial Intelligence* 194: 28-61.

Horrocks, Ian (2013): "What are ontologies good for?", in: Kuppers, Bernd Olaf / Hahn, Udo / Artmann, Stefan (eds.): *Evolution of Semantic Systems*. Berlin / Heidelberg / New York: Springer 175-188.

Kontchakov, Roman / Wolter, Frank / Zakharyaschev, Michael (2010): "Logic-based ontology comparison and module extraction, with an application to DL-Lite", in: *Artificial Intelligence* 174, 15: 1093-1141.

Kowalski, Robert A. (2014): "Logic Programming", in: Siekmann, Jörg (ed.): *Computational Logic* (= Handbook of the History of Logic 9). Amsterdam: Elsevier 523-569.

Lehmann, Jens / Isele, Robert / Jakob, Max / Jentzsch, Anja / Kontokostas, Dimitris / Mendes N., Pablo / Hellmann, Sebastian / Morsey, Mohamed / van Kleef, Patrick / Auer, Sören / Bizer, Christian (2015): "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia", in: *Semantic Web* 6, 2: 167-195.

Pelzer, Björn / Wernhard, Christoph (2007): "System description: E-KRHyper", in: *Automated Deduction* (CADE-21), volume 4603 of LNCS (LNAI). Berlin / Heidelberg / New York: Springer 503-513.

The TEI Consortium (2015): *TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 2.8.0* TEI Consortium http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 9. Oktober 2015].

Toman, David / Weddell, Grant (2011): *Fundamentals of Physical Design and Query Compilation San Rafael*. CA: Morgan and Claypool.

Urban, Josef / Rudnicki, Piotr / Sutcliffe, Geoff (2013): "ATP and presentation service for Mizar formalizations", in: *Journal of Automated Reasoning* 50 (2): 229-241.

Wernhard, Christoph (2004): "Semantic knowledge partitioning", in: *Logics in Artificial Intelligence*: 9th European Conf. (JELIA 04), volume 3229 of LNCS (LNAI). Berlin / Heidelberg / New York: Springer 552-564.

Wernhard, Christoph (2014): *Expressing view-based query processing and related approaches with second-order operators*", Technical Report - Knowledge Representation and Reasoning 14-02, TU Dresden, http://www.wv.inf.tu-dresden.de/Publications/2014/report-2014-02.pdf [letzter Zugriff 9. Oktober 2015].

Attribuierung direkter Reden in deutschen Romanen des 18.-20. Jahrhunderts. Methoden zur Bestimmung des Sprechers und des Angesprochenen

Krug, Markus
markus.krug@uni-wuerzburg.de
Universität Würzburg, Deutschland

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reger, Isabella
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Macharowsky, Luisa
luisa.macharowsky@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

Weimer, Lukas
lukas.weimer@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

Puppe, Frank
frank.puppe@uni-wuerzburg.de
Universität Würzburg, Deutschland

Problembeschreibung

Im Folgenden wird ein Verfahren vorgestellt, das die automatische Zuordnung von direkter Rede in Erzähltexten sowohl zur sprechenden als auch zur angesprochenen Figur ermöglicht. Kann man eine solche automatische Zuordnung vornehmen, ermöglicht dies die Extraktion eines sozialen Netzwerks aus einem Text, wobei die Figuren als Knoten und die direkte Rede als Kanten modelliert werden (Elson / Dames 2010), aber sie kann auch eine wichtige Informationsquelle für andere analytische Schritte sein, z.B. zur Verbesserung der Koreferenzresolution oder zur Analyse der Quelle der Zuschreibung von Figurenattributen.

Stand der Forschung

Eine der ersten Arbeiten auf diesem Gebiet ist das System ESPER (Zhang et al. 2003), das direkte Reden innerhalb von Kindergeschichten erkennen soll. Das System extrahiert zunächst die direkten Reden im Text und klassifiziert diese mit einem Entscheidungsbaum in zwei Kategorien, Sprecherwechsel bzw. kein Sprecherwechsel. Evaluiert werden die Ergebnisse mit zwei manuell annotierten, sehr unterschiedlichen Geschichten. Sie berichten eine Genauigkeit, gemeint ist hier die Anzahl der korrekt bestimmten Sprecher für alle direkten Reden, von 47.6% und 86.7%. Glass und Bangay (2006), ebenfalls regelbasiert, bestimmen zunächst für eine direkte Rede das Kommunikationsverb und anschließend eine Menge von Akteuren, woraus letztendlich der Sprecher bestimmt wird. Sie evaluieren ihre Techniken auf 13 englischsprachigen fiktionalen Werken und berichten eine Genauigkeit von 79.4% (Glass / Bangay 2007). Iosif und Mishra (2014) folgen im Prinzip dem Schema von Glass und Bangay (2007), ergänzen es aber durch eine aufwendigere Vorverarbeitung einschließlich Koreferenzresolution. Sie erreichen eine Genauigkeit von ca 84.5% und zählen damit zu den besten bisher veröffentlichten Ergebnissen. Ruppenhofer und andere (Ruppenhofer et al. 2010) berichten einen F-Score von 79% in der Zuordnung von Politikern zu ihren Aussagen in deutschsprachigen Kabinettsprotokollen aus den Jahren 1949-1960.

Neben diesen regelbasierten Ansätzen werden auch maschinelle Lernverfahren eingesetzt. Zu den ersten erfolgreichen Systemen zählt das von Elson und McKeown (2010). Ihre Daten für die Sprecherzuordnung ließen sie über Amazons *Mechanical Turk* System bearbeiten. Ihr System klassifiziert zunächst regelbasiert eine direkte Rede in eine von fünf syntaktischen Kategorien. Für jede dieser Kategorien wurden anschließend eigenständige maschinelle Lernverfahren trainiert. Insgesamt erreichen sie eine Genauigkeit von etwa 83%, ausgewertet anhand von englischen Romanen des 19 Jahrhunderts. O'Keefe und andere (O'Keefe et al. 2012), die an Elson und McKeowns Ansatz die Erstellung des Goldstandards und auch die praxisferne Verwendung von Informationen aus dem Goldstandard kritisieren, betrachten die Zuordnung als Sequenzproblem. Sie nutzen die Klassifikationsangaben von vorhergehenden direkten Reden als Features für die gesamte Sequenz. In ihrer Evaluation vergleichen Sie drei Verfahren mit einer sehr einfachen regelbasierten Baseline. Ihre Ergebnisse bei der Anwendung des Systems auf zwei Zeitungskorpora - Wall Street Journal und Sydney Morning Herald - sowie die Sammlung literarischer Texte aus der Arbeit von Elson und McKeown zeigen einen großen Unterschied zwischen den Domänen. Sie erreichen auf den beiden Zeitungskorpora 84.1% (WSJ) bzw. 91.7% (SMH) Genauigkeit. Auf dem literarischen Korpus erreichen sie dagegen lediglich eine maximale Genauigkeit von 49%. (He et al. 2013) erreichen mit einem auf Ranking basierten maschinellem Lernverfahren unter der Ausnutzung von Features des Actor-Topic Modells (Celikyilmaz et al. 2010) auf dem Elson und McKeown-Korpus eine Genauigkeit zwischen 74.8% und 80.3%. Almeida und andere gehen von einer engen Verflechtung von Koreferenzresolution und Sprecherattribution aus und integrieren dabei beide Verfahren in ihrem Ansatz; die Ergebnisse der beiden einzelnen Lernverfahren werden in einem dritten Schritt verbunden. Sie erreichen damit 88.1% Genauigkeit (Almeida et al. 2014). Neuere Versuche mit Deep Learning-Verfahren aufgrund der Sprache der Figuren haben nur Genauigkeiten von unter 50% erreicht (Chaganty / Muzny 2014).

Die Zuordnung einer angesprochenen Figur wurde unserer Wissens noch in keiner anderen Arbeit untersucht.

Daten und Annotation

Für diese Arbeit verwenden wir Abschnitte des frei zugänglichen Korpus DROC. DROC besteht aus 89 Romanausschnitten, jeweils 130 Sätze lang, in denen alle Figurenreferenzen (mit und ohne Namen) und Koreferenzen annotiert sind. Aus dem Korpus wurden 77 Ausschnitte ausgewählt und mit einem eigens entwickelten Tool alle direkten Reden sowie die zugehörigen Sprecher und angesprochenen Figuren eingetragen. Jeder Text wurde von einem Annotator bearbeitet; eine zweite Annotation ist vorgesehen. Insgesamt wurden so 2264 direkte Reden mit Sprecher und Angesprochenen annotiert. Für die in Abschnitt 5 diskutierten Experimente wurde das Korpus in drei zufällige Mengen aufgeteilt:

Korpus	# Tokens	# Direkte Reden	# Romanfragmente
Trainingskorpus	107141	1185	37
Entwicklungskorpus	58709	615	20
Testkorpus	64330	464	20

Tab. 1: Überblick über die Auftrennung des in dieser Arbeit verwendeten Korpus.

Methoden

Wir verwenden regelbasierte Verfahren und maschinelle Lernverfahren, aber anders als in (He et al. 2013) oder (O'Keefe et al. 2012) dienen erstere nicht nur als Baseline-Verfahren, sondern wurden soweit wie möglich optimiert.

Wir verwenden die Techniken 2-Way Klassifikation und N-Way Klassifikation wie in (O'Keefe et al. 2012) vorgeschlagen. Zusätzlich evaluieren wir MaxEnt2WayToMatch, bei dem Kandidaten nur bis zum ersten tatsächlichen Sprecherkandidaten erzeugt werden.

Für die Sprecherzuordnung und Zuordnung eines Angesprochenen sind die in dieser Arbeit verwendeten Features in Tabelle A1 im Anhang zusammengefasst.

Für diese Aufgabe haben sich regelbasierte Verfahren als konkurrenzfähig mit den aktuellen ML-Verfahren erwiesen. Sie besitzen außerden den Vorteil, dass sie nicht so viele Trainingsbeispiele benötigen. Die Grundstruktur des Algorithmus ist der Idee des regelbasierten Koreferenzsystems von Stanford (Lee et al. 2011) angelehnt. Es werden eine Reihe von Regelpässen nacheinander ausgeführt. Die Regelpässe sind gemäß ihrer *Precision* geordnet, d. h. Regeln mit einer hohen *Precision* werden zuerst ausgeführt. Eine spätere Regel kann eine Entscheidung einer früheren Regel nicht revidieren. Tabelle A2 im Anhang zeigt die in dieser Arbeit verwendeten Regelpässe.

Mit Hilfe der Trainingsdaten konnte eine optimale Reihenfolge der Ausführung der Regeln empirisch ermittelt werden, bei der einige Regeln auch mehrfach angewendet werden.

(1)→(2)→(3)→(4)→(5)→(6)→(7)→(5)→(6)→(8)→(9)→(5)→(6)→(7)→(10).

Evaluation

Die Parameter für die ML-Verfahren wurden auf dem Development-Anteil der Daten optimiert und anschließend gegen die Testmenge evaluiert. Für die regelbasierten Verfahren gibt es keine Unterscheidung zwischen Trainings- und Development-Korpus. Ein Sprecher gilt als korrekt bestimmt, wenn sich der vom System bestimmte Kandidat in der selben Koreferenzkette befindet, wie die Entität, die von unserem Annotator als korrekt markiert wurde. Tabelle 2 beschreibt die Ergebnisse bei der Anwendung der Verfahren auf das Testkorpus.

Verfahren	Sprecher Precision	Sprecher Recall	Sprecher Genauigkeit	Angesprochener-Precision	Angesprochener Recall	Angesprochener Genauigkeit
MaxEnt2Way	72.4%	86.0%	62.3%	37.3%	86.2%	32.2%
MaxEntNWay	50.3%	97.6%	49.1%	38.5%	94.0%	36.2%
MaxEnt2WayToMatch	67.1%	81.3%	54.6%	32.3%	57.3%	18.5%
SVM2Way	58.8%	88.4%	52.0%	39.4%	37.1%	14.6%
CRFNWay	50.0%	96.6%	48.3%	35.8%	92.7%	33.2%
Regelbasiert (vorläufig)	**79.6%**	**98.5%**	**78.4%**	**63.0%**	**94.2%**	**59.3%**

Tab. 2: Ergebnisse der einzelnen Verfahren auf dem Testkorpus, bestehend aus 20 zufällig gewählten Romanfragmenten.

Unsere Experimente bestätigen die Aussagen von O'Keefe (O'Keefe et al. 2012), dass 2Way ML-Verfahren bessere Ergebnisse in der Sprechererkennung liefern, als korrespondierende NWay Verfahren. Analoges gilt für die Evaluation der CRFs, die sogar beinahe den selben Wert für die Sprechererkennung liefern wie in (O'Keefe et al. 2012). Sowohl auf dem Developmentkorpus, als auch auf dem Testkorpus zeigen regelbasierte Ansätze deutliche Vorteile gegenüber den in dieser Arbeit verwendeten ML-Verfahren. Es ist weiterhin ersichtlich, dass die Bestimmung des Sprechers einfacher ist, als die Bestimmung des Angesprochenen. Wahrscheinlich liegt das daran, dass im Fall der Sprecherzuschreibung mehr Information vorliegt, nämlich die direkte Rede und der Kontext, während bei der Ermittlung des Angesprochenen die direkte Rede selbst nur hilfreich ist, wenn ein Angesprochener direkt darin vermerkt ist.

Ein direkter Vergleich mit dem besten in der Literatur zu findenden Verfahren (Almeida et al. 2014) kann direkt nicht durchgeführt werden. Berücksichtigt man den Unterschied, der Verfahren von O'Keefe auf den Texten des WSJ und den literarischen Texten, könnte eine Qualität von 90% Genauigkeit erreicht werden und damit ein mit der state of the art vergleichbares, sogar möglicherweise besseres Ergebnis. Im Gegensatz zu ihrem Verfahren ermitteln wir zudem auch noch eine angesprochene Entität.

Diskussion und Ausblick

Die Ergebnisse zeigen, dass das regelbasierte Verfahren für diese Aufgabe deutlich bessere Ergebnisse erzielen kann als alle ML-Verfahren, die in dieser Arbeit getestet wurden. Es ist geplant, die hier erstellte Zuordnung in die regelbasierte Koreferenzauflösung von (Krug et al. 2015) einzuarbeiten, um diese damit zu verbessern. Weil unsere Hauptmotivation die Verbesserung der Koreferenzresolution ist, diese aber im Ansatz von Almeida nicht wirksam verbessert werden konnte, haben wir darauf verzichtet, deren komplexes Lernverfahren nachzuvollziehen. Gerade die Ergebnisse, die in Tabelle 2 zu sehen sind, zeigen, dass mögliche Dialogsequenzen genauer untersucht werden müssen, um diese zuverlässig erkennen und auflösen zu können. Eine genaue Dialoganalyse vereinfacht wiederum die Korefenzauflösung, so dass eine Extraktion von Beziehungen zwischen Personen und Attributen zu Entitäten innerhalb der Romane möglicher erscheint.

Anhang

Featurebeschreibung (zwischen Kandidat und direkten Rede)	Verwendung für Sprecher-zuordnung	Zuordnung des Angesprochenen
1. Ist der Kandidat Subjekt	+	-
2. Das Verb in der Dependenzstruktur, auf das sich der Kandidat bezieht	+	+
3. Das POS-Tag des Kandidaten	-	-
4. Ist der Kandidat ein Pronomen	-	-
5-6. Befindet sich der Kandidat im Akkusativ/Dativ	+/+	+/+
7. Kandidat befindet sich in einer direkten Rede	+	-
8. Kandidat erscheint in der aktuellen direkten Rede	-	-
9. Kandidat befindet sich im selben Satz wie die direkte Rede	+	-
10. Die Direkte Rede beginnt mit einem kleingeschriebenem Wort	+	-
11. Zwischen Kandidat und direkter Rede befindet sich ein Doppelpunkt	-	-
12-14. Distanz zw. Kandidat und direkter Rede in Sätze/Wörter/ Entitäten	-/-/+	-/-/-
15-16. Wort an Position +1/-1	-/-	-/-
17-18. Wort an Position +1/-1 ist Satzzeichen	+/+	-/-
19-20. Wort an Position +1/-1 ist in direkter Rede	+/+	-/-
19-20. Kandidat ist Sprecher der direkten Rede an Position -1/-2	-/-	+/-
21-22. Kandidat ist Angesprochener der direkten Rede an Position -1/-2	-/-	-/-

Tab. A1: Ein Überblick über die in dieser Arbeit verwendeten Features. Durch + und - ist angegeben, ob dieses Feature gewinnbringend eingesetzt werden konnte. Zur Wahl der Features vgl. auch (Elson / McKeown 2010) und (He et al. 2013).

Regelbezeichnung	Regelbeschreibung
(1) Explizite Sprechererkennung	Nutzt Pattern-Matching und grammatikalische Regeln um explizite Erwähnungen eines Sprechers im direkten Umfeld einer direkten Rede zu erkennen.
(2) Explizite Erkennung des Angesprochenen	Nutzt Pattern-Matching und grammatikalische Regeln um explizite Erwähnungen eines Angesprochenen innerhalb der direkten Rede zu erkennen.
(3) Explizite Erkennung des Angesprochenen II	Wie (1) nur für den Angesprochenen
(4) Explizite Sprechererkennung II	Wie (1), nur der Kontext wird um 1 Satz außerhalb der direkten Rede erweitert.
(5) Vorwärtspropagierung	Zwei direkt aufeinanderfolgenden direkten Reden wird der Sprecher/ Angesprochener der ersten direkten Rede zugeordnet, wenn beide direkte Reden innerhalb des selben Satzes liegen
(6) Rückwärtspropagierung	wie (5) nur mit entgegengesetzter Richtung der Ausführung
(7) Nachbarschaftspropagierung	Direkten Reden, die keinen eingeschobenen Kontext aufzeigen, wechseln den Sprecher/Angesprochenen (falls vorhanden)
(8) Fragenpropagierung	Nach einer Frage wechseln Sprecher/Angesprochener
(9) Dialogpropagierung	Direkte Rede mit maximal einem zwischenliegenden Satz wechseln ihren Sprecher/ Angesprochenen
(10) Default-Sprecher/ Angesprochener	Als Sprecher wird das letzte Subjekt außerhalb direkter Reden gesetzt, als Angesprochener das letzte Subjekt, das nicht Sprecher der aktuellen direkten Rede ist.

Tab. A2: Überblick über die Regelpäse für das in dieser Arbeit vorgestellte regelbasierte Verfahren zur Sprecherzuordnung bzw. Zuordnung eines Angesprochenen. Optimale Reihenfolge der Ausführung der Regeln aufgrund der Auswertung des Trainingssatzes:
(1)→(2)→(3)→(4)→(5)→(6)→(7)→(5)→(6)→(8)→(9)→(5)→(6)→(7)→(10).

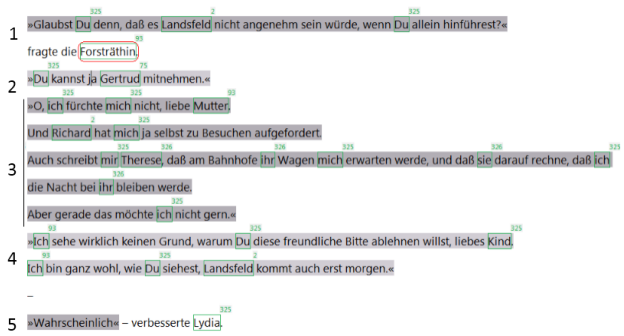


Abb. A3: Auszug aus Aston Louise "Lydia": Beispiel für die Erkennung von Sprecher und Angesprochenem in direkten Reden gemäß den Regeln in Tabelle A2. Im ersten Durchlauf wird mit der Regel (1) die Sprecherin für die direkten Reden **1** und **5** erkannt. Anschließend erkennt Regel (7) in Rückwärtsrichtung jeweils abwechselnd Sprecherin **4** und **2** und Angesprochene in **3**. Schließlich erkennt Regel (7) in Vorwärtsrichtung die Sprecherin in **3** und die Angesprochene in **4** und **2**.

Bibliographie

Almeida, Mariana S.C. / Almeida, Miguel B. / Martins, André F.T. (2014): "A joint model for quotation attribution and coreference resolution", in: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden* 39-48.

Bohnet, Bernd / Kuhn, Jonas (2012): "The best of both worlds: a graph-based completion model for transition-based parsers." In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: 77-87.

Chaganty, Arun / Muzny, Grace (2015): *Quote Attribution for Literary Text with Neural Networks* https://cs224d.stanford.edu/reports/ChagantyArun.pdf [letzter Zugriff 08. Februar 2016].

Celikyilmaz, Asli / Hakkani-Tur, Dilek / He, Hua / Kondrak, Greg / Barbosa, Denilson (2010): "The actortopic model for extracting social networks in literary narrative.", in: *Proceedings of the NIPS 2010 Workshop Machine Learning for Social Computing* https://webdocs.cs.ualberta.ca/~denilson/files/publications/nips2010.pdf [letzter Zugriff 08. Februar 2016].

Elson, David K. / Dames, Nicholas / McKeown, Kathleen R. (2010a): "Extracting social networks from literary fiction", in: *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics http://www1.cs.columbia.edu/~delson/pubs/ACL2010-ElsonDamesMcKeown.pdf [letzter Zugriff 08. Februar 2016].

Elson, David K. / McKeown, Kathleen R. (2010b): "Automatic Attribution of Quoted Speech in Literary Narrative", in: *Proceedings of AAAI* 1013-1019.

Glass, Kevin / Bangay, Shaun (2006): "Hierarchical rule generalisation for speaker identification in fiction books", in: *Proceedings of the 2006 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*. South African Institute for Computer Scientists and Information Technologists: 31-40.

Glass, Kevin / Bangay, Shaun (2007): "A naive salience-based method for speaker identification in fiction books", in: *Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA'07)* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.494.3729&rep=rep1&type=pdf [letzter Zugriff 16. Februar 2016].

He, Hua / Barbosa, Denilson / Kondrak, Grzegorz (2013): "Identification of Speakers in Novels", in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria: 1312-1320.

Iosif, Elias / Mishra, Taniya (2014): "From Speaker Identification to Affective Analysis: A Multi-Step System for Analyzing Children's Stories", in: *EACL* 2014: 40-49.

Jannidis, Fotis / Krug, Markus / Reger, Isabella / Toepfer, Martin / Weimer, Lukas / Puppe, Frank (2015): "Automatische Erkennung von Figuren in deutschsprachigen Romanen", in: *Digital Humanities im deutschsprachigen Raum (Dhd 2015), Graz, Austria*.

Joachims, Thorsten (2002): *Learning to classify text using support vector machines*. Methods, theory and algorithms (= The Springer International Series in Engineering and Computer Science 668). New York: Springer.

Krug, Markus / Puppe, Frank / Jannidis, Fotis / Macharowsky, Luisa / Reger, Isabella / Weimer, Lukas (2015): "Rule-based Coreference Resolution in German Historic Novels", in: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* 98-104.

Lee, Heeyoung / Peirsman, Yves / Chang, Angel / Chambers, Nathanael / Surdeanu, Mihai / Jurafsky, Dan (2011): "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task", in: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics http://nlp.stanford.edu/pubs/conllst2011-coref.pdf [letzter Zugriff 08. Februar 2016].

McCallum, Andrew Kachites (2002): *MALLET: A Machine Learning for Language Toolkit* http://mallet.cs.umass.edu [letzter Zugriff 08. Februar 2016].

Mikolov, Tomas / Sutskever, Ilya / Chen, Kai / Corrado, Greg / Dean Jeffrey (2013): "Distributed representations of words and phrases and their compositionality", in: *Advances in neural information processing systems* 26 http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf [letzter Zugriff 08. Februar 2016].

O'Keefe, Tim / Pareti, Silvia / Curran, James R. / Koprinska, Irena / Honnibal, Matthew (2012): "A sequence labelling approach to quote attribution", in: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012: 790–799.

Rahman, Altaf / Ng, Vincent (2011): "Narrowing the modeling gap: a cluster-ranking approach to coreference resolution", in: *Journal of Artificial Intelligence Research* 40: 469-521.

Ruppenhofer, Josef / Sporleder, Caroline / Shirokov, Fabian (2010): "Speaker Attribution in Cabinet Protocols", in: *The seventh international conference on Language Resources and Evaluation (LREC)* 2510-2515.

Schmid, Helmut (1999): "Improvements in part-of-speech tagging with an application to German", in: Armstrong, Susan / Church, Kenneth / Isabelle, Pierre / Manzi, Sandra / Tzoukermann, Evelyne / Yarowsky, David (eds.): *Natural language processing using very large corpora* (= Text, Speech and Language Technology 11). New York: Springer 13-25.

Schmid, Helmut / Laws, Florian (2008): "Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging", in: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* 777–784.

Sutton, Charles / McCallum, Andrew (2006): "An introduction to conditional random fields for relational learning", in: Getoor, Lise / Taskar, Ben (eds.):*Introduction to statistical relational learning*. Cambridge, MA / London: The MIT Press 93-128.

Zhang, Jason Y. / Black Alan W. / Sproat, Richard (2003): "Identifying speakers in children's stories for speech synthesis", in: *EUROSPEECH* 2041-2044.

Moving around the City of Glass

Laubrock, Jochen
laubrock@uni-potsdam.de
Universität Potsdam, Deutschland

Hohenstein, Sven
hohenstein@uni-potsdam.de
Universität Potsdam, Deutschland

Thoß, Alexander
athoss@uni-potsdam.de
Universität Potsdam, Deutschland

Graphische Romane und Comics vereinen als hybride Gattung Aspekte von Literatur und bildender Kunst und werden deshalb auch als sequenzielle Kunst bezeichnet. Man kann erwarten, dass sich die psychologische Wirkung graphischer Romane von der rein wortbasierter Romane unterscheidet. Einerseits sagt ein Bild mehr als tausend Worte, was die deutlich geringeren Zahl von Wörtern bei graphischen Adaptionen klassischer Romane erklärt, andererseits hat der Leser weniger Freiheitsgrade bei der visuellen Ausgestaltung der Szene und wird durch die erforderliche Integration von Bild und Text möglicherweise vor besondere Aufgaben gestellt. Wie interagieren Bild und Text beim Lesen graphischer Literatur und ermöglichen das Verstehen des Gesamtwerkes? Welche besonderen Herausforderungen stellt das multimediale Format an den Leser? Wie unterscheidet sich die Narrativität graphischer von der herkömmlicher Romane?

Im Kontext der interdisziplinär ausgerichteten Nachwuchsgruppe "Hybride Narrativität: Digitale und Kognitive Methoden zum Leseverständnis Graphischer Literatur" wurde die *Comic Book Markup Language* (CBML) erweitert zur *Graphic Novel Markup Language* (GNML) und ein zugehöriges Annotationstool entwickelt, das auch die semiautomatische Erkennung von Elementen ermöglicht. Durch die graphische Benutzeroberfläche wird es ermöglicht, auf einfache Weise die Strukturen einer Comic-Seite zu annotieren. Die Annotation einer Vielzahl von Werken erlaubt beispielsweise vergleichende Untersuchungen über die in kulturellen Traditionen oder bei bestimmten Autoren vorherrschende Stilistik. Um auch die kognitive Verarbeitung auf Seite der Rezipienten abzubilden, sammeln wir empirische Daten über die Wirkung graphischer Literatur. Neben der Erhebung psychologischer Größen wie Reaktionszeiten und subjektiven Maßen werden dazu insbesondere auch Blickbewegungen einer größeren Zahl von Leserinnen und Lesern einiger kanonischer Graphic Novels aus unterschiedlichen Kulturkreisen registriert. Die Bewegungen des Auges haben sich in einer Vielzahl an Studien als valides, nichtreaktives Maß für die Verarbeitung und das Verstehen von Text und Bild erwiesen, in dem sich zudem auch unbewusste Verarbeitungsprozesse niederschlagen. Für die Zuordnung der Blickbewegungsdaten auf das Material, Weiterverarbeitung und statistische Analyse der XML-Daten wird ein R-Paket entwickelt. Damit wird es auch möglich sein, erhobene Daten zu visualisieren.

Im vorliegenden Beitrag illustrieren wir den potenziellen Nutzen einer solchen Kombination anhand einer Analyse der Eye-tracking-Daten von (a) einer Sammlung kürzerer Passagen aus mehreren kanonischen graphischen Romanen – einem repräsentativen Korpus – und (b) Passagen der Graphic-Novel-Adaptation von Paul Austers Roman "City of Glass" (Auster 1985; Karasik / Mazzucchelli / Auster 1994). Für (a) berichten wir eine Analyse des relativen Anteils von Fixationen auf Text vs. Bild. Effekte der Wortlänge sowie statistischen Worthäufigkeit in der geschriebenen Sprache auf die Fixationsdauern zeigen, dass der Text auch tatsächlich gelesen und rezipiert wird. Analysen der Fixationsmuster zeigen zudem, dass der Text meist vor dem Bild angeschaut wird und das Bild oft entweder gar nicht oder rein im peripheren Sehfeld analysiert wird. Interessante Objekte wie Personen oder Gesichter werden mit höherer Aufmerksamkeit bedacht als Objekte des Hintergrundes. Ob das Bild überhaupt betrachtet wird, ist unter anderem vom Informationsgehalt des Bildes abhängig, der sich wiederum je nach Art des Überganges zwischen zwei Panels unterscheidet (McCloud 1993). Wenn sich die bestehende Handlung auf dem nächsten Panel fortsetzt, wird dieses mit höherer Wahrscheinlichkeit übersprungen als ein Panel, das sich deutlicher von seinem Vorgänger unterscheidet und damit einen entscheidenderen Anteil an der Handlung hat, beispielsweise bei einem Szenenwechsel. Bei (b) fokussieren wir insbesondere auf die Frage der Text-Bild-Beziehung. Unterscheidet sich beispielsweise das Blickverhalten, wenn Bild und Text auf gemeinsame vs. unterschiedliche Handlungsstränge fokussieren? Zudem berichten wir über deutliche Zusammenhänge von Leser-Expertise mit graphischer Literatur und explizit gemessener Verständnistiefe bei diesem speziellen Werk sowie implizit gemessenen Blickdauern. Anders als beim Lesen von reinem Text drückt sich Expertise beim Lesen graphischer Literatur nicht in geringeren, sondern in höheren Betrachtungszeiten aus, die sich speziell auf den Bildanteil konzentrieren.

Perspektivisch soll das Material anhand von aus dem computationalem Sehen abgeleiteten Deskriptoren beschrieben und klassifiziert werden. Beispielsweise sollen dazu Farb-Histogramme, lokales Fourier-Spektrum, der SURF-Algorithmus etc. genutzt und Klassifikationsverfahren aus dem Bereich des maschinellen Lernens angewandt werden. Diese werden sicherlich die stilistische Beschreibung anreichern und können als potenzielles Nebenprodukt auch die Suche in Bilddatenbanken ohne explizites verbales Tagging vorbereiten. Im Rahmen unseres Projektes erhoffen wir uns von einer derartigen Anreicherung der Daten jedoch eine Antwort auf die Frage, wie sich die Wechselwirkung solcher Bottom-up-Merkmale mit Top-down-Einflüssen von einfacher Worthäufigkeit bis hin zu narratologischen Elementen auf das Blickbewegungsverhalten und die Rezeption der Literatur auswirkt. Letztlich ist es eine empirische Frage, wie viel des Verhaltens und Verstehens sich durch simple Deskriptoren erklären lässt und welche Anteile sich durch Hinzunahme weiterer, beispielsweise konfigurationaler oder strategisch-aufgabenorientierter Merkmale aufklären lassen.

Zusammenfassend berichten wir über eine von kognitiven und psychologischen Fragen geleitete Analyse graphischer Literatur und darauf erhobenen Blickbewegungsdaten. Zum einen werden dabei allgemeine Prinzipien anhand einer Sammlung verschiedener kanonischer Werke des Genres illustriert.

Zum anderen beschreiben wir eine tiefere Analyse eines spezifischen Exemplars dieser Gattung.

Bibliography

Auster, Paul (1985): *City of glass.* Volume 1 of The New York trilogy. New York: Penguin Books.

Karasik, Paul / Mazzucchelli, David / Auster, Paul (1994): *Paul Auster's City of glass*. New York: Avon Books.

McCloud, Scott (1993): *Understanding comics: the invisible art*. Northampton, MA: Tundra.

Kafkas Stil. Zur Psychostilistik der Tagebücher Kafkas

Lauer, Gerhard
glauer@gwdg.de
Universität Göttingen, Deutschland

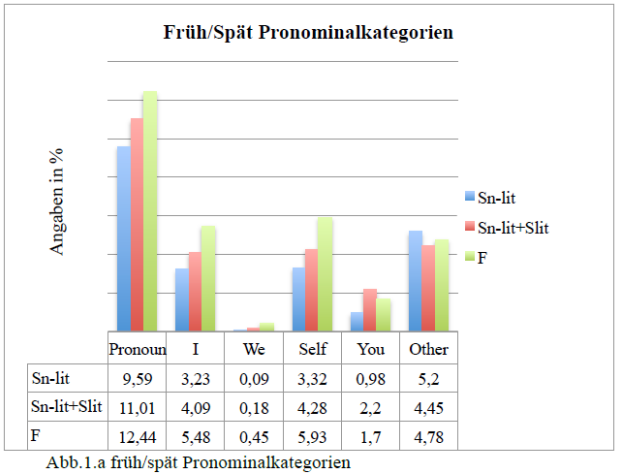
Mattner, Cosima
cosima.mattner@stud.uni-goettingen.de
Universität Göttingen, Deutschland

Herrmann, Berenike
jb.herrmann@phil.uni-goettingen.de
Universität Göttingen, Deutschland

Kafkas Werkbiographie ist Gegenstand divergierender Forschungsmeinungen. Besonders seine Tagebücher und ihr Zusammenhang mit dem übrigen Werk wurden und werden kontrovers diskutiert. Unterscheidet sich sein literarisches Werk von seinen übrigen Aufzeichnungen, gibt es eine Entwicklung in Kafkas Schreibstil und lassen sich aus Kafkas Aufzeichnungen überhaupt Rückschlüsse auf dessen Psyche ziehen, - das sind Fragen, die die Kafka-Forschung seit gut einem halben Jahrhundert beschäftigen. Mehrheitlich geht die Forschung dabei von der These aus, dass sich in Kafkas Schreiben literarisches Werk und Tagebücher nicht unterscheiden lassen, vielmehr sein Schreiben als ein geschlossenes Ganzes gesehen werden müsse, das immer abstrakter und rätselhafter werde. Vor diesem Hintergrund werden Aussagen über Kafkas psychische Verfasstheit teils auf Basis der Gesamtheit der Aufzeichnungen getroffen - ungeachtet formaler Eigenheiten und Differenzen zwischen diesen (Engel / Auerochs 2010). Diese Thesen der Kafka-Forschung wurden von uns in einem neuen Verfahren überprüft. Teilweise finden wir sie bestätigt, teilweise aber auch sind sie zu revidieren.

Auf der Basis des Projekts ziehen wir weiterreichende Folgerungen für einen psychostilistischen Ansatz in der literaturwissenschaftlichen Forschung.

Unser Vortrag stellt einen psychostilistischen Untersuchungsansatz vor, der persönlichkeitspsychologische und stilometrische Methoden verbindet. Ziel unserer Analysen ist die Verbesserung der Reliabilität und Validität der literaturwissenschaftlichen Forschung, hier der Forschung zu Kafka. Dazu haben wir zunächst die bisherige Kafka-Forschung philologisch aufgearbeitet und die verschiedenen Positionen in der Debatte um Kafkas Werke und insbesondere seine Tagebuchaufzeichnungen identifiziert und typisiert. Analog zu Klassifizierungen der Tagebucheinträge in der Kritischen Ausgabe (Koch 1990) haben wir Kafkas Werk in Korpora eingeteilt (frühe vs. späte Schriften, literarische vs. nicht-literarische Schriften). Dann haben wir ausgehend von der sozialpsychologischen Forschung über den Zusammenhang zwischen Wortgebrauch und Persönlichkeit (Chung / Pennebaker 2007; Ireland / Pennebaker 2010) die Korpora genauer untersucht. Mit einem diktionärbasierten Ansatz (Wolf et al. 2008) haben wir den Wortgebrauch Kafkas digital und quantitativ näher analysiert und schließlich die Ergebnisse mit dem ausbalancierten Korpus des Digitalen Wörterbuchs der deutschen Sprache verglichen. Besonderes Augenmerk lag dabei auf dem Gebrauch der Pronomina, der Verwendung von Wörtern, die soziale Beziehungen ausdrücken, der Häufigkeit von positiven und negativen Emotionsworten und einiger weiterer, besonders persönlichkeitspsychologisch signifikanter Kategorien. Die Abbildung zeigt exemplarisch den Gebrauch von Pronominalkategorien in früheren („F“) im Vergleich mit späteren („$S_{n\text{-}lit+}S_{lit}$“) Schriften Kafkas:



	Pronoun	I	We	Self	You	Other
Sn-lit	9,59	3,23	0,09	3,32	0,98	5,2
Sn-lit+Slit	11,01	4,09	0,18	4,28	2,2	4,45
F	12,44	5,48	0,45	5,93	1,7	4,78

Abb.1.a früh/spät Pronominalkategorien

Unsere Ergebnisse zeigen, dass sich Kafkas Stil über die vergleichsweise kurze Zeit seiner Werkbiographie verändert hat. Um nur einige der Entwicklungen zu nennen: Selbst- und sozialbezügliche Äußerungen nehmen ab, negative Emotionswörter nehmen zu.

Diese Veränderungen finden wir jedoch nur in Kafkas Ego-Dokumenten ausgeprägt, sein literarisches Werk weist diese Veränderungen im Wortgebrauch nicht auf. Vielmehr finden sich hier konstante Wortfrequenzen. Vor allem unterscheiden sich die als literarische qualifizierten Dokumente im Pronominalgebrauch von nicht-literarischen Texten. Das quantitativ-digitale Verfahren legt damit neben den Stilentwicklungen Genre-Signale offen, die eingehender untersucht wurden.

Unser Projekt demonstriert, dass und wie eine quantitativ-digitale Methodik eine genauere Erfassung stilistischer Eigenschaften ermöglicht. Damit können Ergebnisse konventioneller literaturwissenschaftlicher Forschung überprüft und gegebenenfalls korrigiert werden. Außerdem erweist sich einmal mehr, dass Texteigenschaften empirisch in Daten abgebildet werden können, Daten, die dann als Grundlage für weitere literaturwissenschaftliche Forschung herangezogen werden können. Die Analyse wurde deshalb durch die Betrachtung anderer Korpus ergänzt. Im Vergleich u. a. mit dem DWDS-Korpus wird Kafkas besonderer Stil näher bestimmt, aber auch aufgezeigt, wie eine literaturwissenschaftliche Forschung psychostilistische Methoden anpassen muss, um den Besonderheiten literarischer Texte gerecht zu werden. Generelle Schlussfolgerungen für die Chancen einer literaturwissenschaftlichen Psychostilistik werden abschließend diskutiert.

Bibliographie

Chung, Cindy K. / Pennebaker, James W. (2007): "The psychological function of function words", in: Konrad Fiedler (ed.): *Social communication*. Frontiers of social psychology. New York: Psychology Press 343-359.

Engel, Manfred / Auerochs, Bernd (2010): *Kafka-Handbuch*. Leben – Werk – Wirkung. Stuttgart: Metzler.

Ireland, Molly E. / Pennebaker, James W. (2010): "Language style matching in writing: Synchrony in essays, correspondence, and poetry", in: *Journal of Personality and Social Psychology* 99: 549-571.

Koch, Hans-Gerd (1990) (ed.): *Franz Kafka*. Gesammelt Werke in Einzelbänden in der Fassung der Handschriften. Frankfurt am Main: Fischer.

Wolf, Markus / Horn, Andrea B. / Mehl, Matthias R. / Haug, Severin / Pennebaker, James W. / Kordy, Hans (2008): "Computergestützte quantitative Textanalyse. Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count", in: *Diagnostica* 54: 85-98.

Die Lehre der digitalen Visualisierung am Beispiel der Architektur

Lengyel, Dominik
lengyel@tu-cottbus.de
Brandenburgische Technische Universität Cottbus-Senftenberg, Deutschland

Toulouse, Catherine
toulouse@tu-cottbus.de
Brandenburgische Technische Universität Cottbus-Senftenberg, Deutschland

Forschungskontext

Visualisierung ist die visuelle Vermittlung nicht sichtbarer Inhalte. Dies sind im Kontext der Architektur zwar auch Planungen, bei der Beschäftigung mit der Geschichte jedoch unmittelbar geisteswissenschaftliche – baugeschichtliche, bauforscherische, archäologische oder kunsthistorische – Inhalte.

Der Vortrag soll darstellen, wie Studierende der Architektur in die Lage versetzt werden, diese geisteswissenschaftlichen Inhalte visuell zu vermitteln. Architektur ist prinzipiell fächerübergreifend, geradezu prädestiniert, unterschiedliche Fächer – beim Bauen Gewerke genannt – miteinander in Einklang zu bringen. Dabei lernen Architektur Studierende nicht nur die Techniken und Methoden der klassischen Darstellung, sondern vor allem die Besonderheiten des Digitalen, dass nämlich im Digitalen aus den vorhandenen Möglichkeiten besonders gezielt selektiert werden muss, um zugleich effizient und überzeugend im Sinne der Wissensvermittlung – und auch der Wissenschaftsvermittlung – vorgehen zu können.

Der Vortrag geht von der allgemeinen Bedeutung der Gestaltung beim digitalen Visualisieren aus, das heißt von der Vermittlung dieser Bedeutung über die Vermittlung der notwendigen grundlegenden Techniken und Methoden bis hin zu Fallbeispielen aus der forschungsinduzierten Lehre. Der Vortrag erläutert damit die Grundlagen – und sieht sich als komplementäre Ergänzung – unseres Vortrags auf der Grazer DHd-Konferenz 2015 „Die Bedeutung architektonischer Gestaltung in der visuellen Vermittlung wissenschaftlicher Unschärfe am Beispiel von Ktesiphon und weiteren archäologischen Stätten“.

Lehre und Forschung

Digitale Visualisierungen sind, anders als analoge Visualisierungen, Ergebnis mehrfacher abstrakter Übersetzungsprozesse, deren kontrollierter, wissenschaftlicher Umgang ein hohes Abstraktionsvermögen erfordert. Es hat sich gezeigt, dass die schrittweise Heranführung an die zunehmende Komplexität des Visualisierens Studierende gezielt darauf vorbereitet, in jedem Schritt das übergeordnete Ziel der Visualisierung im Auge zu behalten. In der Architektur allgemein als Gestaltung bezeichnet, bedeutet dies, dass alle Schritte auf dem Weg zur Visualisierung unter Berücksichtigung sämtlicher Gesichtspunkte erfolgt und im Idealfall nichts dem Zufall überlassen wird. Zahlreiche realisierte Forschungsprojekte haben zeigt, dass erst die systematische und streng methodengebundene Vorgehensweise Visualisierungen ermöglicht, die als Forschungsinstrumente Mehrwerte sowohl in der Vermittlung als auch in der Forschung selbst generieren (vgl. Laufer 2011; Lengyel / Toulouse 2011a, c; Lengyel / Schock-Werner / Toulouse 2011).

Die Reihe der Übersetzungen geisteswissenschaftlicher Inhalte in eine digitale Visualisierung beginnt im Allgemeinen auf einer textlichen Grundlage. Im Allgemeinen wird dieser Text von Illustrationen flankiert, die bereits den Anspruch haben, den Inhalt zu konkretisieren. Aus beiden wird ein digitales Modell generiert, das in der Lage sein muss, die sprachlich bedingte, und zwar durchaus intendierte, Unschärfe im Wissen, aufzunehmen. Das Modell schließlich wird dann zu einer Visualisierung transformiert, die die modellierten Inhalte visuell wahrnehmbar macht. Damit entsteht beim Betrachter im besten Fall ein Modell im Kopf, das dem Gedanken des ursprünglichen geisteswissenschaftlichen Ursprungs entspricht.

Architektenlehre als Fallbeispiel

Zu allen oben genannten Schritten der Übersetzung geisteswissenschaftlicher Inhalte gibt es in der Architektur Entsprechungen. Auch hier entsteht zunächst im Kopf des Planers eine Vision eines Gebäudes, das über mehrere Transformationsstufen schließlich zu einer Visualisierung gelangt, die die Grundlage für eine Beauftragung werden kann. Insofern ist die Übersetzung geisteswissenschaftlicher Inhalte für die Ausbildung von Architekten nicht wesensfremd und wird seit Jahren erfolgreich erprobt (vgl. v. a. Lengyel / Toulouse 2011b).

Der Anspruch an die Architektenlehre ist immer auch ein Anspruch an eine hohe Gestaltungsqualität. Diese so zu vermitteln, dass deutlich wird, dass Gestaltung vor allem eine hohe Konsistenz der Einzelteile bedeutet, nicht etwa eine Geschmacksache ist, sondern eben aus Erfahrung, intensiver Auseinandersetzung mit Grundlagen und auch der eigenen Arbeit sowie der Einbeziehung fächerübergreifender Perspektiven erwächst, ist Inhalt der Lehre auch der digitalen Visualisierung.

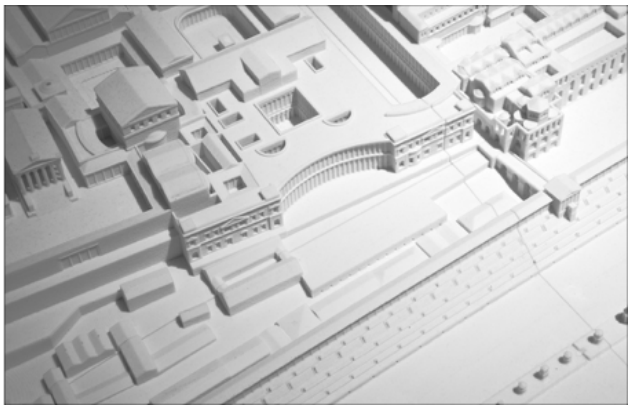
Das Curriculum

Die Grundlage der Gestaltung bildet die Rezeption von Gestaltung. Erreicht werden soll vor allem die Erkenntnis, dass Gestaltung beurteilt und bewertet werden kann. Die individuell gewählte Gestaltungsrichtung mag subjektiv sein, diese ist jedoch unabhängig von der Gestaltungshöhe. Um die Unabhängigkeit der Richtung der Gestaltung zu vermitteln, werden Beispiele aus der bildenden Kunst als Referenz verwendet. Das allerdings setzt ein offenes Kunstverständnis voraus, durch das über den unmittelbaren ästhetischen Eindruck des Betrachters die Inspiration für die eigene Arbeit entsteht.

Eine kurze Darstellung der verwendeten Techniken gibt einen Überblick über die Grundlage der Ausdrucksmöglichkeiten für die digitalen Visualisierungen. Die erste ist die Darstellende Geometrie, da ein geometrisches Verständnis Voraussetzung für digitale Visualisierungen ist, sobald diese räumliche Konstruktionen beinhalten (vgl. Lengyel 2005).

Das sich anschließende CAD – nur vermeintlich der Kern der digitalen Visualisierung – ist tatsächlich nur der konstruktive Teil, der zudem ergänzt wird durch nicht konstruierte Daten wie 3D-Scans. Diese sind völlig anders aufgebaut als konstruierte Geometrie. Die Herausforderung besteht darin, Konstruiertes und 3D-Scans aufeinander abzustimmen, wie dies beispielsweise bei Rekontextualisierungen antiker Skulpturen der Fall ist (vgl. Lengyel / Toulouse 2011c und 2014). Auch die Ausgabe als 3D-Druck gehört zum weiteren Bereich der Visualisierung, doch steht hier nicht etwa die Technik im Vordergrund. Zwar muss die Statik des Materials beachtet werden, die eigentliche Herausforderung ist aber auch hier die Gestaltung des 3D-Drucks. Die Geometrie muss trotz

der technisch bedingten Anpassung den architektonischen Charakter beibehalten. Auch hierfür sind stringente Gestaltungsrichtlinien zu berücksichtigen.



Spätestens in dieser Phase muss das Modell zurückgeführt werden in den Bereich der Wahrnehmung. Die Bedeutung der gezielten Projektion des Modells in ein sichtbares Medium wird durch die Analogie virtuelles Modell und virtuelle Fotografie unterstrichen (vgl. v. a. Lengyel / Schock-Werner / Toulouse 2011). Dies soll vor allem in der Lehre verankern, dass es sich bei der Projektion um bewusste gestalterische Prozesse handelt. Um den intensiven Arbeitsprozess der Gestaltung auf die Spitze zu treiben, haben sich strenge Rahmenbedingungen als zielführend erwiesen. Sie zwingen zu einer Auseinandersetzung in der Tiefe, wodurch Erkenntnisse generiert werden, die sich problemlos auf andere Projekte übertragen lassen.

Das Curriculum in der Architekturlehre ist bestimmt durch die Steigerung der Komplexität. Am Beginn steht die visuelle Reflexion vorgegebener architektonischer Gestaltung. Die Auseinandersetzung mit bereits realisierter Architektur verlangt es, fremde Gestaltungsmotive zu reflektieren, zunächst also ein den Geisteswissenschaften geläufiger Vorgang, diese dann aber in eine eigenständig gestaltete Darstellung umzusetzen. So werden hier die Grundlagen der digitalen Visualisierung gelegt, denn das zugrunde liegende Modell wird im virtuellen Raum dreidimensional konstruiert. Hierfür werden nicht nur die digitalen Werkzeuge und auch die digitalen Methoden der Konstruktion, Rotation, Vervielfältigung, Konstruktionspunkte usw. geübt, sondern die Analyse der geometrischen Strukturen führt darüberhinaus zu einem inhaltlichen Verständnis der vorgefundener Motive.

Mit zunehmender Komplexität erhält entweder der Kontext der Architektur oder auch die eigenständige Gestaltung ein höheres Gewicht. In der Seminarreihe „Perspektiven Gestalten“ vermittelt die Architekturfotografie als praktisches Vorbild Komposition als Gestaltungskonzept (vgl. Lengyel 2008). Reflektiert werden neben der Komposition natürlich auch Beleuchtung, Brennweite, natürliche Augenhöhe, senkrechte Bildebene. Gleichzeitig werden auch Begriffe der inhaltlichen Ebene wie Kontext, Tiefenstaffelung und Detaillierungsgrad behandelt.

Aufgaben der nächsten Stufe erhalten zusätzliche Anforderungen. In der Fotomontage wird eine eigenständige Raumgestaltung mit einem vorgegebenen architektonischen Kontext konfrontiert. Durch den Dialog zwischen realer und virtueller Welt wird ein wichtiger Anspruch an die digitale Visualisierung deutlich, nämlich die angemessene Abstraktion. Wie definiert nämlich muss ein Modell sein, um erstens neben dem Foto bestehen zu können und zweitens trotz seiner unvermeidlichen Abstraktion immersiv und überzeugend zu wirken. Diese Kombination entspricht natürlich auch dem üblichen Arbeiten der Architekten, Abstrakte Inhalte überzeugend zu präsentieren. Vermittelt wird also die Stärke der Abstraktion als Weg, bestimmte Komponenten besonders in der Vordergrund zu stellen. Die Abstraktion umfasst alle Stufen von der geringfügigen geometrischen Vereinfachung über die Typenbildung bis hin zu diagrammatischen Strukturen (vgl. Lengyel / Toulouse 2013).

Um die Bildaussage in höchstem Maße auf das gesteckte Ziel abzustimmen, wie es bei der Visualisierung geisteswissenschaftlicher Inhalte der wissenschaftliche Anspruch verlangt, verfolgen die Aufgaben die Raumgestaltung allein zum Zweck der Raumwirkung. Diese Fokussierung ist kein Eingeständnis, sondern die bewusste Betonung eines einzelnen Aspektes, auch dies also die Auseinandersetzung in die Tiefe als Voraussetzung für intensives gestalterisches Arbeiten in der digitalen Visualisierung.

Im Masterstudium wird es dann möglich, unmittelbar forschungsinduziert zu lehren. Allgemeine Fragestellungen zum geometrischen Abstraktionsprozess etwa lauten, wie weit eine Abstraktion gehen kann, ohne das die Wiedererkennbarkeit leidet. Oder wie sich Farbigkeit auf Raumwahrnehmung auswirkt, auf die Tiefenwahrnehmung, auf das Licht, die Plausibilität des Raumes usw.

Abgeschlossen wird die Heranführung der werdenden Architekten an die digitalen Visualisierungen als studentische Hilfskraft in laufenden Forschungsprojekten. Sie kommen so besonders nah an die Komplexität des Themas heran. Gerade mehrjährige Projekte in Forschungsgruppen oder Clustern erlauben es, einen besonderes tiefen Eindruck zu gewinnen. Das Excellence Cluster TOPOI, bei dem der Lehrstuhl mit dem Thema der Darstellung von Unschärfe im archäologischen Wissen – oben erwähnt als Beitrag zur DHd-Konferenz 2015 – beteiligt ist, beispielsweise besteht fast ausschließlich aus Geisteswissenschaftlern.

Fazit

Die digitale Visualisierung ist eine gestalterische Tätigkeit, die durch die wissenschaftliche Reflexion

erst einen Mehrwert in den Geisteswissenschaften entfalten kann. Obgleich die Beherrschung grundlegender Techniken unverzichtbar ist, sind diese noch kein Garant für eine weiterführende Anwendung. Technik und Methoden sind noch am ehesten erlernbar. Gestaltung aber ist vor allem eine Sache von Praxis und Erfahrung. Erst der kritische Umgang mit den Werkzeugen, der intensive Dialog mit den beteiligten Wissenschaften und vor allem Selbstreflexion erlauben es der digitalen Visualisierung, als Instrument sowohl der Forschung selbst als auch deren Vermittlung zu dienen. Das vorgestellte Curriculum aus der Architektenlehre sei als Beispiel gedacht für eine mögliche Heranführung an das Thema digitale Visualisierungen als eine Möglichkeit der Aufarbeitung geisteswissenschaftlicher Daten.

Bibliographie

Laufer, Eric / Lengyel, Dominik / Pirson, Felix / Stappmanns, Verena / Toulouse, Catherine (2011): "Die Wiederentstehung Pergamons als virtuelles Stadtmodell", in: Scholl, Andreas, Kästner, Volker / Grüssinger, Ralf (eds.): *Antikensammlung Staatliche Museen Berlin*. Pergamon. Panorama der antiken Metropole. Petersberg: Imhof 82–86.

Lengyel, Dominik (2005): "Anschauliche Geometrie", in: *Positionen der Geometrieausbildung*. Tagungsband der ersten Tagung der Deutschen Gesellschaft für Geometrie und Grafik 45–52.

Lengyel, Dominik / Toulouse, Catherine (2008): *Perspektiven Gestalten*. Cottbus: Brandenburgische Technische Universität.

Lengyel, Dominik / Toulouse, Catherine (2011a): "Die Gestaltung der Vision Naga - Designing Naga's Vision", in: Kröper, Karla / Schoske, Sylvia / Wildung, Dietrich (eds.): *Königsstadt Naga - Naga, Royal City*. Grabungen in der Wüste des Sudan - Excavations in the Desert of the Sudan. München / Berlin: Naga-Projekt Berlin / Staatliches Museum Ägyptischer Kunst München 163-175.

Lengyel, Dominik / Toulouse, Catherine (2011b): "Darstellung von unscharfem Wissen in der Rekonstruktion historischer Bauten", in: Heine, Katja / Rheidt, Klaus / Henze, Frank / Riedel, Alexandra (eds.): *Von Handaufmaß bis High Tech III*. 3D in der historischen Bauforschung. Darmstadt / Mainz: Philipp von Zabern 182-186.

Lengyel, Dominik / Toulouse, Catherine (2011c): "Ein Stadtmodell von Pergamon - Unschärfe als Methode für Darstellung und Rekonstruktion antiker Architektur", in: Petersen, Lars / Hoff, Ralf von den (eds.): *Skulpturen in Pergamon*. Gymnasion, Heiligtum, Palast. Katalog zur gleichnamigen Ausstellung, Archäologische Sammlung der Universität Freiburg, 6. Mai 2011 - 31. Juli 2011: 22–26.

Lengyel, Dominik / Toulouse, Catherine (2012): "Rekonstruktionszeichnungen der Bauphasen des kaiserzeitlichen Palatin in Rom", in: Märtin, Ralf-Peter (ed.): *Jenseits des Horizonts*. Raum und Wissen in den Kulturen der Alten Welt. Stuttgart: Verlag Konrad Theiss 14–20.

Lengyel, Dominik / Toulouse, Catherine (2013): "Die Bauphasen des Kölner Domes und seiner Vorgängerbauten: Gestaltung zwischen Architektur und Diagrammatik", in: Boschung, Dietrich / Jachman, Julian (eds.): *Diagrammatik der Architektur*. Tagungsband Internationales Kolleg Morphomata der Universität zu Köln. Paderborn: Verlag Wilhelm Fink 327-352.

Lengyel, Dominik / Toulouse, Catherine (2014): "3D-Scans für die Rekontextualisierung antiker Skulptur", in: Bienert, Andreas / Helmsley, James / Santos, Pedro (eds.): *EVA Berlin 2014*. Elektronische Medien & Kunst, Kultur und Historie. Konferenzband. Darmstadt / Berlin: Staatliche Museen zu Berlin / Stiftung Preußischer Kulturbesitzt / Fraunhofer-Institut für Grafische Datenverarbeitung 135-142.

Lengyel, Dominik / Schock-Werner, Barbara / Toulouse, Catherine (2011): *Die Bauphasen des Kölner Doms und seiner Vorgängerbauten / Cologne Cathedral and Preceding Buildings*. Köln: Verlag Kölner Dom e.V.

Die Geowissenschaftliche Analyse von großen Mengen historischer Texte: Die Visualisierung geographischer Verhältnisse in deutschen Familienzeitschriften

McIsaac, Peter
pmcisaac@umich.edu
Literature, Sciences and the Arts, Universitaet von Michigan, USA

Jamin, Sugih
sugih@umich.edu
Electrical Engineering and Computer Science, Universitaet von Michigan, USA

Ibanez, Ines
iibanez@umich.edu
School of Natural Resources, Universitaet von Michigan, USA

Singer, Oskar

oskarsinger@gmail.com
Electrical Engineering and Computer Science, Universitaet von Michigan, USA

Bray, Benjamin

benrbray@umich.edu
Literature, Sciences and the Arts, Universitaet von Michigan, USA

In diesem Vortrag werden die Verarbeitung und Visualisierung von geowissenschaftlichen Daten in populären, in Deutschland zwischen 1853 und 1918 publizierten Familienzeitschriften wie *Die Gartenlaube*, *Deutsche Rundschau* und *Westermanns Illustrirte Deutsche Monatshefte* präsentiert. Nach einer einleitenden Diskussionen über die Fragestellungen, die eine geowissenschaftliche Analyse dieser Druckerzeugnisse aus älterer und „digitaler" geisteswissenschaftlicher Sicht motivieren, werden unsere Herangehensweisen erläutert. Neben kuratorischen Aspekten behandelt die Präsentation die von uns entwickelten Techniken des maschinellen Lernens und einer historisierenden Visualisierung, die großen Mengen von historischen Texten gerecht werden. Darüber hinaus werden erste Ergebnisse der Visualisierung gezeigt, die neue Antworten auf noch ungelöste Fragen bieten.

Die Fragestellungen, die diesem DH-Projekt zugrunde liegen, entsprechen manchen zentralen Fragen der älteren geisteswissenschaftlichen Forschung. Diese interessierte sich für die Darstellung spezifischer geographischer Orte und Gebiete zunächst im Zusammenhang mit der Entwicklung einer modernen deutschen Nationalidentität, die als überregional und allen Deutschen gemeinsam verstanden wird (Belgum 1998: xi-xv). Familienzeitschriften befassten sich bekanntlich nicht nur programmatisch mit der Formulierung und Verbreitung der historischen, sprachlichen und geographischen Konturen einer solchen Nationalidentität (Barth 1975: 205-12), sie unternahmen dies als die ersten Druckerzeugnisse, deren Verbreitung ein annähernd nationales Ausmaß annahm (McIsaac 2014: 186-8; Belgum 1998: 1-27). Während ihre relativ erschwinglichen Preise und ihre breit angelegte inhaltliche Thematik ein unerhört zahlenreiches und breites Publikum ansprachen (McIsaac 2014: 186-8; Daum 2002), ermöglichten technische Entwicklungen die zeitgleiche wöchentliche bzw. monatliche Belieferung des gesamten geographischen Gebietes, das als territoriale Basis für Deutschland als politische Nation kritisch in Frage steht (Belgum 1998: 1-27). Innenpolitisch dürften diese Druckerzeugnisse also zum Nationalgefühl im Sinne von Benedict Andersons Begriff der Nation als „vorgestellte Gemeinschaft" beigetragen haben (Anderson 2006). Zugleich war die Frage nach der geographischen Darstellung aber stets auch eine globale, indem die Familienzeitschriften Deutschlands Rollen als wichtiges Emigrationsland, später dann als aufstrebende Kolonial- und Weltmacht mit gezielten Beiträgen bewusst reflektierten (Belgum 1998: 142-82). Es geht bei diesen Fragen also um die lokalen und globalen territorialen Be-, Ein- und Entgrenzungen in ihrem Verhältnis zum deutschen Nationalgefühl.

Im Zeitalter der Globalisierung und Massenmigration haben diese Fragen nach der nationalen Identität in lokalen und internationalen Kontexten nichts an Brisanz eingebüßt, auch wenn (oder gerade weil) ihre Beantwortung mittels traditioneller Methoden nur in Ansätzen gelungen ist. Dass dies mit herkömmlicher Analyse nicht mehr zu erreichen ist, hängt im großen Maße mit der Fülle an Lesematerial zusammen, der mit normaler Lektüre nicht beizukommen ist (McIsaac 2014: 185). Erst mit der Digitalisierung ganzer Zeitschriftenauflagen, wie dies Google in Zusammenarbeit mit dem US-amerikanischen HathiTrust-Consortium unternommen hat, ist es möglich geworden, mit computerbasierten Techniken an diese Fragen heranzugehen. Diese Techniken bergen insbesondere die Möglichkeit einer kartographischen Visualisierung der geowissenschaftlichen Daten in den Familienblättern in sich, und zwar eine, die das langjährige Erscheinen der Blätter in regelmäßigen Zeitabständen historisch zu verwerten trachtet. In Bezug auf die angestrebte historisierende Visualisierung geowissenschaftlicher Daten gibt es allerdings technische, finanzielle und methodische Probleme, deren Lösung für große Mengen von historischen Texten weder trivial noch vollkommen ist.

Auf welche Weise diese Probleme sich bewältigen lassen, wird Gegenstand des Vortrags anhand von einem Korpus (*Deutsche Rundschau* 1873-1918) sein. Geschildert werden zunächst Techniken, die nicht nur zur Behebung von Problemen historischer und kuratorischer Natur (z. B. Verbesserung der optischen Zeichenerkennung bei Fraktur; Algorithmen zur passenden Gliederung der Zeitschriften) dienten, sondern auch zur Entwicklung einer skalierbaren Datenbank beitrugen. Diese ist so konzipiert worden, dass Metadaten und Annotationen verschiedener Art mit den jeweiligen Korpora assoziiert werden können und als die Basis für Anwendungen des maschinellen Lernens verwendet werden. Bei diesen Anwendungen geht es uns besonders um eine automatisierte Auflösung von Ortsnamen (eine Form von automated toponym resolution) im Zusammenhang von Named-Entity-Recognition (NER), die die vorkommenden Ortsnamen mit hoher Präzision in großen Mengen von Texten identifizieren. Um unseren Beitrag klarer darzustellen, werden unsere Anwendungen von Methoden und Programmbibliotheken anderer Forschunggruppen (allen voran Statros et al/Kim; Wing & Baldrige; Speriosu & Baldrige; DeLosier) erläutert.

Um der historischen Spezifizität unserer Texte gerecht zu werden, werden die Ortsnamen aus einer speziell von uns zusammengestellten Datenbank von geokodierten historischen Ortsnamen gespeist (Datenquelle: Mini-Gov Datenbank 2015). Zum Schluss wird mittels eines

Open-Source-Plug-Ins der Omeka-Plattform (neatline) eine Visualisierung der geowissenschaftlichen Daten ermöglicht, die nicht nur synchronische geographische Verhältnisse zwischen Zeitschriftentext, Thema und Ort bzw. Region darstellen, sondern auch diachronische. Die Grenzen dieser Methode im Vergleich zu jenen eines GIS-Systems werden kurz besprochen. Somit wird eine solide Basis für die Möglichkeit neuen geisteswissenschaftlichen Wissens gestellt, die dann zuschließend mit ersten Ergebnissen gezeigt wird.

Bibliographie

Anderson, Benedict (2006): *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. New York: Verso.

Barth, Dieter (1975): "Das Familienblatt — ein Phänomen der Unterhaltungspresse des 19. Jahrhunderts. Beispiele zur Gründungs- und Verlagsgeschichte", in: *Archiv für Geschichte des Buchwesens* 15: cols. 205-12.

Belgum, Kirsten (1998): *Popularizing the Nation: Audience, Representation, and the Production of Identity in Die Gartenlaube, 1853-1900*. Omaha: U Nebraska P.

Daum, Andreas (2002): *Wissenschaftspopularisierung im 19. Jahrhundert: bürgerliche Kultur, naturwissenschaftliche Bildung und die deutsche Öffentlichkeit, 1848-1914*. Munich: Oldenbourg.

McIsaac, Peter (2014): "Rethinking Non-Fiction: Distant Reading the Nineteenth-Century Science-Literature Divide," in: Tatlock, Lynne / Erlin, Matt (eds.): *Distant Readings: Topologies of German Literature in the Long Nineteenth Century*. Rochester: Camden House 185-208.

Mini-Gov Datenbank (2015): *Mini-GOV (Genealogisches Ortsverzeichnis)*. Daten des Genealogischen Ortsverzeichnisses GOV , Verein für Computergenealogie e. V. http://wiki-de.genealogy.net/GOV/Mini-GOV [letzter Zugriff 28. Dezember 2015].

Sprache als Netz: Diagnostik durch Visualisierung

Meindl, Claudia
Meindl@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Rausch, Alexandre
Rausch@rz.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Bis jetzt haben sich Sprachwissenschaftler viel mit der Frage beschäftigt, wie Menschen (sprachliches) Wissen erwerben. Aber auch wie wir Wissen verlieren, erlaubt einen Einblick in die Struktur unseres Geistes. Darüber ist aber vergleichsweise wenig bekannt, obwohl es bei einer Vielzahl von Erkrankungen auch zu sprachlichen Einschränkungen kommt, die von den Betroffenen und ihren Angehörigen als sehr belastend empfunden werden. Bedenkt man bspw. die Häufigkeit dementieller Erkrankungen (11% der über 65jährigen sind betroffen), ist es umso erstaunlicher, dass sich die Geisteswissenschaften mit ihrer Expertise bis jetzt an der Entwicklung diagnostischer Verfahren kaum beteiligt haben. In der klinischen Praxis der Demenz-Diagnostik kommt dabei der Spontansprache eine große Bedeutung zu. Hier ersetzt oft aus Kostengründen ein Gespräch eine umfangreiche, testpsychologisch basierte Diagnostik mit zum Teil gravierenden Folgen für die Betroffenen.

Dabei wird vorausgesetzt, dass die Sprache den kognitiven Abbau widerspiegelt, und sich der Untersucher auf der Basis seiner Wahrnehmung ein reliables Urteil bilden kann. Unsere Wahrnehmung lässt es aber nicht zu, beim Lesen und Hören Erzählverläufe zu erfassen und auf mögliche Auffälligkeiten zu achten. Auch eine Mustererkennung ist auf diese Weise nicht möglich. Bereits nach wenigen Probanden bleibt lediglich ein Eindruck haften, und es lassen sich zudem auch schwerwiegende Wahrnehmungsfehler nachweisen (bspw. in der Einschätzung der Kohärenz).

In der sprachlichen Diagnostik werden aber auch die Grenzen eines rein kategorialen, frequenzorientierten Ansatzes schnell sichtbar. Untersucht man typische Phänomene der Spontansprache und der Textproduktion, zeigt sich, dass die dort eingesetzten Kategorien zwar theoretisch gut abgesichert sind, die Trennschärfe aber wegen der großen Varianz viel zu gering ausfällt. Und die traditionellen, regelbasierten Modelle der generativen Grammatik können zwar zur Entwicklung von Testmaterial auf der Satzebene gewinnbringend herangezogen werden, ihr Beitrag zur Analyse von Texten und Diskursen ist aber gering. Gerade den größeren Domänen wird aber eine höhere ökologische Validität zugeschrieben (der Mensch als Dialogwesen). Weigand (2003: X) fordert auch deshalb für die Linguistik, dass Erklären nicht automatisch auf das Zurückführen einer Regel gesehen werden sollte:

> „(…) Das Problem liegt nicht mehr in der Relation zwischen regelgeleiteter Kompetenz und die Regel überschreitender, chaotischer Performanz, sondern bezieht sich auf einen komplexen Zusammenhang von Ordnung und Chaos, auf Kompetenz-in-der-Performanz. Dies ist unser Gegenstand. Hier trifft sich die Linguistik mit anderen Disziplinen (…). Es gilt, sich der Komplexität zu stellen, anstatt von ihr zu abstrahieren. Orthodoxes methodologisches Denken ist zu überprüfen und gegebenenfalls aufzugeben." (Weigand 2003: X).

Eine Möglichkeit, sich dieser Komplexität zu stellen, liegt u. E. in der Anwendung der Graphentheorie auf psycholinguistische Fragestellungen und Datensätze, genauer: in der Visualisierung sprachlicher Daten. Im Rahmen einer umfangreichen Studie mit 60 Demenzpatienten und gesunden älteren Kontrollprobanden wurden unterschiedliche Datensätze in den Domänen Wort, Satz und Text/Diskurs erhoben und ausgewertet. Im Bereich Text soll an drei verschiedenen Textsorten (Wegbeschreibung, Bildergeschichte und Interview) beispielhaft gezeigt werden, wie solche sprachlichen Daten visualisiert werden können und welche Analysemöglichkeiten sich damit für den Bereich Diagnostik eröffnen. Die drei Textsorten unterscheiden sich in verschiedenen Aspekten: Wegbeschreibungen und -auskünfte sind bspw. durch eine zielorientierte Ökonomie gekennzeichnet. Untersuchungen zeigen, dass eine ideale Wegauskunft sich auf sehr wenige, gut überschaubare Lokalisationen konzentriert und Wiederholungen und unnötige Spezifikationen vermeidet. Das Erzählen einer Bildergeschichte bietet dem Sprecher schon deutlich mehr Gestaltungsmöglichkeiten, hier kann bspw. von einer source-path-goal-Strategie abgewichen werden. Oft werden solche elizitierten Geschichten auch inhaltlich stark angereichert. Damit wird auch die Aufgabe der Modellierung und Visualisierung anspruchsvoller. Interviews schließen lassen sich zwar in Frage-Antwort-Paare strukturieren, die aber oft quer zur inhaltlichen Strukturierung liegen. Außerdem muss hier zusätzlich die Gesprächsorganisation abgebildet werden können.

Mit einem graphentheoretisch basierten Ansatz eröffnet sich ein neuer Zugang im Bereich Diagnostik: Während die Kodierung einzelner Informationen nur die Aggregation einzelner Variablen zulässt (also bspw.: Kommt Proposition x häufiger vor als Proposition y; kommt Proposition x immer mit Proposition y vor etc.?), erlauben Netzwerkmatrizen eine synoptische Visualisierung und damit einen direkten Zugang zur Struktur, zum propositionalen Aufbau und bspw. auch zur Analyse von Linearisierungsstrategien. Die folgenden Abbildungen zeigen beispielhaft Visualisierungen der transkribierten Bildergeschichten, die mit dem Programm *VisuaLyzer* erzeugt wurden. Zunächst wurden die einzelnen Propositionen aufgelistet und die Matrizen aufgebaut. Die Knoten sind von links nach rechts in ihrer logischen (zeitlichen) Abfolge angeordnet. Sind mehrere Propositionen einem Zeitpunkt zugeordnet, werden sie übereinander angeordnet dargestellt. Die horizontalen Schichten des Graphen stellen die Kontexte dar, die vertikalen Schichten die Zeitpunkte. Die Nummern an den Pfeilen, also formal gesehen die gerichteten Kanten (die Bogen) des Graphen, geben die Position in der Erzählsequenz an. Unsere Wahrnehmung lässt es nicht zu, beim Lesen oder Hören Erzählverläufe zu erfassen und auf mögliche Unterschiede bei den Probandengruppen zu achten. Bereits nach wenigen Bildergeschichten bleibt lediglich ein Eindruck haften. Mit dem Graphen wird es aber möglich, die Struktur der Bildergeschichten zu visualisieren und damit analysieren zu können. Einige Beispiele sollen zeigen, wie unterschiedlich Erzählverläufe aussehen können:

Abb. 1: kreb, Alzheimer Demenz

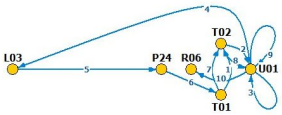


Abb. 2: oeh, Alzheimer Demenz

Die Geschichte des Alzheimer Patienten „kreb“ besteht nur aus drei Sequenzen, die kaum verknüpft sind. Bei der Abbildung 2 (Alzheimer-Patientin „oeh“) wird deutlich, dass sich diese Demenzpatientin nur auf das Ende der Geschichte konzentriert und die ersten Sequenzen nicht realisiert werden. Auch die Perseverationen werden als Loops sichtbar. Allerdings produzieren auch die gesunden Kontrollprobanden Erzählungen mit nur wenigen Propositionen und minimalem kognitiven Planungsaufwand, wie die Abbildung 3 (gesunde Kontrollpatientin „rei“) beispielhaft zeigt. Die Geschichte der gesunden Probandin „fcon“ (Abbildung 4) dagegen realisiert zwar viele Propositionen, häufige Rückverweise lassen sie aber als wenig strukturiert erscheinen.

Abb. 3: rei, Kontrollgruppe

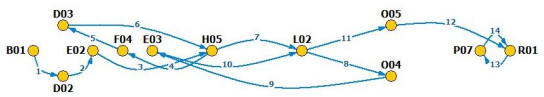


Abb. 4: fcon, Kontrollgruppe

Aggregiert man einzelne Graphen über Gruppen wird auch die Variabilität bei bestimmten Items sichtbar, was hilfreich bei der Entwicklung von diagnostischen Instrumentarien sein könnte (vgl. Abbildung 5). Für die Beschreibung des Zeitpunktes 17, der für die

Probanden das Kernereignis repräsentiert, werden 26 verschiedene Versionen (P-Knoten) gewählt, bei den Eingangssequenzen der Geschichte gibt es dagegen kaum Variation. Sichtbar wird auch, welches Weltwissen als geteilt angenommen wird.

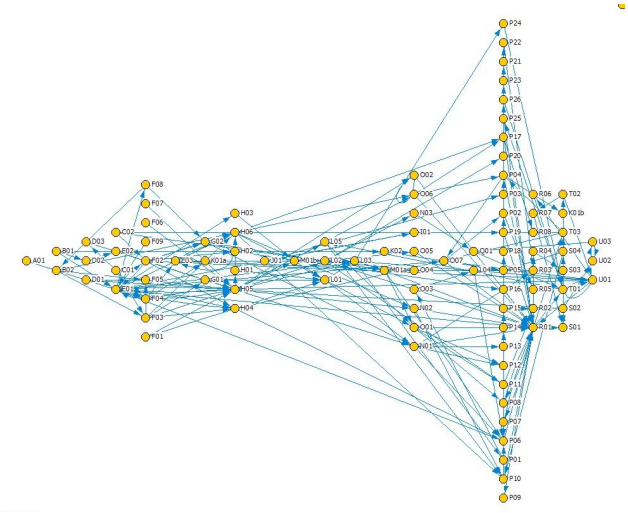


Abb. 5

Abbildung 6 zeigt abschließend beispielhaft die Visualisierung eines Interviews mit einem Alzheimer-Patienten. Hier kann man bspw. den thematischen Aufbau, die Turnorganisation, die Redeanteile oder auch die Rückverweise erkennen.

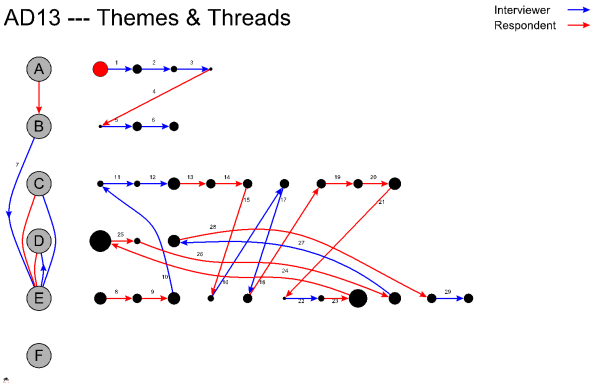


Abb. 6: Interview Alzheimer-Patient

Dieser Ansatz ist allerdings nicht für alle Datensätze geeignet. Abschließend sollen deshalb die folgenden Probleme kurz diskutiert werden: Wo könnten die diagnostischen Grenzen liegen, wenn man mit Visualisierungen in diesem Bereich arbeitet (bspw. in der Komplexität der Graphen)? Wie lässt sich eine solche, auch auf Visualisierungen basierende Diagnostik theoretisch und durch externe Kriterien empirisch absichern? Und wie kann eine Brücke geschlagen werden zwischen dem oft deduktiven Vorgehen der Psycholinguistik und dem zumeist induktiven Vorgehen der Korpuslinguistik?

Bibliographie

Weigand, Edda (2003): *Sprache als Dialog*. Sprechakttaxonomie und kommunikative Grammatik. Tübingen: Niemeyer-Verlag.

Die datengeleitete Ermittlung des gemeinsamen sprachlichen Inventars der Geisteswissenschaften

Meißner, Cordula
cordula.meissner@uni-leipzig.de
Universität Leipzig, Deutschland

Wallner, Franziska
f.wallner@rz.uni-leipzig.de
Universität Leipzig, Deutschland

Hintergrund

Sprache ist in der Wissenschaft nicht nur ein Instrument, um Sachverhalte zu vermitteln, sondern spielt für das wissenschaftliche Denken eine konstitutive Rolle. Dies gilt insbesondere für die geisteswissenschaftlichen Disziplinen, da hier selbst die Gegenstände der Forschung größtenteils sprachlich verfasst sind (vgl. Kretzenbacher 2010). Die nicht-terminologische, disziplinenübergreifend verwendete Wissenschaftssprache spiegelt dabei in besonderem Maße die in Sprache niedergelegten Erkenntnisprozesse wider und ist somit von wesentlicher wissenschaftsmethodologischer Bedeutung. Zu ihr gehören beispielsweise Ausdrucksmittel des Voraussetzens, des Begründens, des Folgerns, des Einschränkens, des Übertragens und Vergleichens. Für diesen Bereich, der unter dem Begriff der allgemeinen oder auch alltäglichen Wissenschaftssprache zusammengefasst wird (Schepping 1976; Ehlich 1999), steht eine systematische lexikographische Erschließung und Beschreibung jedoch bislang noch aus. Der einzige vorliegende Ansatz zu einer lexikografischen Erfassung der allgemeinen Wissenschaftssprache nimmt das gesamte Spektrum akademischer Fächer in den Blick und erlaubt so eine nur geringe Beschreibungsdetailliertheit (Erk 1972, 1975, 1982, 1985).

Das Projekt GeSIG (Das gemeinsame sprachliche Inventar der Geisteswissenschaften) setzt sich daher zum Ziel, erstmals das Inventar der allgemeinen Wissenschaftssprache der Geisteswissenschaften auf empirischer Grundlage zu bestimmen und damit den

Grundstein für seine umfassende Erschließung zu legen. Ein auf diese Weise bestimmtes Inventar stellt eine wertvolle Grundlage für die Dokumentation und Erforschung der Sprache der Geisteswissenschaften dar und bietet die Ausgangsbasis für die Reflexion spezifisch geisteswissenschaftlicher Erkenntnisprozesse. Das Projekt ist als Pilotprojekt angelegt und soll Vorarbeiten liefern für den Aufbau einer umfassenden elektronischen lexikographischen Ressource dieses Sprachbereichs.

Vorgehen

Das Inventar der allgemeinen Wissenschaftssprache der Geisteswissenschaften wird datengeleitet ermittelt. Die Datenbasis bilden Korpora verschiedener geisteswissenschaftlicher Fachbereiche. Zur Operationalisierung der „Geisteswissenschaften“ wird dabei die Umfangsbestimmung des Wissenschaftsrates (2010) zugrunde gelegt, der sich an die Systematik des statistischen Bundesamtes anlehnt und Fächergruppen wie Philosophie, Sprach- und Literaturwissenschaften, Geschichtswissenschaften, Regionalstudien, religionsbezogene Wissenschaften, die bekenntnisgebundenen Theologien, die Ethnologien sowie die Medien-, Kunst-, Theater- und Musikwissenschaften umfasst (vgl. Statistisches Bundesamt 2013).
Die zugehörigen Disziplinen sind in 19 Gruppen zusammengefasst, die für die Bildung von Teilkorpora herangezogen werden. Dabei werden für jeden Bereich mindestens 10 Dissertationen und mindestens 1 Mio. Token erhoben. Die Analysegrundlage bilden somit Teilkorpora in einem Gesamtumfang von ca. 19 Mio. Token.

Um einen systematischen Zugriff auf den Wortschatzbestand der allgemeinen Wissenschaftssprache der Geisteswissenschaften zu ermöglichen, werden aktuelle korpusmethodologische Werkzeuge und Erschließungsverfahren eingesetzt. Die Sprachdaten werden zunächst für die korpuslinguistische Analyse bereinigt. Um eine systematische Auswertung auf Lemmaebene und im Hinblick auf Wortarten durchzuführen, werden sie anschließend mit Hilfe des TreeTaggers (Schmid 1995) nach Wortarten annotiert sowie lemmatisiert. Dabei liegen die Richtlinien des STTS zugrunde (Schiller et al. 1999). Zusätzlich erfolgen weitere Nachbearbeitungsschritte zur Desambiguierung automatisch ermittelter Homonyme sowie zur Lemmatisierung der Partikelverben und unvollständiger Wortformen.

Auf der Grundlage der so aufbereiteten Teilkorpora wird der allgemeinwissenschaftliche Wortschatz der Geisteswissenschaften ermittelt. Dieser wird operationalisiert durch das disziplinübergreifende Vorkommen von Lemmata. Hierzu wird für jedes Teilkorpus eine Lemmaliste erstellt und eine Schnittmenge aus diesen 19 Listen gebildet.

Ergebnisse

Das allgemeinwissenschaftliche sprachliche Inventar der Geisteswissenschaften setzt sich aus den Lemmata zusammen, die in allen Teilkorpora vorkommen. Es umfasst damit jene sprachlichen Mittel, die der Form nach in geisteswissenschaftlichen Disziplinen übergreifend gebraucht werden. Die quantitative Auswertung zeigt jedoch deutliche Frequenzunterschiede für einzelne Lemmata in bestimmten Disziplinen. Dies deutet darauf hin, dass einige der übergreifend gebrauchten Lexeme in den geisteswissenschaftlichen Disziplinen einen unterschiedlichen Stellenwert haben und möglicherweise fachterminologisch geprägt sind.

Die Frequenzwerte weisen zudem darauf hin, dass einzelne Fachbereiche hinsichtlich der gebrauchten sprachlichen Mittel einander näher stehen und größere Überschneidungsmengen bilden, als andere. Nimmt man diese frequenzindizierten Ähnlichkeiten als Ausgangspunkt, ergeben sich alternative Möglichkeiten der Fachbereichsgruppierung, welche sich letztendlich auch auf Umfang und Ausprägung des zu ermittelnden gemeinsamen Inventars der Geisteswissenschaften auswirken.

Der Vortrag stellt die Ergebnisse unterschiedlicher Erschließungs- und Auswertungsverfahren gegenüber und diskutiert diese im Hinblick auf das Konzept einer allgemeinen Wissenschaftssprache der Geisteswissenschaften und ihrer lexikografischen Erfassung.

Bibliographie

Ehlich, Konrad (1999): "Alltägliche Wissenschaftssprache", in: *Informationen Deutsch als Fremdsprache* 26: 3-24.

Erk, Heinrich (1972): *Zur Lexik wissenschaftlicher Fachtexte*. Verben, Frequenz und Verwendungsweise (= Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 4). München: Hueber.

Erk, Heinrich (1975): *Zur Lexik wissenschaftlicher Fachtexte*. Verben, Frequenz und Verwendungsweise (= Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 4). München: M. Hueber.

Erk, Heinrich (1982): *Zur Lexik wissenschaftlicher Fachtexte*. Verben, Frequenz und Verwendungsweise (= Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 4). München: M. Hueber.

Erk, Heinrich (1985): *Wortfamilien in wissenschaftlichen Texten*. Ein Häufigkeitsindex (= Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 9). München: M. Hueber.

Kretzenbacher, Heinz (2010): "Fach- und Wissenschaftssprachen in den Geistes- und Sozialwissenschaften", in: Krumm, Hans-Jürgen / Fandrych, Christian / Hufeisen, Britta /

Riemer, Claudia (eds.): *Deutsch als Fremd- und Zweitsprache* (= Handbücher zur Sprach- und Kommunikationswissenschaft 35.1). Berlin, New York: de Gruyter 493-501.

Schepping, Heinz (1976): "Bemerkungen zur Didaktik der Fachsprache im Bereich des Deutschen als Fremdsprache", in: Rall, Dietrich / Schepping, Heinz / Schleyer, Walter (eds.): *Didaktik der Fachsprache*. Beiträge zu einer Arbeitstagung der RWTH Aachen vom 30.9. bis 4.10.1974. Bonn-Bad Godesberg: DAAD 13-34.

Schmid, Helmut (1995): "Improvements In Part-of-Speech Tagging With An Application To German", in: *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland [letzter Zugriff 02. Oktober 2015].

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Technischer Bericht. Universitäten Stuttgart & Tübingen.

Weibliches Erzählen im Expressionismus? Eine Stilometrie von Mela Hartwigs Prosa

Mihm, Melanie
Melanie.Mihm@germanistik.uni-giessen.de
Justus-Liebig-Universität Gießen, Deutschland

Hintergrund

„Die Stilanalyse ist eine Schlüsselqualifikation literarturwissenschaftlicher Arbeit“ (Meyer 2007: 70). So die Einführung, die man im zweiten Band *Methoden und Theorien* des Handbuchs Literaturwissenschaften erhält. Aktuelle Beiträge wie beispielsweise Christof Schöchs (2014) „Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik“ und die korpusgestützte Untersuchung von McIntyre / Walker (2010) zu ausgewählter Lyrik von William Blake sowie Drehbüchern zeigen, dass die alte, aber fortwährende Debatte aus den 1970ern zwischen „Anwälte[n] einer historisch-hermeneutischen Literaturwissenschaft“ und Vertretern einer „ahistorischen Quantifizierung von Stileigenheiten“ (Meyer 2007: 70) ausgetragen zu sein scheint, und dass im Zuge moderner Verfahren die computergestützte Stilometrie (*statistical stylistic*) als „a good helpmate“ (Tuldava 2008: 369) herangezogen werden kann. Die Gegenstände der Stilometrie sind die Charakterisierung und der Vergleich des Stils von Autoren, Gattungen und einzelnen Werken. [1] Schöch (2014: 132) weist darauf hin, dass die Stilometrie für die Literaturgeschichtsschreibung und Gattungstheorie neue Perspektiven eröffnen könnte. Ein Vorschlag, den Franco Moretti unterbreitet, lautet: „we know how to read texts, now let’s learn how *not* to read them“ (Moretti 2013 / 2000: 48). Des Weiteren führt er den Begriff „distant reading“ ein, der seit geraumer Zeit in den Digital Humanities kursiert, und erklärt: „where distance, let me repeat it, *is a condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems“ (Moretti 2013: 48-49).

Diese vorliegende Stilometrie setzt sich zum Ziel, mithilfe statistischer Analysetechniken sowohl eine quantitative als auch qualitative Stiluntersuchung von Prosa der österreichischen Autorin Mela Hartwig zu realisieren. Bislang liegen keine computergestützten Studien für diese Autorin der Zwischenkriegsjahre vor. Dies soll zum Anlass genommen werden, exemplarisch die Verwendung einer Herzmetaphorik als spezifischen Stil der Autorin zu untersuchen. Es sollen mithilfe der Stilometrie intuitive Annahmen überprüft und in Zusammenhang mit ausgewählten Prosatexten des literarischen Expressionismus gesetzt werden. Hartwig verwendet das Herz mit einer tieferen Bedeutung, sie meint nicht nur das Organ, das den Körper mit Blut versorgt. Bei den expressionistischen Texten von Hartwig kann das Herz bei der Frau beispielsweise für das Pendant des männlichen Gehirns stehen. Hierbei wird der Frau scheinbar das Vermögen des Denkens abgesprochen und an Stelle des Gehirns, also dem rationalen Vermögen des Mannes, arbeitet das Herz der Frau im Sinne eines affektiven, emotionalen, spontanen Teils des weiblichen Körpers. Es scheint auffällig, dass Hartwig ein binäres System von Herz – Verstand / Gehirn und damit einhergehend eine Gegenüberstellung von Frau – Mann herausarbeitet. Diese Dichotomie und das überwiegend weibliche konfliktgeladene Figurenrepertoire sowie die Erzähltextperspektive aus Sicht von weiblichen Figuren könnte als typisches ‚weibliches Erzählen‘ gefasst werden. Die erste Hypothese lautet, dass in Mela Hartwigs gesamter Prosa die Herzmetaphorik Verwendung findet. Die zweite Hypothese lautet, dass die Herzmetapher bei Mela Hartwig wesentlich mit Weiblichkeit verbunden ist.

Ziele des Vortrages sind das zielgruppenorientierte Präsentieren des methodischen Vorgehens und der Zwischenergebnisse, die Erläuterung der interdisziplinären Fragestellungen, Arbeitshypothesen und das in Zusammenhangsetzen der Zwischenergebnisse der Stilometrie für Prosatexte von Mela Hartwig mit dem von Moretti (2013) etablierten *distant reading*.

Vorgehen

Es wurden zwei literarische Textkorpora manuell erstellt und für das dritte Vergleichskorpus eine Kombination von zwei Korpora aus COSMAS II

[2] herangezogen. Die Kriterien für die manuelle Generierung lauten, dass (i) die Texte gemeinfrei und elektronisch verfügbar sind [3]; (ii) die Texte aus der Gattung Epik stammen; (iii) die Autoren als kanonisierte Vertreter_innen des Expressionismus gelten. Das erste Korpus umfasst Hartwigs publizierte Texte (vgl. Hartwig 2001 / 1931, 2002 / 1929; 2004 / 1982; fortan KMH). Das zweite Korpus besteht aus 122 Texten des Expressionismus (fortan KEX). Für das dritte Vergleichskorpus wird auf das Institut für Deutsche Sprache (IDS) und die Volltextdatenbank COSMAS II zurückgegriffen. [4] Allerdings wurde dieses mächtige Korpus insofern verkleinert, als dass nur das Korpus „lit-pub - Belletristik / Trivialliteratur“ mit 148 Texten und das „loz-pub - Belletristik des 20. und 21. Jahrhunderts“, das 115 Texte beinhaltet, verwendet und ausgewertet wurden. Insgesamt umfasst dieses Vergleichskorpus 263 Texte (fortan COSMAS II). Das Vorgehen der Analyse der Herzmetapher wurde in drei Schritten vollzogen: (i) Für jeden Text von Hartwig wurde eine Konkordanzliste zum Wort 'Herz' erstellt, um dann manuell zu überprüfen, ob das Wort 'Herz' als Metapher im oben eingeführten und erläuterten Verständnis verwendet wird. Ebenso wurde dies für die Vergleichskorpora überprüft. (ii) Es sollte das Hartwig-Korpus (KMH) mit dem Korpus des Expressionismus (KEX) verglichen werden, um die Signifikanz der Herzmetapher bei Hartwig mit dem *log-likelihood*-Test (G2) zu berechnen. Es galt zu prüfen, ob allgemein die Herzmetapher bei Texten ein beliebtes Stilmittel in der schöngeistigen Literatur darstellt. Ob ein Ergebnis nun als signifikant eingestuft werden kann, hängt von der Höhe des G²-Wertes ab. Als Signifikanzniveau wurden 5 % angenommen. Wurde ein G²-Wert erreicht, der den kritischen Wert von 3,84 oder größer ausgibt, dann wurde dies als signifikant gewertet.

Ergebnisse

In dem Korpus von Mela Hartwig (KMH) kommt die Herzmetapher in neun von elf Texten vor. Die Auswertung des Korpus mit den expressionistischen Texten ergab, dass von 122 Texten 37 eine Herzmetapher aufweisen. Die Ergebnisse der Berechnung zeigen, dass Hartwig in 81,82 % ihrer publizierten Texte die Herzmetapher verwendet. Im Falle der expressionistischen Texte ist berechnet worden, dass 30,33 % der Prosatexte eine ähnliche Herzmetapher nachweisen. Da der G²-Wert 5,78 beträgt und der zuvor festgelegte kritische Wert bei 3,84 oder größer liegt (bei einem Signifikanzniveau von 5 %), ist die Herzmetapher bei Hartwigs Texten gegenüber den expressionistischen Texten im Korpus signifikant. Auch gegenüber dem größeren, epochenübergreifenden COSMAS II-Korpus tritt die Herzmetapher signifikant häufig auf: Die Berechnung ergab einen G²-Wert von 8,42 und übersteigt den kritischen Punkt. Dieser G²-Wert bestätigt meine zu Beginn dieser Arbeit geäußerte Vermutung, dass es sich bei der Herzmetapher, wie sie Mela Hartwig verwendet, um eine Besonderheit ihres Schreibstils handelt. Betrachtet man den niedrigen G²-Wert von 0,97, der bei der Berechnung der expressionistischen Texte (KEX) und COSMAS II herauskommt, fällt auf, dass hier eine ähnliche prozentuale Verteilung der Herzmetapher (KEX = 30,33 % und COSMAS II = 24,71 %) vorliegt. Der Wert unterschreitet den festgelegten kritischen Wert von 3,84. Dennoch sei darauf aufmerksam gemacht, dass in den Texten des Expressionismus ca. 6 % häufiger die Herzmetapher vorkommt.

Für die zweite Hypothese, dass Mela Hartwig die Herzmetapher nur in Texten verwendet, bei der sie eine weibliche Erzähltextperspektive und einen weiblichen autodiegetischen Erzähler einsetzt, wurde ein Konkordanz-Plot im Barcode-Format für das Personalpronomen 'ich' erstellt. Je dunkler der Barcode-Streifen des Konkordanz-Plots ausfällt, desto häufiger tritt an der schwarzen Stelle das angesteuerte Wort der Suchanfrage auf. Das Ergebnis der Visualisierung zeigt, dass das Personalpronomen in den Texten, in denen die Herzmetapher vorkommt, sehr viel häufiger verwendet wird. In Texten, in denen die Herzmetapher nicht vorkommt, findet sich das Pronomen, wenn überhaupt, nur in der wörtlichen Rede.

Fazit

Diese Stilometrie brachte das Zwischenergebnis zu Tage, dass man durch die Kombination von *distant* und *close reading* und computergestützten Verfahren dazu in der Lage ist, stilistische Alleinstellungsmerkmale in Prosatexten herausfiltern und visualisieren zu können. Für die Prosatexte der österreichischen Autorin der Zwischenkriegsjahre Mela Hartwig (vgl. KMH) konnte so berechnet werden, dass die Herzmetaphorik signifikant häufig im Vergleich mit anderen literarischen Texten aus den Vergleichskorpora (vgl. KEX und COSMAS II) verwendet wird. Dieses Ergebnis kann im Hinblick auf die Interpretation der weiblichen Erzählstruktur und der sprachlichen Darstellung von Weiblichkeit in Hartwigs Prosa herangezogen werden. Zusätzlich wurde mithilfe eines Visualisierungstools das Personalpronomen 'ich' untersucht und dadurch als ein Indikator für einen weiblichen autodiegetischen Erzähler in Hartwigs Prosa ausgemacht. Bei der Visualisierung im Barcode-Format sowie bei der Ermittlung der G²- Werte soll deutlich gemacht werden, dass es hierbei immer noch einer manuellen Überprüfung, Auswertung und Interpretation der Werte und schließlich der einzelnen Textstellen bedarf.

Ausblick

Es werden weitere korpusbasierte Analysen angestrebt. Diese erfordern aber eine Voraussetzung: Im Zuge der „digitale[n] Wende“ (Schöch 2014: 130) ist es weiterhin wünschenswert und erforderlich, dass immer mehr literarische Texte digital zur Verfügung stehen oder diese durch leichte und praktikable Verfahren der Texterkennung (OCR, *optical character recognition*) digitalisierbar gemacht werden können. Für den literarischen Expressionismus müssten dafür noch vorbereitende Maßnahmen getroffen werden. Es wäre interessant, ein weiteres Korpus für die Gattung Lyrik zu erstellen und über Konkordanzlisten beispielsweise das Wort 'Herz' anzusteuern. Sowohl für solch ein Korpus der Lyrik als auch für die bereits erstellten Korpora der Prosatexte ergeben sich weiterführende Fragen: Ist die Herzmetaphorik bei Hartwig auch weiterhin signifikant, wenn lediglich Texte untersucht werden, die nicht nur von weiblichen Autoren geschrieben wurden, sondern auch aus weiblicher Erzählperspektive dargestellt werden? Lässt sich eine Epoche festmachen, oder handelt es sich hierbei um ein epochenübergreifendes und genderunabhängiges sprachlich-stilistisches Phänomen? Könnte man eine Parallele ziehen und von „weiblichem Erzählen“ sprechen, wie es Gabriele Otto (2009) für die Nachkriegsliteratur von Ingeborg Bachmann konstatiert? Dabei soll generell betrachtet werden wie ‚die Frau‘ literarisiert und versprachlicht wird.

Notes

1. Die historische Entwicklung von stilometrischen Prinzipien lassen sich bei Schöch (2014: 133-135) nachlesen.
2. COSMAS II steht für ‚Corpus Search, Management und Analysis System‘. Es ist das Nachfolgesystem von COSMAS I (1991-2003). Das kostenlose und für wissenschaftliche und nichtkommerzielle Zwecke konzipierte Deutsche Referenzkorpus umfasst mehr als 3,9 Mrd. Wörter.
3. Recherchiert über Gutenberg-DE-Projekt, die Volltextbibliothek Zeno.org, das Deutsche Textarchiv (DTA) sowie im Internet Archive.
4. Es beinhaltet Zeitungsartikel, Sach-, Fach- und schöngeistige Literatur aus Deutschland, Österreich und der Schweiz.

Bibliographie

Institut für Deutsche Sprache (2015): *COSMAS II* http://www.ids-mannheim.de/cosmas2/uebersicht.html [letzter Zugriff 22. Dezember 2015].

Berlin-Brandenburgische Akademie der Wissenschaften (2007-2015): *Deutsches Textarchiv* (DTA) http://www.deutschestextarchiv.de/ [letzter Zugriff 22. Dezember 2015].

Gutenberg-DE-Projekt http://gutenberg.spiegel.de/ [letzter Zugriff 22. Dezember 2015].

Hartwig, Mela (2001 / 1931): *Bin ich ein überflüssiger Mensch?* Wien: Droschl Graz Verlag.

Hartwig, Mela (2002 / 1929): *Das Weib ist ein Nichts*. Wien: Droschl Graz Verlag.

Hartwig, Mela (2004 / 1928): "Das Verbrechen", "Der phantastische Paragraph","Aufzeichnungen einer Häßlichen", "Die Hexe", "Das Kind", "Die Kündigung", "Der Meineid", "Das Wunder von Ulm", "Georgslegende", in: Hartwig, Mela: *Das Verbrechen*. Novellen und Erzählungen. Wien: Droschl Graz Verlag.

Internet Archive: https://archive.org/ [letzter Zugriff 22. Dezember 2015].

McIntyre, Dan / Walker, Brian (2010): "How can corpora be used to explore the language of poetry and drama?", in: O#Keeffe, Anne / McCarthy, Michael (eds.): *The Routledge Handbook of Corpus Linguistics*. London: Routledge 516-530.

Meyer, Urs (2007): "Stilanalyse", in: Anz, Thomas (ed.): *Handbuch Literaturwissenschaft*. Band 2: Methoden und Theorien. Stuttgart / Weimar: J. B. Metzler 70-81.

Moretti, Franco (2013 / 2000): "Conjectures on World Literature", in: Moretti, Franco: *Distant Reading*. London: Verso 43-62.

Otto, Gabriele E. (2009): *Weibliches Erzählen?* Entwicklung der Erzählverfahren in Ingeborg Bachmanns Prosa. Würzburg: Königshausen & Neumann.

Schöch, Christof (2014): "Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik", in: Schöch, Christof / Schneider, Lars (eds.): *Literaturwissenschaft im digitalen Medienwandel* (= Philologie im Netz Beiheft 7) 130-157 http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf [letzter Zugriff 25. Dezember 2015].

Tuldava, Juhan (2008): "Stylistics, author identification (Stilistik und Autorenbestimmung)", in: Köhler, Reinhard / Altmann, Gabriel / Piotrowski, G. Rajm (eds.): *Quantitative Linguistik / Quantitative Linguistics*. Ein internationales Handbuch / An International Handbook. Berlin / New York: de Gruyter 368-387.

Software-Einsatz in der geisteswissenschaftlichen Forschungspraxis: Ergebnisse einer Umfrage

Müller-Birn, Claudia
clmb@inf.fu-berlin.de
Freie Universität Berlin, Deutschland

Schlegel, Alexa
alexa.schlegel@inf.fu-berlin.de
Freie Universität Berlin, Deutschland

Baillot, Anne
anne.baillot@hu-berlin.de
Humboldt-Universität zu Berlin

Klawitter, Jana
klawitter@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Zielstellung

Mit unserem Vortrag wollen wir folgende Beiträge für die Digital Humanities leisten:

1. Vorstellung und kritische Reflexion der Forschungsergebnisse anhand der abgeschlossenen Umfrage für Berlin/Brandenburg: Hier werden wir vor allem aufzeigen, zu welchen Ergebnissen wir mit unserer Umfrage als Basis für unsere umfassender angelegte Studie gelangt sind und welche Anforderungen an zukünftige Forschungsarbeiten in den Geisteswissenschaften und der Informatik sich aus der gesamten Studie ableiten lassen können.

2. Reflexion über die Nutzung der TaDiRAH-Taxonomie als methodische Basis für die Umfrage: In unserem Vortrag werden wir unsere Probleme aufzeigen, die bei der Nutzung der Taxonomie im Rahmen des Fragebogens entstanden sind. Diese Probleme sind einerseits auf die Verwendung bestimmter Fachbegriffe aber auch auf eine teilweise fehlende systematische Abgrenzung zwischen Forschungstätigkeiten zurückzuführen. Wir werden konkrete Vorschläge für die Verbesserung der Taxonomie vorstellen.

3. Reflexion über den Einsatz einer Umfrage als Forschungsinstrument: In diesem Teil des Vortrags werden wir vor allem auf statistische Anforderungen, die aus der Nutzung einer solchen Forschungsmethode erwachsen, eingehen und zeigen, inwiefern unsere Ergebnisse verallgemeinerbar sind.

Methodischer Ansatz

Die Umfrage-basierte Studie orientiert sich an folgendem Vorgehen (Müller et al. 2014): (1) Definition der Forschungsziele, (2) Bestimmung der Zielgruppe und der möglichen Stichprobe, (3) Spezifizierung des Fragenbogendesigns, (4) Überprüfung und Pre-Tests, (5) Umsetzung und Einführung sowie (6) Datenanalyse. Wir werden auf diese einzelnen Schritte im Folgenden kurz eingehen.

Definition der Forschungsziele

Der Anlass für diese Studie ist der Sammelband *#bbdh – Berliner Beiträge zu den Digital Humanities*, welcher vom Einstein-Zirkel Digital Humanities im Januar 2016 veröffentlicht wird. Der darin erscheinende Beitrag „Forschungspraxis in den Geisteswissenschaften oder wieviel Digital Humanities gibt es in den Geisteswissenschaften?“ (Müller-Birn et al. im Druck) liefert einen Überblick über den Einsatz von Software in der Forschungspraxis in den Geisteswissenschaften. Wir wollen dabei unter anderem folgende Fragen beantworten:

- In welchem Umfang wird Software eingesetzt?
- Inwiefern gibt es einen Zusammenhang zwischen dem Softwareeinsatz und dem Forschungskontext?
- Welche Software wird eingesetzt und welche typischen Softwarenutzungsmuster lassen sich identifizieren?

Darüber hinaus soll die Studie einen Beitrag für den Forschungsbereich E-Research leisten. Durch den Einsatz von Software ändert sich, wie Wissenschaftler_innen forschen. Besonders in den Geisteswissenschaften scheint diesbezüglich ein Wandel stattzufinden. Unser Ziel ist es, besser zu verstehen, welche Anforderungen an zukünftige Software gestellt werden und wie die Informatik Forscher_innen in ihren Wissensschaffungsprozessen besser unterstützen kann. Daher sind wir explizit auch an Umfrageteilnehmer_innen interessiert, die noch keine oder wenig über die gängigen Office-Programme hinausgehende Software in ihrer persönlichen Forschungspraxis einsetzen.

Bestimmung der Zielgruppe

Wir haben eine geographische Beschränkung vorgenommen, indem der Fragebogen vor allem an Wissenschaftler_innen in den geisteswissenschaftlichen Disziplinen in Berlin und Brandenburg gerichtet ist.

Spezifizierung des Fragenbogendesigns

Eine große Herausforderung war es, den Begriff der Forschungspraxis näher zu definieren. Hierfür haben wir bestehende Fragenbögen zu Arbeitspraktiken im Bereich DH analysiert. Diese überwiegend im angloamerikanischen Raum durchgeführten Studien konzentrieren sich vor allem auf text-zentrierte Forschungsmethoden und zeigen die zunehmende Durchdringung von digitalen Methoden in der textbezogenen (insbesondere philologischen, linguistischen und historischen) Quellenarbeit (vgl.

Unsworth 2000; Houghton et al. 2004; Toms / O'Brien 2008; Ceccarelli et al. 2011; Kemman et al. 2014). Als Studien für den deutschsprachigen Raum wurden Burghardt et al. (2015) und Stiller et al. (2015) herangezogen. In der Auswertung dieser Studien stellte sich heraus, dass die Begrifflichkeiten variieren bzw. sehr unterschiedlich verwendet wurden, wodurch die Ergebnisse nicht oder nur schwer vergleichbar waren.

Daher haben wir uns entschieden, die Forschungsaktivitäten aus TaDiRAH (Taxonomy of Digital Research Activities in the Humanities) zugrunde zu legen (Perkins et al. 2014), welche unter anderem auf Arbeiten von Unsworth (2000) sowie Gasteiner und Haber (2010) basiert. Die in TaDiRAH definierten Forschungsaktivitäten (research activities) sind in Unteraktivitäten unterteilt. Wir haben diese als Tätigkeiten bezeichnet. Jeder einzelnen Tätigkeit wurde in einem ersten Schritt die jeweils für die Tätigkeit geeignete Software zugeordnet.

Die einzelnen Konzepte wurden wie folgt in der Umfrage zusammengeführt:

Forschungsaktivitäten sind alle Aktivitäten, die auf die Untersuchung von bestimmten Phänomenen in der Forschungsarbeit abzielen. Solche Phänomene können beispielsweise die Pariser Jahrhundertwende oder dramatische Strukturen im elisabethanischen Theater sein. Um solche Phänomene näher zu untersuchen, werden beispielsweise textuelle, bildliche oder vertonte Quellen genutzt. Diese Quellen werden mithilfe bestimmter Tätigkeiten bearbeitet, die wiederum durch Software unterstützt werden.

Der Umfrageteilnehmer bzw. die Umfrageteilnehmerin wird dabei zunächst nach einer Priorisierung der Aktivitäten befragt und dann nach der Häufigkeit der Anwendung ausgewählter Tätigkeiten sowie der zugehörigen Software. Darüber hinaus werden eine Reihe weiterer Fragen zur Zusammenarbeit mit anderen Geisteswissenschaftler_innen gestellt sowie demographische Angaben abgefragt.

Überprüfung und Vortesten

Im Rahmen der Fragebogenerstellung wurde extensives Pre-Testing betrieben. Es wurden 15 Wissenschaftler_innen aus dem Bereich der Geisteswissenschaften gewonnen, eine Word-Version des Fragebogens zu begutachten. In drei Iterationen wurde der Fragebogen immer weiter adaptiert und vor allem die Wortwahl an die Zielgruppe angepasst.

Des Weiteren wurde die Übersetzung von TaDiRAH immer weiter geschärft und bestehende Inkonsistenzen in der Beschreibung entfernt. Neben der Überprüfung des Fragebogens durch die Zielgruppe wurde ebenfalls eine Evaluation des Fragebogens aus statistischen Gesichtspunkten durchgeführt, wodurch weitere Anpassungen in der Art der Fragestellung erforderlich waren.

Umsetzung und Einführung

Die Umfrage wurde mit Hilfe der Software Questback / Unipark erstellt und war vom 17. September 2015 bis zum 2. November 2015 aktiv.

Weitere Informationen zur Online-Umfrage und der gesamten geplanten Studie sind auf der Webseite verfügbar: https://practices4humanities.wordpress.com/ (Müller-Birn 2015).

Deskriptive Datenanalyse

Mit dem Stand Dezember 2015 wurde der Fragebogen von 270 Personen beantwortet, darunter von über 100 Wissenschaftler_innen (N = 123) aus Berlin / Brandenburg. Nach Abschluss der Umfragephase werden die erhobenen Daten analysiert. Aus den Umfrageergebnissen konnten wir fünf typische Softwarenutzungsmuster ableiten, die im Kontext bestehender Literatur diskutiert werden: 1) Word+, 2) Suchmaschine, 3) Annotationen, 4) Social Media und 5) Basissoftware. Drei Nutzungsmuster beschreiben Personen und ihre Forschungspraxis mit dem Einsatz unterschiedlicher Software, zwei Nutzungsmuster zeigen einen eher selektiven Einsatz von Software. Diese Nutzungsmuster werden genauer beschrieben und ihre Bedeutung analysiert.

Ausblick aus DH-Perspektive

Als Referenzrahmen bietet diese Umfrage die Möglichkeit, in regelmäßigen Abständen und an unterschiedlichen Orten Entwicklungen aufzunehmen, zu analysieren und zu begleiten. Eine derart gestaltete, breitere Umsetzung ist im Rahmen von DARIAH-EU anvisiert (Anne Baillot in Zusammenarbeit mit Laurent Romary).

Die Umfragedaten liefern eine solide Basis für Antworten auf so zentrale Fragen der europäischen Digital Humanities. Dies betrifft an erster Stelle die Verortung des Bedarfs der Forscher_innen im Hinblick auf ihre Arbeitsprozesse und die Software, die sie täglich benutzen. Die Umfrageergebnisse spiegeln ebenfalls die Selbstwahrnehmung der wissenschaftlichen Gemeinschaft als digital forschend wider und leisten in diesem Sinne einen Beitrag zur reflexiven Entwicklung digitaler Methoden in den europäischen Geisteswissenschaften.

Ausblick aus informatischer Perspektive

Die Umfrage ist Teil einer umfassenderen Studie, die uns über die wissenschaftliche Arbeit und die

bestehenden Praktiken Aufschluss gibt und konkretisiert, wie die Forschungsarbeit in den Geisteswissenschaften durch Softwareanwendungen unterstützt wird und innerhalb welcher Kontexte Technologien eingebettet sind. Die Erkenntnisse aus der Umfrage bilden die Basis für Interviews, die im Rahmen von sogenannten Arbeitsplatzstudien durchgeführt werden. Die Frage, ob der Umfrageteilnehmer bzw. die Umfrageteilnehmerin auch für ein Interview zur Verfügung stehen würde, haben über 40 Personen in der vollständigen Stichprobe (N = 270) positiv beantwortet und ihre E-Mail-Adresse angegeben. Ziel ist es herauszufinden, wie sich ausgewählte Technologien und Infrastrukturen in die tagtägliche Arbeit von Forscher_innen einbetten. Dabei soll bewusst die angemahnte Ingenieursperspektive auf Software überwunden werden (Fuller 2008) und Software dahingehend untersucht werden, wie Forscher_innen diese für ihre epistemologische Prozesse und ihrer tagtäglichen Forschungspraxis verwenden (Berry 2011). Die dabei gesammelten Einsichten sollen vor allem für die (Weiter-)Entwicklung von Software für die geisteswissenschaftliche Forschungspraxis genutzt werden.

Bibliographie

Berry, David M. (2011): „The computational turn: Thinking about the digital humanities“, in: *Culture Machine* 12: 1-22.

Burghardt, Manuel / Wolff, Christian / Womser-Hacker, Christa (2015): „Informationswissenschaft und Digital Humanities“, in: *Information – Wissenschaft & Praxis* 66, 5-6: 287–294.

Ceccarelli, Diego / Gordea, Sergiu / Lucchese, Claudio / Nardini, Franco Maria / Tolomei, Gabriele (2011): „Improving Europeana Search Experience Using Query Logs“, in: *Proceedings oft he 15th International Conference on Theory and Practice of Digital Libraries Research and Advanced Technology for Digital Libraries* 384-395.

Fuller, Matthew (2008): *Software studies.* A lexicon. MIT Press.

Gasteiner, Martin / Haber, Peter (eds.) (2010): *Digitale Arbeitstechniken fu#r die Geistes- und Kulturwissenschaften*. Vienna: UTB.

Houghton, John W. / Steele, Colin / Margaret Henty: (2004): „Research practices and scholarly communication in the digital environment“, in: Grewal, Bhajan S. / Kumnick, Margarita (eds.): *Engaging the New World*. Responses to the Knowledge Economy. Melbourne: Melbourne University Publishing 169-203.

Kemman, Max / Kleppe, Martijn / Scagliola Stef (2014): „Just Google It - Digital Research Practices of Humanities Scholars“, in: *Proceedings of the Digital Humanities Congress*.

Müller, Hendrik / Sedley, Aaron / Elizabeth Ferrall-Nunge (2014): "Survey research in HCI", in: Olson, Judith S. / Kellogg, Wendy A. (eds.): *Ways of Knowing in HCI*. New York: Springer 229-266.

Müller-Birn, Claudia (2015): *practices4humanities*. Wissenschaftliche Forschungspraxis in den Geisteswissenschaften https://practices4humanities.wordpress.com/ [letzter Zugriff 09. Februar 2016].

Müller-Birn, Claudia / Schlegel, Alexa / Baillot, Anne / Klawitter, Jana (im Druck): „Forschungspraxis in den Geisteswissenschaften oder wie digital sind die Geisteswissenschaften?", in: Baillot, Anne / Schnöpf, Markus (eds.): #bbdh – Berliner Beiträge zu den Digital Humanities. Berlin.

Perkins, Jody / Dombrowski, Quinn / Borek, Luise / Schöch, Christof (2014): „Building bridges to the future of a distributed network: From DiRT categories to TaDiRAH, a methods taxonomy for digital humanities", in: *Proceedings of the International Conference on Dublin Core and Metadata Applications*.

Stiller, Juliane / Thoden, Klaus / Leganovic, Oona / Heise, Christian / Höckendorff, Mareike / Gnadt, Timo (2015): *Nutzungsverhalten in den Digital Humanities*. Technical Report. DARIAH-DE Projektdokumentation R 1.2.1/ M 7.6. https://wiki.de.dariah.eu/display/publicde/Reports+and+Milestones [letzter Zugriff 09. Februar 2016].

Toms, Elaine G. / O'Brien, Heather L. (2008) „Understanding the information and communication technology needs of the e-humanist“, in: *Journal of Documentation* 64, 1: 102–130.

Unsworth, John (2000): "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common and How Might Our Tools Reflect This?", in: *Humanities Computing, Formal Methods, Experimental Practice*. Symposium, Kings College, London.

HistStadt4D - Multimodale Zugänge zu historischen Bildrepositorien zur Unterstützung stadt- und baugeschichtlicher Forschung und Vermittlung

Münster, Sander
sander.muenster@tu-dresden.de
Medienzentrum/TU Dresden, Deutschland

Niebling, Florian
florian.niebling@uni-wuerzburg.de
Lehrstuhl HCI/Universität Würzburg, Deutschland

Nutzungsszenarien und Forschungsmehrwerte

Ähnlich virtuellen 3D-Modellen adressieren auch digitale Bildrepositorien einen breiten Nutzerkreis mit höchst unterschiedlichen Anforderungen, welche von einer Forschungsunterstützung und damit verbundenen geistes- und informationswissenschaftlichen Fragestellungen über eine Wissensvermittlung in akademischen und musealen Kontexten bis hin zu touristischen Anwendungen reichen (Münster 2011a). In Abhängigkeit vom Nutzerkreis existieren dabei eine Reihe teilweise gegensätzlicher Anforderungen: Für geschichtswissenschaftliche Forschungsaufgaben stehen beispielsweise Aspekte einer Vergleich- und Kontextualisierbarkeit von Quellen (Wohlfeil 1986; Brandt 2012; Münster et al. 2015) oder des Bezugs zwischen Quelle und Repräsentation (Favro 2004; Niccolucci / Hermon 2006) ebenso wie eine Identifikation beispielsweise von formalen Mustern, Singularitäten, Brüchen und Devianzen in Architektur und Stadtbild (Andersen 2007; Bürger 2011) im Vordergrund. Zu den damit verbundenen geschichts- und kulturwissenschaftlichen Fragestellungen gehören beispielsweise:

- Wie verändern sich Bauten und Städte im Laufe der Zeit?
- Welche zeitlichen und örtlichen Zäsuren und Brüche lassen sich erkennen?
- In welchen Kontexten findet die Transformation eines historischen Stadtbilds statt?
- Welche baukulturellen Standards und Notwendigkeiten, Bauvorschriften sowie Zeit-, Regional- und Personalstile und Singularitäten von historischen Objekten lassen sich erkennen?
- Welche Ansichten weisen die historisch höchste Überlieferungsdichte auf?
- Welche Blickwinkel lassen auf eine bestimmte Betrachterwirkung und ein abhängiges Betrachterverhalten schließen?
- Wie wird durch die Motive der Ansichten bzw. durch die Ansichten selbst regionale Identität erzeugt und transportiert?

Demgegenüber sind für eine Vermittlung geschichtswissenschaftlichen Wissens - neben der Attraktion von Aufmerksamkeit beispielsweise durch spielerische Elemente (Jacobson et al. 2009) - dessen mentale Anknüpfbarkeit und Verständlichkeit relevant. Mit Blick auf eine Nutzung sind zwei wesentliche Vorgehensweisen der Informationserschließung erkennbar: Einerseits ein selbstgesteuertes Durchsuchen von Sammlungen historischer Fotografien, Zeichnungen und Pläne, andererseits eine orts- oder kontextbezogene Informationsvermittlung beispielsweise im Zuge stadträumlicher oder musealer Präsentation. Mit Blick auf diesen letzten Aspekt hat insbesondere mit dem Aufkommen leistungsfähiger personal devices wie Smartphones oder Tablets eine Vor-Ort-Darstellung geschichtswissenschaftlicher Information als Augmented Reality Bedeutung gewonnen und wurde vielfältig erprobt und untersucht (Livingston et al. 2008; Zöllner et al. 2010; Walczak 2011). Dabei stehen umfassende Untersuchungen eines derartigen interaktiven Zugangs zu großen und heterogenen historischen Medienrepositorien sowohl aus technischer und gestalterischer Sicht als auch mit Blick auf Wissenstransfers und Lerneffekte jedoch noch aus.

Klassifikation von Bildern

Mit Blick auf eine Bildklassifikation gehen bisherige Ansätze vor allem von einer – durch Experten oder eine Nutzercommunity – vorgenommenen Verschlagwortung von Bildern aus. Insbesondere bei fotografischen Aufnahmen lassen sich über derartige manuelle Klassifikationen hinaus technische Ansätze einer Erkennung von Bild-Features oder Bildkompositionen (Hanzl / Káňa 2012) anwenden und daraus beispielsweise Ähnlichkeiten von Motiven ableiten (Hoiem / Savarese 2011; Endres 2013). Analog nutzen photogrammetrische Verfahren die Möglichkeit, anhand von fotografischen Aufnahmen von unterschiedlichen Standpunkten mit überlappenden Bildinhalten räumlich dreidimensionale Strukturen abzuleiten (Pierrot-Deseilligny et al. 2011; Kersten et al. 2012). Gerade umfangreiche Fotorepositorien ermöglichen die Erzeugung komplexer dreidimensionaler Landschafts- und Stadtmodelle, welche unter Nutzung bekannter ortsbezogener Objekte wie Bauwerke oder Straßenverläufe im Weltkoordinatensystem verortet werden können. Damit erlauben derartige 3D-Modelle wiederum einen ortsbezogenen und intuitiven Zugriff auf einzelne Bilder. Während solche Ansätze für zeitgenössische Fotografien erprobt sind und vielfältig Anwendung finden (Structure-from-Motion, z. B. Virtual Rome, MS Photosynth), stellt sich bei historischem Bildmaterial nicht nur das Problem höchst unterschiedlicher technischer Qualitäten und einer mangelnden Reproduzierbarkeit der Aufnahmesituation (Brenningmeyer / Begg 2006; Stojakovic / Tepavcevic 2009) sondern auch die Anforderung einer zeitlichen Dimensionierung der Aufnahmen. Die Forderung nach einer Auslotung derartiger Ansätze im Dienste der Humanities stammt dabei insbesondere aus den Reihen der Archäologie und im Kontext kulturellen Erbes (Ioannides et al. 2013). Aus diesen Bereichen stammen ebenfalls umfangreiche Vorarbeiten insbesondere zu technischen Algorithmen und Vorgehensweisen, welche jedoch zumeist anhand zeitgenössischer Aufnahmen oder anhand spezifisch ausgewählter historischer Bilddatenbestände erprobt wurden (Ioannides et al. 2013).

4D- Informationssysteme

Durch die Verknüpfung von Orts- und Zeitbezügen entstehen Informationssysteme, in welche sich neben Fotografien eine Vielzahl weiterer, ortsbezogener Daten integrieren lassen. Gerade im deutschsprachigen Raum fokussieren derartige Systeme zumeist auf eine räumlich-zweidimensionale, zeitbezogene Kartierung historischer Artefakte sowie damit verbundene Relations- und Aggregationsinformationen. Dies spiegelt sich nicht nur in einer Vielzahl von Projekten sondern auch im als diesbezügliches Infrastrukturangebot entwickelten Europeana 4D Interface wider, welches die Basis des Dariah-Geobrowsers bildet und trotz seines Namens auf eine vorrangig zweidimensionale Kartierung abzielt. Darüber hinausgehend eröffnen perspektivisch korrekte Darstellungen dreidimensionaler Daten, beispielsweise als virtuelle Stadt- und Landschaftsmodelle, gerade mit Blick auf die Verknüpfung und Veranschaulichung komplexer historischer Informationen gegenüber Kartierungen eine Reihe von Möglichkeiten (Prechtel et al. 2013).

Die BMBF-eHumanities-Nachwuchsgruppe HistStadt4D

Die vorgeschlagene Präsentation stellt die konzeptionellen sowie empirischen Vorarbeiten der schon angeführten eHumanities-Nachwuchsgruppe HistStadt4D vor.

Die Nachwuchsgruppe adressiert mit Blick auf die beschriebene Problemstellung drei Fragenkomplexe. Ein geschichtlich-architektonischer Komplex behandelt am Beispiel der baugeschichtlichen Entwicklung der Stadt Dresden im 20. Jahrhundert Fragen, deren Bearbeitung mit einer intensiven Nutzung von Bild- und Planquellen verbunden ist und für welche eine Zusammenführung und technische Unterstützung Forschungs- und Vermittlungsmehrwerte verspricht. Dies umfasst Aspekte zeitlicher Entwicklung und Transformation einer Stadtlandschaft und formaler Stile ebenso wie kulturelle Aspekte wie beispielsweise ein fotografisches Dokumentationsverhalten und dessen geschichtsbildende Wirkung aber auch spezifische Fragen wie die nach einer Rekonstruktion des Aufnahmestandpunktes von historischen Fotografien.

Damit verknüpft ist ein zweiter, methodischer Komplex, welcher diesbezüglich forschungsmethodische Anforderungen an digitale Bild- und Planquellenrepositorien und sich daraus ableitenden technische Unterstützungsoptionen behandelt. Dazu gehören die Systematisierung von Unterstützungsbedarfen beispielsweise einer Identifikation und Kontextualisierung sowie eines visuellen Vergleichs zwischen derartigen Quellen und die Entwicklung von Nutzungsszenarien.

Darauf aufbauend behandelt ein informationell-technischer Komplex eine bedarfsgerechte Informationsmodellierung und deren technische Umsetzung am Beispiel der Deutschen Fotothek . Dazu gehören Aspekte einer Prozessierung und Verknüpfung von historischen Medien- und Wissensbeständen unter Einbeziehung von Ort und Zeit zu einer virtuellen Forschungsumgebung, sowie die Untersuchung und Entwicklung von Visualisierungs- und Informationszugängen mittels 4D-Browser sowie als interaktive ortsbezogene Augmented Reality.

Präsentationsinhalte

Die vorgestellte Nachwuchsforschergruppe HistStadt4D befindet sich momentan in der Vorbereitungsphase. Zum Zeitpunkt der Konferenz werden präsentierbar sein:

- Die im Rahmen zweier beginnender Habilitationsvorhaben erarbeiteten Darlegungen der Forschungsstände (1) zu Forschungsansätzen sowie Arbeitstechniken der Visual Humanities (Münster / Prechtel 2014; Münster et al. 2015; Münster in Vorbereitung) sowie (2) zur Visualisierung umfangreicher Bildrepositorien.
- Ein darauf aufbauendes Forschungskonzept der Nachwuchsgruppe.

Darüber hinaus liegen zum Zeitpunkt der Konferenz Ergebnisse im Kontext der Nachwuchsgruppe entstandener flankierender Forschungsarbeiten vor und sollen in eine Präsentation einfließen:

- In einer Diplomarbeit gewonnene und getestete Erkenntnisse zur Abstraktion und Wiedererkennbarkeit virtueller Gebäuderepräsentationen (Weller 2013; Münster et al. submitted paper).
- Eine im Rahmen einer Masterarbeit ab Mitte 2015 erfolgte wissenschaftlich fundierte Bewertung und Konzeption sowie prototypische Testung von technischen Unterstützungsoptionen stadträumlicher Forschung und Vermittlung.
- Die im Rahmen einer weiteren Masterarbeit ab Mitte 2015 erfolgte Untersuchung der dreidimensionalen Verortung historischer Fotografien unter Zuhilfenahme eines fotogrammetrisch erstellten zeitgenössischen 3D-Modells.
- Mit Blick auf eine Bandbreite der geschichtswissenschaftlichen Nutzung von Medienrepositorien erfolgte im Rahmen einer Dissertation (Münster 2014) anhand einer Analyse von 578 - in erweiterter Form knapp 3000 - internationalen Konferenzbeiträgen der Spatial Humanities eine umfassende Aufstellung sowohl

von Projekttypen und Objekten (Münster / Köhler 2012) als auch von verwendeten Technologien sowie Forschungs- und Vermittlungskontexten (Münster 2011; Münster / Ioannides 2015; Münster et al. im Druck).

- Im Kontext einer Untersuchung von Kooperationstechniken wurde durch die Autoren nachgewiesen, dass bildliche Darstellungen eine wesentliche Bedeutung zur Unterstützung von Kommunikations- und Wissenstransferprozessen sowohl für eine wissenschaftlich fundierte Modellgenese (Münster 2013) als auch zur Wissensvermittlung besitzen (Weller 2013; Köhler et al. 2014; Münster et al. 2014).
- Vor diesem Hintergrund wurden durch die Autoren im Rahmen mehrerer internationaler Drittmittelprojekte räumliche, geobasierte 3D-Informationssysteme für geschichtswissenschaftliche Inhalte untersucht und hinsichtlich Gebrauchstauglichkeit und Informationsvermittlung getestet (Köhler et al. 2013). Dabei wurden innovative Informations- und Softwarearchitekturen sowohl für Geobrowser (Schubert 2013; Schubert 2013; Kröber 2014) als auch Plugin-freie und browserbasierte mobile 3D-Augmented Reality Verfahren (Prechtel et al. 2013; Schietzold 2013) entwickelt. Darüber hinaus wurden derartige Techniken für ein ubiquitäres Bildungs- und Informationssystem verwandt, welches individuelle Nutzerpräferenzen, Social Media-Aktivitäten und Online-Inhalte über eine semantische Wissensbasis verknüpft und dafür verwendet, dem Nutzer personalisierte Lerninhalte auf unterschiedlichen Endgeräten bereitzustellen (Funke et al. 2013).

Bibliographie

Andersen, Kirsti (2007): *The Geometry of an Art*. The History of the Mathematical Theory of Perspective from Alberti to Monge. New York: Springer.

Brandt, Ahasver von (2012): *Werkzeug des Historikers*. Eine Einführung in die Historischen Hilfswissenschaften. Stuttgart: Kohlhammer.

Brenningmeyer, Todd / Begg, Ian D. J. (2006): "Reconstructing Tebtunis: Assembling a Site Model Using Archived Aerial Photography", in: *Proceedings of the 34th Conference on Computer Applications and Quantitative Methods in Archaeology* (CAA), Fargo, USA.

Bürger, Stefan (2011): "Unregelmässigkeit als Anreiz zur Ordnung oder Impuls zum Chaos. Die virtuose Steinmetzkunst der Pirnaer Marienkirche", in: *Zeitschrift für Kunstgeschichte* 74: 123-132.

Burke, Peter (2003): *Augenzeugenschaft*. Bilder als historische Quellen. Berlin: Klaus Wagenbach Verlag.

DARIAH-DE (2015): *DARIAH-DE Geobrowser* http://dev2.dariah.eu/e4d/ [letzter Zugriff 9. Februar 2016].

DRESDEN-concept (2016): *Organisational Structure* http://www.dresden-concept.de/en/alliance/structures.html [letzter Zugriff 9. Februar 2016].

Endres, Ian / Shih, Kevin J. / Jiaa, Johnston / Hoiem, Derek (2013): "Learning Collections of Part Models for Object Recognition", in: *26th IEEE Conference on Computer Vision and Pattern Recognition* (CVPR).

Favro, Diane (2004): "In the eyes of the beholder. Virtual Reality re-creations and academia", in: Haselberger, Lothar / Humphrey, Jon / Abernathy, D.C. (eds.): *Imaging ancient Rome: Documentation, visualization, imagination*. Proceedings of the 3rd Williams Symposium on Classical Architecture, Rome. Portsmouth: Journal of Roman Archaeology.

Funke, Alexandra / Brunk, Sören / Kühn, Romina / Schlegel, Thomas (2013): "An Ontology-based Interaction Concept for Social-aware Applications Human-Computer Interaction. Towards Intelligent and Implicit Interaction", in: *Proceedings of the 15th International Conference on Human-Computer Interaction 2013, Las Vegas*. New York: Springer.

Hanzl, Vlastimil / Káňa, David (2012): "Application of computer vision methods and algorithms in documentation of cultural heritage", in: *Geoinformatics FCE CTU* 9 https://ojs.cvut.cz/ojs/index.php/gi/article/viewFile/gi.9.3/2407

.

Hoiem, Derek / Savarese, Silvio (2011): *Representations and Techniques for 3D Object Recognition & Scene Interpretation*. San Rafael: Morgan & Claypool Publishers.

Ioannides, Marinos / Hadjiprocopis, Andreas / Doulamis, Nikolaos / Doulamis, Anastasios / Protopapadakis, Eftychios / Makantasis, Kostas / Santos, Pedro / Fellner, Dieter / Stork, Andre / Balet, O./ Julien, Martine Julien / Weinlinger, Günther / Johnson, Paul S. / Klein, Michael / Fritsch, Dieter (2013): "Online 4D Reconstruction using Multi-Images available under Open Access", in: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5/W2 (XXIV International CIPA Symposium) 169-174.

Jacobson, Jeffrey / Handron, Kerry / Holden, Lynn (2009): "Narrative and Content Combine in a Learning Game for Virtual Heritage", in: *Proceedings of the 37th Computer Applications and Quantitative Methods in Archaeology Conference*, Williamsburg.

Justus-Liebig-Universität Gießen (2012-2015): *GeoBib*. GeoBib - Virtueller Atlas und Online-Bibliographie der frühen Holocaustliteratur http://www.uni-giessen.de/fbz/zmi/projekte/geobib [letzter Zugriff 9. Februar 2016].

Kersten, Thomas P. / Lindstaedt, Maren / Mechelke, Klaus / Zobel, Kay (2012): "Automatische 3D-Objektrekonstruktion aus unstrukturierten digitalen Bilddaten für Anwendungen in Architektur, Denkmalpflege und Archäologie", in: Seyfert, Eckhardt

(ed.): *Erdblicke - Perspektiven für die Geowisseschaften*. Vorträge: 32. Wissenschaftlich-technische Jahrestagung der DGPF, Potsdam (= Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V. 21). Oldenburg: DGPF.

Kohle, Hubertus / Bry, Francois(2015): *ARTigo*. Play4Science Projekt http://www.artigo.org/ [letzter Zugriff 9. Februar 2016].

Köhler, Thomas / Münster, Sander / Schlenker, Lars (2013): "Didaktik virtueller Realität. Merkmale einer zielgruppengerechten Gestaltung im Kontext akademischer Bildung", in: Reinmann, Gabi / Ebner, Martin / Schön, Sandra (eds.): *Hochschuldidaktik im Zeichen von Heterogenität und Vielfalt*. Doppelfestschrift für Peter Baumgartner und Rolf Schulmeister. Norderstedt: Books on Demand http://bimsev.de/ [letzter Zugriff 16. Februar 2016].

.

Köhler, Thomas / Münster, Sander / Schlenker, Lars (2014): "Smart communities in virtual reality. A comparison of design approaches for academic education", in: *Interaction Design and Architecture(s) Journal (IxD&A)* 22 (Special issue on "Social Behaviors and Learning in Smart Communities): 48-59.

Kröber, Cindy / Münster, Sander / Prechtel, Nikolas / Schietzold, Sebastian / Schubert, Christian (2014): "GEPAM – Eine interaktive Informationsplattform zur "Landschaft des Gedenkens (Poster)", in: *1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014)*, Passau.

Kwastek, Katja (2014): "Vom Bild zum Bild. Digital Humanities jenseits des Texts (Keynote)", in: *1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014)*. Passau.

Livingston, Mark A. / Bimber, Oliver / Saito, Hideo (2008): *Proceedings of the 7th IEEE International Symposium on Mixed and Augmented Reality*. Cambridge, UK. Piscataway, N.J.: IEEE Xplore.

Münster, Sander (2011a): "Entstehungs- und Verwendungskontexte von 3D-CAD-Modellen in den Geschichtswissenschaften", in: Meißner, Klaus / Engelien, Martin (eds.) *Virtual Enterprises, Communities & Social Networks*. Dresden: TUDpress.

Münster, Sander (2011b): "Militärgeschichte aus der digitalen Retorte - Computergenerierte 3D-Visualisierung als Filmtechnik", in: Kästner, Alexander / Mazerath, Josef (eds.): *Mehr als Krieg und Leidenschaft*. Die filmische Darstellung von Militär und Gesellschaft der Frühen Neuzeit (= Militär und Gesellschaft in der frühen Neuzeit 15, 2). Potsdam: Universitätsverlag.

Münster, Sander (2013): "The role of images for a virtual 3D reconstruction of historic artifacts", in: *Annual Meeting of the International Communication Association (ICA)*. London.

Münster, Sander (2014): *Interdisziplinäre Kooperation bei der Erstellung virtueller geschichtswissenschaftlicher 3D-Rekonstruktionen*. Dissertation. Technische Universität Dresden.

Münster, Sander / Hegel, W. / Kröber, Cindy (in Vorbereitung): "A classification model for digital reconstruction in context of humanities research", in: Münster, Sander / Pfarr-Harfst, Mieke / Ioannidis, Marinos / Quack, Ewald (eds.) *The 2nd International Workshop on ICT for the Preservation and Transmission of Intangible Cultural Heritage "How to exchange Cultural Heritage 3D objects and knowledge in Digital Libraries?"* Cham: Springer LNCS.

Münster, Sander / Ioannides, Marinos (2015): "The scientific community of digital heritage in time and space", in: *Digital Heritage 2015*.

Münster, Sander / Jahn, Peter Heinrich / Wacker, Markus (2015): "Von Plan- und Bildquellen zum virtuellen Gebäudemodell. Zur Bedeutung der Bildlichkeit für die digitale 3D-Rekonstruktion historischer Architektur", in: Ammon, Sabine / Hinterwaldner, Inge (eds.): *Bildlichkeit im Zeitalter der Modellierung*. Operative Artefakte in Entwurfsprozessen der Architektur und des Ingenieurwesens. München: Wilhelm Fink Verlag.

Münster, Sander / Köhler, Thomas (2012): "3D reconstruction of Cultural Heritage artifacts. A literature based survey", in: *Proceedings of CHCD2012 Conference*. Beijing.

Münster, Sander / Köhler, Thomas / Hoppe, Stephan (2013): "3D modeling technologies as tools for the reconstruction and visualization of historic items in humanities. A literature-based survey", in: Traviglia, Arianna (ed.): *Across Space and Time*. Selected Papers from the 41st Computer Applications and Quantitative Methods in Archaeology Conference, Perth. Amsterdam: Amsterdam University Press.

Münster, Sander / Kröber, Cindy / Schlenker, Lars / Weller, Heide (submitted paper): "Virtual Reconstructions of Historical Architecture as Media for Visual Knowledge Representation", in: *International Communication Association (ICA) Annual Meeting, 9-13 June 2016*. Fukuoka.

Münster, Sander / Prechtel, Nikolas (2014): "Beyond Software. Design Implications for Virtual Libraries and Platforms for Cultural Heritage from Practical Findings", in: Ioannidis, Marinos / Magnenat-Thalmann, Nadja / Fink, Eleanor / Žarnić, Roko / Yen, Alex-Yianing / Quak, Ewald (eds.): *Digital Heritage*. Progress in Cultural Heritage: Documentation, Preservation, and Protection. Cham: Springer.

Münster, Sander / Schlenker, Lars / Köhler, Thomas (2014): "Common grounds and representations in cross-disciplinary processes", in: Carlussi, Daniela, Schiuma, Giovanni / Spender, JC (eds.): *Knowledge and Management Models for Sustainable Growth*. Basilicata: Institute of Knowledge Asset Management.

Niccolucci, Franco / Hermon, Sorin (2006): "A Fuzzy Logic Approach to Reliability in Archaeological Virtual Reconstruction", in: Niccolucci, Franco / Hermon, Sorin (eds.): *Beyond the Artifact*. Digital Interpretation of the Past. Budapest: Archaeolingua.

Paul, Gerhard (2006): *Visual History*. Ein Studienbuch. Göttingen: Vandenhoeck & Ruprecht.

Pérez-Gómez, Alberto / Pelletier, Louise (1997): *Architectural Representation and the Perspective Hinge*. Cambridge / London: University Press.

Pierrot-Deseilligny, Marc / de Luca, Livio / Remondino, Fabio (2011): "Automated Image-Based Procedures for Accurate Artifacts 3D-Modeling and Orthoimage Generation", in: *Geoinformatics CTU FCE* 291-299.

Prechtel, Nikolas / Münster, Sander / Kröber, Cindy / Schubert, Christian / Schietzold, Sebastian (2013): "Presenting Cultural heritage Landscapes - From GIS via 3D Models to Interactive Presentation Frameworks", in: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. XL-5/W2 (XXIV International CIPA Symposium) 253-258.

Schietzold, Sebastian (2013): *Augmented Reality mit 3D-Inhalten auf mobilen Endgeräten*. Studienarbeit, betreut duch Florian Niebling / Sander Münster. TU Dresden.

Schubert, Christian (2013a): *Verbesserte Graphik und Integration für das 3D-Landschaftsmodell des Ethno-Nature Parks „Uch-Enmek" (Altai) unter Verwendung von OpenWebGlobe*. Studienarbeit, betreut duch Nikolas Prechtel / Sander Münster. TU Dresden.

Schubert, Christian (2013b): *Vom 3D – Landschaftsmodell zu einer integrativen Web-basierten Informationsapplikation für ein archäologisches Schutzgebiet* (Uch Enmek, Republik Altai). Diplomarbeit, betreut duch Nikolas Prechtel / Sander Münster. TU Dresden.

Stojakovic, Vesna / Tepavcevic, Boran (2009): "Optimal methods for 3d modeling of devastated architectural objects", in: Remondino, Fabio / El-Hakim, Sabry F. / Gonzo, Lorenzo (eds.): *3D-ARCH 2009*. 3D Virtual Reconstruction and Visualization of Complex Architectures. Trento, Italy.

Walczak, Krzysztof / Cellary, Wojciech / Prinke, Andrzej (2011): "Interactive Presentation of Archaeological Objects Using Virtual and Augmented Reality", in: Jerem, Erszébet / Redö, Ferenc / Szeverényi, Vajk (eds.): *On the Road to Reconstructing the Past*. Proceedings of the 36th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA). Budapest: Archaeolingua.

Weller, Heide (2013): *Generalisierte 3D-Gebäuderepräsentation im Spannungsfeld von Primärinformation, Modellierungsaufwand und Wiedererkennbarkeit am Beispiel eines 3D-Stadtmodells von Dresdens um 1940 (Diploma thesis)*. Diploma Thesis, TU Dresden.

Wohlfeil, Rainer (1986): "Das Bild als Geschichtsquelle", in: *Historische Zeitschrift* 243: 91–100.

Zöllner, Michael / Becker, Mario / Keil, Jens (2010): "Snapshot Augmented Reality - Augmented Photography", in: Artusi, Alessandro / Joly-Parvex, Morwena / Lucet, Genevieve / Ribes, Alejandro / Pitzalis, Denis (eds.): *11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2010)*. Paris: Eurographics Association.

Small Data: Wissensproduktion und -vermittlung bei digitalen Visualisierungen in der Kunstgeschichte

Neubauer, Susanne
susanne.neubauer@fu-berlin.de
Freie Universität Berlin, Deutschland

Der aktuelle Diskurs in den Digital Humanities dreht sich vorwiegend um die Frage nach der Fruchtbarmachung digitaler Daten für wissenschaftliche Erkenntnisse. Dabei wird zunehmend die Maxime laut, dass geisteswissenschaftliche Kernmethoden informationstheoretische und –praktische Ansätze aufzunehmen hätten. Zu einseitig sei noch die Tätigkeit der Kunstgeschichte, nur digitale Quellen zu generieren und eine „digitized" anstatt eine „digital art history" (Drucker 2013) zu betreiben. Die für die Geisteswissenschaften typische hermeneutische Herangehensweise sollte beispielsweise Niederschlag in den Aushandlungsprozessen der „neuen" Forschungsthemen der Digital Humanities finden (Gaehtgens 2013). In nahezu allen Bereichen – den Visualisierungen und Modellierungen sowie den zahlreichen Bild-Text-Datenbanken der bildbasierten Wissenschaften – lassen sich jedoch weder Ansätze einer selbstreflexiven Forschungs- und Vermittlungstätigkeit noch die Thematisierung prozessualer Vorgänge ausmachen. Gerade die Prozessualität von Forschung und Vermittlung in den Künsten, die historisch bedeutende Vorgänger aus dem analogen Bereich des Museums- und Publikationswesens hervorgebracht hat, müsste verstärkt Eingang in die Problematisierung von computergestützter Forschung in der Kunstwissenschaft finden.

Mein Beitrag will mit historischen analogen und aktuellen digitalen Beispielen aus den objektbezogenen Bereichen der digitalen Kunstgeschichte einen Gegenentwurf zu „distant reading" (Moretti) und „big data" (beispielsweise Lev Manovichs vorgeschlagenen Data-Visualisierungen) vorlegen und ein besonderes Augenmerk auf die Vermittlung und Dokumentation gerade räumlicher Kunst (Ausstellungen, Installationen) werfen. Ein erster Teil thematisiert Projekte, die einen besonderen Bezug zwischen der historischen Gegenwart im Raum (Ausstellung) und deren

Dokumentation (Ausstellungskatalog) entwickelt haben. Als besonderes Beispiel in diesem Kontext dient die erste monografische Ausstellung und Publikation des amerikanischen Künstlers Richard Tuttle, die Marcia Tucker 1975 für das Whitney Museum of American Art eingerichtet hat. Dieses Beispiel dient als Grundlage zu weiterführenden Überlegungen, wie rahmende Formate („framing devices", Doulkaridou, 2015) nicht nur Zugang zu Daten bieten, sondern auch deren Interpretation beeinflussen. Mögliche Strukturen der Vermittlung sind idealerweise durch eine Verräumlichung, wie sie bereits auch in Ausstellungen angelegt ist, geprägt. Dabei gilt es, nicht nur diese Strukturen auf ihre Nutzbarkeit in den Digital Humanities hin zu prüfen, sondern zugleich auch die gerne zur Wissensvermittlung herangezogenen Instrumente wie Modell, Modellierung oder Visualisierung kritisch zu befragen. Ein zentraler Diskussionspunkt ist beispielsweise die Frage nach dem Umgang historischer Lücken, die in Modellierungen oder Visualisierungen bisher keine allgemein eingeführten Formate gefunden haben, jedoch von großer wissenschaftlicher Relevanz sind. Gerade in bilderzeugenden Verfahren liegt der Schwerpunkt in der Visualisierung von vorhandenen Daten oder gar der Generierung neuer Daten, nicht aber in der Thematisierung der Datenlücken bzw. der prinzipiell fragmenthaften Quellenlage historischer Ereignisse, die sich an Dokumenten, Artefakten oder Kunstwerken festmachen lassen. An dieser grundlegenden Frage nach den Möglichkeiten digitaler Verfahren unterschieden sich die textbezogenen von den objektbezogenen Geisteswissenschaften in dem Sinne, dass bei den ersteren die Vorstellung einer potentiellen Fragmenthaftigkeit der zumindest gedruckten Literatur vernachlässigt wird.

In einem zweiten Teil stellt der Vortrag drei unterschiedliche Ansätze der digitalen Kunstgeschichte vor, die sich einer kritischen Interpretation von bildbasierten Daten annähern: Das hypermediale Text-Bild-Archiv zu Anna Oppermanns Werken (Leuphana Universität Lüneburg; Wedemeyer; Warnke; Terstegge), das von mir entwickelte digitale Dokumentationsprojekt PTPROJECT.NET zum amerikanischen Künstler Paul Thek sowie die von Catherine Dossin konzipierte kritische Umsetzung eines Mapping-Projekts zur Rezeption amerikanischer Kunst im westlichen Nachkriegseuropa. Alle drei Projekte verbinden eine strukturelle Anlage, die über eine reine Bilddatenbank hinausgeht. Sie zeigen einen Umgang mit Quellenmaterial auf, das sich als „small data" bezeichnen ließe und das Hinweise auf seine Entstehungsgeschichte bietet.

Meine Überlegungen zielen folglich auf eine Umkehrung von den Verfahren der „big data" zu einem Konzept der „small data" im Kontext der kunsthistorischen Wissensbildung ab. Das Konzept der „small data" fokussiert auf die kritische öffentliche Vermittlung von digitalen Daten, die gerade ihre Entstehungsgeschichte, Einzigartigkeit sowie ihre übersetzten inhärenten Qualitäten der abgebildeten künstlerischen Objekte und ihrer Verfahren in die Vermittlung miteinbeziehen und nicht durch einen distanzierten Blick auf die digitalisierte Masse ungreifbar machen wollen. Die Herausforderung dieses Ansatzes besteht darin, auf theoretischer Ebene einen Diskurs anzuregen, der sich mit der Problematik des „lückenhaften Quellennetzes" (Reinhard Koselleck) in Bezug zu wissenschaftlichen Fragestellungen befasst, und der zugleich auf praktischer Ebene eine Umsetzung in veränderbaren, digitalen Wissensstrukturen ermöglicht.

Bibliographie

Dossin, Catherine (2012): „Mapping the Reception of American Art in Postwar Western Europe", in: *Artl@a Bulletin* 1, 1: Article 3.

Doulkaridou, Elli (2015): "Reframing Art History", in: *DAH-Journal* 1: 67-83 http://dx.doi.org/10.11588/dah.2015.1.21638 [letzter Zugriff 11. Oktober 2015].

Drucker, Johanna (2013): „Is There a „Digital" Art History?", in: *Visual Resources: An International Journal of Documentation* 29, 1-2: 5-13.

Gaeghtens, Thomas W. (2013): „Thoughts on the Digital Future oft he Humanities and Art History", in: *Visual Resources: An International Journal of Documentation* 29, 1-2: 22-25.

Kosselleck, Reinhard (1995): „Ist Geschichte eine Fiktion?", Interview von Hasso Spode mit Reinhard Kosselleck, in: *NZZ Folio* 3.

Manovich, Lev (2015): „Data Science and Digital Art History", in: *DAH-Journal* 1: 13-25.

Moretti, Franco (2013): *Distant Reading*. London: Verso.

Neubauer, Susanne (2005): *PTPROJECT.NET. Outline of the Research Project* http://www.ptproject.net/introduction_more.php [letzter Zugriff 11. Oktober 2015].

Tucker, Marcia (1975): *Richard Tuttle* New York: Whitney Museum of American Art.

Warnke, Martin / Wedemeyer, Carmen (2011): "Documenting Artistic Networks", in: *Leonardo*. Journal of the International Society for the Arts, Sciences and Technology 44, 3: 258-259.

Digitale Editionen als Web-Services

Normann, Immanuel
immanuel.normann@pagina-tuebingen.de
pagina GmbH, Deutschland

Verstehen wir unter einer digitalen Edition eine „erschließende Wiedergabe historischer Dokumente", welche dem digitalen Paradigma folgt, indem sie die gegenwärtigen technischen Möglichkeiten berücksichtigt

(cf. Sahle 2013: 138, 148), dann stellt sich die Frage, welche technischen Möglichkeiten zu welchem Zweck eingesetzt werden sollen. In diesem Beitrag wird die Überzeugung vertreten, dass digitale Editionen als zentraler Bestandteil von Forschungsumgebungen der Textwissenschaft von weit größerem Nutzen sein können, wenn sie über standardisierte semantische Web-Schnittstellen verfügen. Digitale Editionen wären dann primär als Web-Services zu verstehen, die über ihre Web-Schnittstellen mit anderen Web-Services oder mit Web-Anwendungen kommunizieren. Es wäre erst die Web-Anwendung (welche im Browser ausgeführt wird), mit der der menschliche Nutzer interagiert, wogegen alle übrige Kommunikation von Maschine zu Maschine liefe. Herkömmliche digitale Editionen sind primär auf eine Nutzung durch den Menschen allein ausgerichtet. Die im Folgenden zu begründende These ist, dass Werkzeuge der Forschungsumgebungen mit diesen herkömmlichen digitalen Editionen deshalb nur unbefriedigend ineinandergreifen, weil sie programmatisch abgeschlossen sind. Dieser Zustand ist insofern unbefriedigend, als dadurch Textforschung weit weniger vernetzt und kollaborativ vonstatten geht als dies möglich wäre.

Eine Verbesserung dieses Zustands kann natürlich nicht allein von technischen Neuerungen digitaler Editionen erhofft werden. Es sind ebenso technische Neuerungen bei allen Komponenten bestehender Forschungsumgebungen nötig (und bei Initiativen wie TextGrid auch im Gange). Dabei besteht eine wechselseitige Abhängigkeit des Entwicklungsfortschritts: Nur wenn die eine Komponente das eine neue Feature anbietet, besteht bei der anderen Komponente die Chance eines Entwicklungssprungs. Mit Blick auf diese *Koevolution* müssen also diejenigen Komponenten einer Forschungsumgebung berücksichtigt werden, die mit einer digitalen Edition im Datenaustausch stehen oder stehen sollten. Dabei ist es zielführend, sich nicht ausschließlich von der Frage leiten zu lassen, wie man digitale Editionen möglichst interoperabel zu den am weitestverbreiteten Werkzeug der Textwissenschaftler (z. B. der dominierenden Textverarbeitungssoftware) machen kann. Vielmehr sollte die Aufmerksamkeit darauf gerichtet werden, welche nützlichen Werkzeuge man schaffen könnte, wenn man die digitalen Editionen mit bestimmten technischen Neuerungen ausstatten würde.

Im Folgenden wird daher das Umfeld digitaler Editionen innerhalb einer textwissenschaftlichen Forschungsumgebung in den Blick kommen und zwar in einer Weise, die auch noch nicht existierende Systeme mitdenkt. Dies ist möglich, wenn man eine solche Umgebung zu diesem Zweck nicht als eine Ansammlung bestehender Tools auffasst, sondern die textwissenschaftlichen Tätigkeiten identifiziert, für die man sich ohne Rücksicht auf bestehende Fertiglösungen technische Unterstützung überhaupt vorstellen kann.

Die aus informationstechnischer Sicht relevanten Tätigkeiten lassen sich in diesem Kontext sinnvoll unterteilen in: das *Lesen*, *Schreiben* und *Verwalten* von Text. Während das Lesen und Schreiben von Text in diesem Rahmen keiner weiteren Erklärung bedarf, muss näher darauf eingegangen werden, was mit Textverwaltung alles gemeint sein kann. Eine positive Definition dieses Begriffs würde wahrscheinlich keine allgemeine Zustimmung finden, daher sollen ein paar paradigmatische Beispiele zur Begriffsklärung ausreichen: Exzerpieren, Organisieren von Textschnipseln in Zettelkästen, Anlegen von Literaturlisten, Zusammenstellen eines Semesterapparats, Sortierung von Büchern, Klassifikation von Texten, Erstellen von Registern und vieles mehr – für all diese und ähnliche Tätigkeiten soll der Begriff Textverwaltung hier stehen. Zwar wird in all diesen Fällen auch geschrieben und gelesen, aber das ist nicht das Wesentliche an der Textverwaltung, sondern die in diesen Tätigkeiten erzeugten Ordnungen oder Relationen.

Fragen wir uns nun, zu welchen dieser drei Tätigkeitsfeldern (Lesen, Schreiben, Verwalten) eine digitale Edition eine unmittelbare und eine mittelbare Unterstützung liefern kann. Traditionell dienen digitale Editionen (wie ihre gedruckten Vorfahren) in erster Linie dazu gelesen zu werden. Zwar sind die in ihr enthaltenen Texte und ihre Metadaten natürlich auch Ergebnis einer Textverwaltung. Jedoch bieten sie dem Nutzer nur in seltenen Fällen und da auch nur rudimentär die Möglichkeit selbst Text zu verwalten (cf. z. B. Arbeitsmappen bei Jung 2015). Eine außergewöhnliche Ausnahme ist ein Editionsprojekt zu Pessoas „Buch der Unruhe“ (cf. Silva / Portela 2015). Hier ist das Lesen, Schreiben und Verwalten gleichermaßen möglich und ermöglicht den Nutzern aus dem vorhandenen Textmaterial und eigenen Kommentaren eine eigene virtuelle Edition kollaborativ zu erstellen. In diesem Sinne ist diese Plattform nicht mehr eine Edition im traditionellen Sinne, sondern selbst eine in sich abgeschlossene Forschungsumgebung – allerdings für eine ganz spezielle Aufgabe über ein abgegrenztes Textkorpus.

All diesen digitalen Editionen ist jedoch gemeinsam, dass, sofern sie eine Textverwaltung unterstützen, diese dann nur für die im System vorhandenen (oder darin erzeugten) Texte ermöglichen. Im Allgemeinen ist der Textwissenschaftler aber nicht mit einem einzelnen Textkorpus befasst, sondern mit mehreren. Eine Textverwaltung kann dann nur ihren Nutzen entfalten, wenn sie als eigenständiger Service auf mehrere digitale Editionen zugreifen kann.

Nehmen wir als einfaches Beispiel die Zusammenstellung der Literatur zu einem Germanistikseminar, in dem Texte verschiedener Autoren behandelt werden. Von einer komfortablen Textverwaltung würde man jetzt nicht die URL der jeweiligen digitalen Editionen erwarten, sondern man möchte am besten die Texte selbst per Mausklick zur Verfügung gestellt bekommen ohne dabei auf die Web-Seiten der jeweiligen digitalen Editionen gehen zu müssen. Schon dieser einfache Fall zeigt den Nutzen,

den eine programmatische Schnittstelle von digitalen Editionen haben könnte: Ein eigenständiger Service zur Aggregation von Semesterapparaten ließe sich mit geringem Aufwand implementieren.

Tatsächlich bieten manche digitale Editionen (z. B. das Deutsche Textarchiv) ihre Texte (sogar in verschiedenen Formaten: TEI, HTML, plain text) zum Download an, so dass man die entsprechenden Links schon als Web-API auffassen könnte. Allerdings beschränkt sich diese Möglichkeit entweder auf den Download einer einzelnen Seite oder des gesamten Textdokuments. Für eine brauchbare Textverwaltung wäre es jedoch wesentlich praktischer, wenn man Texte nicht nach Paginierungsgrenzen sondern bezüglich semantischer Sinneinheiten beziehen könnte. Es fällt nicht schwer, sich entsprechende Szenarien vorzustellen: Für eine Anthologie möchte man etwa Balladen einer bestimmten Epoche zusammenstellen.; für eine Theaterprobe möchte jeder Schauspieler eine Zusammenstellung derjenigen Szenen, in der seine Rolle vorkommt; ein Übersetzungsforscher möchte alle deutschen Übersetzungen des Monolog der ersten Szene im dritten Aufzug von Shakespeares Hamlet. Die Zahl weiterer Szenarien ist unbegrenzt. Als entscheidende Anforderung an eine digitale Edition wäre festzuhalten: die Adressierbarkeit und Auffindbarkeit von Texten in allen üblichen Struktureinheiten (z. B. Kapitel, Absatz, Drama, Akt, Szene, Gedicht, Strophe, Vers, etc.). Da in den meisten digitalen Editionen die Texte im TEI-XML vorliegen, welche die Kodierung solcher Struktureinheiten erlauben, dürfte es prinzipiell nicht schwierig sein, diese auch über eine Web-API adressierbar zu machen. Was die Auffindbarkeit betrifft, wäre es wünschenswert, die Möglichkeitender in der Backend-Datenbank verwendeten Anfragesprachen weitgehend in der Web-API abzubilden. Das ganze Feld der Suchmöglichkeiten ist allerdings so umfangreich, dass es einen eigenen Beitrag rechtfertigen würde und daher hier nicht weiter vertieft werden soll. Allein die Adressierbarkeit aller textspezifischen Struktureinheiten (s. o.) mittels der Web-API von digitalen Editionen wäre eine große Chance zur Entwicklung nützlicher Textverwaltungsdienste. Allerdings sollten neben den vorgegebenen Struktureinheiten auch vom Nutzer frei definierte Textauswahlen von einer digitalen Edition adressierbar sein. Damit soll die verbreitete Praxis, Textausschnitte mit einem Textmarker zu markieren, im digitalen Medium nicht nur die Funktion erhalten, etwas farblich hervorzuheben, sondern die so ausgezeichneten Textpassagen sollen durch eine generierte Adresse permanent referenzierbar gemacht werden. Damit wäre beispielsweise eine Sammlung von Exzerpten referenzierbar, die ein Benutzer mit einem virtuellen Textmarker erzeugt hat.

Bis hierin wurde die Adressierbarkeit von jeglichen Textausschnittenin den oben angeführten Szenarien ausschließlich für die Erstellung von Textsammlungen verwendet. Das ist aber nur eine einfache Form der Textverwaltung. Denn eine Textsammlung ist zunächst eine in sich unstrukturierte Menge von Texten. Ziel einer Textverwaltung ist es aber meist, in eine Textsammlung eine bestimmte Ordnung zu bringen. Das ist unter anderem der Fall, wenn man die gesammelten Texte nach forschungseigenen Kriterien klassifiziert; z. B. als Linguist nach grammatischen Eigenschaften, als Literaturwissenschaftler nach Motiven, als Übersetzer nach Idiomen, etc.

Textklassifikation wäre eine Relation zwischen Texten und Sammelbegriffen. Darüber hinaus wäre es wichtig, in einer Textverwaltung die Beziehung der Texte untereinander explizit machen zu können. So könnte man beispielsweise explizit erfassen, dass eine bestimmte Textpassage eine Anspielung auf einen anderen Text ist; oder dass die eine Textfassung aus jener Skizze hervorgegangen ist, etc. Soweit würde man Textausschnitte aus digitalen Editionen in Beziehung zueinander setzen. Man würde aber in einer Textverwaltung insbesondere auch die Texte der digitalen Editionen in Beziehung zu selbstverfassten Texten setzten wollen. Auch würde man Texte zu nicht textartigen Gegenständen wie Personen, Orte oder Ereignissen in Beziehung setzen wollen; beispielsweise wenn man in historischen Romanen den Bezug zu historisch belegten Sachverhalten herstellen möchte.

Eine Textverwaltung, die all die skizzierten Funktionalitäten bereitstellen würde, könnte einen Textwissenschaftler bei der Arbeit am Text bzw. der Organisation der eigenen Texte erheblich unterstützen. Sie würde darüber hinaus das kollaborative Arbeiten erleichtern, indem sie eine auf Austausch von Dokumenten basierte Arbeitsweise durch eine Praxis der direkten Vernetzung von Inhalten im Netz ersetzen würde. Sie könnte aber nur funktionieren, wenn die Texte digitaler Editionen in aller Granularität über Web-APIs adressierbar wären.

Abschließend soll erwähnt werden, das eine ganze Reihe von Anstrengung von verschiedenen Seiten schon unternommen wurden, die durch eine geeignete Zusammenführung ein solides Fundament zur Umsetzung dieser Visionen bilden könnten. Allgemeine technische Grundlage wären die Semantic-Web-Technologien. Darauf aufbauend wären folgende theoretische und praktische Arbeiten hervorzuheben: Von Silvio Peroni (2014) zu „Semantic Publishing“ , Fabio Ciottis und Francesca Tomasis (2014) Entwurf zu „Formal ontologies, Linked Data and TEI semantics“, das semantic annotation Tool Pundit (2013-*) und die Open Annotation Initiative : http://www.openannotation.org/.

Bibliographie

Ciotti Fabio, Tomasi Francesca (2014): *Formal ontologies, Linked Data and TEI semantics*. TEI Conference and Members Meeting 2014. Evanston (IL), October 22-24, 2014. http://tei.northwestern.edu/

files/2014/10/Ciotti-Tomasi-22p2xtf.pdf [letzter Zugriff 09. Januar 2016].

Jung, Joseph (ed.) (2015): *Digitale Briefedition Alfred Escher*. Version: Juli 2015. Zürich: Alfred Escher-Stiftung. http://www.briefedition.alfred-escher.ch/ [letzter Zugriff 09. Januar 2016].

Peroni, Silvio (2014): *Semantic Web Technologies and Legal Scholarly Publishing*. Switzerland: Springer International Publishing http://www.springer.com/us/book/9783319047768 [letzter Zugriff 09. Januar 2016].

Pundit (2013-*): *Pundit* net7 http://thepund.it/ [letzter Zugriff 09. Januar 2016].

Sahle, Patrick (2013): *Digitale Editionsformen*. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels: Befunde, Theorie und Methodik (= Schriften des Instituts für Dokumentologie und Editorik 8). Norderstedt: Books on Demand.

Silva, António Rito / Portela, Manuel (2015): "TEI4LdoD: Textual Encoding and Social Editing in Web 2.0 Environments", in: *Journal of the Text Encoding Initiative* 8 http://jtei.revues.org/1171 [letzter Zugriff 09. Januar 2016].

A Visual Approach to the History of Swiss Federal Law

Ourednik, André
Andre.Ourednik@bar.admin.ch
Schweizer Bundesarchiv

Nellen, Stefan
Stefan.Nellen@bar.admin.ch
Schweizer Bundesarchiv

Fleer, Peter
Peter.Fleer@bar.admin.ch
Schweizer Bundesarchiv

The complexity of law is a recurring problem faced by government bodies, judges, lawyers, jurists and, in fact, all citizens confronted with legal issues. Ideally, the application of laws should be in line with the ideals of the society that formulates them. Whether this ideal is satisfied or not depends on many factors, among which the structure of the legal system, the quantity of its resources, or the competence of law interpreters. But it also depends on the form of the laws themselves. The equilibrium of law-writing consists in remaining intelligible while accounting for the great diversity of social situations subject to legislation. It must “find an optimal balance between rigorousness and flexibility, formality and understandability, that is, between expressing authority and yet being grounded in the everyday life of those affected by it” (Höfler / Piotrowski 2011). Problems induced by excessive complexity of law are well identified in legal literature (Epstein 1995; Kades 1996; Katz / Bommarito 2013; Palmirani / Cervone 2014). They have also been drawn to public attention. Complaints about “overlegislation”, meaning the rising quantity and complexity of law, have been formulated in recent years by national media in Switzerland, partially reflecting public opinion and the views of several politicians (e.g. Zurlinden 2013-2015c).

Our question, focused on the Swiss Federal Law, is both formal and historical. We ask whether the degree of its complexity can be measured over time and how it can be represented. How could we, for example, identify specific domains of the Federal Law in which complexity has evolved more than in others? What would this evolution mean in historical terms? In a larger scope, we ask to which extent the history of legislation can be understood with the help of quantitative methods.

Since 1848, the Swiss federal government publishes legal instruments in 3 national languages (German, French, Italian) and in 3 principal publications: the *Federal Gazette*, a weekly report of the Federal Council and law drafts proposed to the Federal Assembly, the legally binding *Official Compilation* (OC), a record of actual changes to the federal law in chronological order, updated every week; finally the trimestral *Systematic compilation*, in thematic order and representing the current state of federal law. Texts are partitioned in structural units (chapters, sections, articles, paragraphs, sentences and enumeration items) unequivocally identified by a reference code (e.g. SR 313.0, art. 29, §1, al. a) equivalent for all language versions. Working with the OC allows us to observe the chronological development of law on a weekly basis.

The OC is available since 1998 in digital format convertible to hierarchically structured XML by an XSLT transformation. Prior to 1998, the Swiss federal archives (SFA) dispose of paper version of the OC. This repository is currently being scanned in the scope of our project. Fields recognition (Figure 1) and OCR is applied to obtain, here also, structured XML (Figure 2).

This XML representation of the introduction of a new law at a given time allows a data analysis with the R statistical programming tool, the extraction of its results to a data format (JSON, cf. ECMA 2013) exploitable by the visualization framework D3. We can visualize laws in the form of a hierarchical tree proceeding from the whole body of the text – at the center of the figure – deployed into lower level structural units (Figure 3). This gives us a structural overview of any given law.

The next step of our work consisted in determining a means to *measure* the growth or regression of any given law over time. What we are interested in is not only the absolute size of the law but, more generally, the degree of difficulty it can represent for a comprehensive overview

by individuals deemed to conform to it or to apply it in the case of legal litigation. Beyond this practical aspect, it is also our wish, and our theoretical posture, to understand laws as organic actants (Callon 1986; Latour 1984, 1996), who prosper or regress in a dialectical relationship with the evolution of their "environment" (Ourednik 2010: §2.3.4.1, §2.2.4.2.), i.e. in relation to the context of other laws, superior (e.g. international) laws, as well as in relation to the social, cultural and economic climate. This posture allows us to speak about growth and regression of laws. Further, considering the dialectical relationship between an actant and its environment, one *relatum* can be taken as an index (Peirce et al. 1934) of the other. More specifically, observing the evolution of a law allow us to formulate historical hypotheses about the evolution of its environment. The question is what quantitative variables can be used to measure the evolution of a law.

In literature, we can distinguish between two approaches: measured and based on *textual statistics* partially inspired by information theory and measures based on the *spatiotemporal extent of law application*.

Among the textual measures, Katz and Bommarito (2013) suggest the use of *total text length*, *average word length*, *Shannon entropy of word use*, *depth of the hierarchical structure* and *density of external references*. Höfler and Piotrowski (2011: 87) further consider the structural complexity by observing the *number of subunits per structural unit* (e.g. paragraphs per article), down to *semantic units* per sentence, separated by expressions like "whereby" or "with the exception of" etc. Palmirani and Cervone (2014) take a diachronic posture, and propose a comprehensive indicator of *dynamic complexity* based on the amount and the impact of consecutive amendments to an original law.

Among the extent-based measures, Tribou and Collins (2015) consider the number of USA States in which a given law applies before reaching federal application. Other suggestions imply more extensive approximations. Epstein (1995: 22), for instance, suggests to measure it in terms of the cost of obedience, in other words: "[T]he cheaper the cost of compliance, the simpler we can say the rule is".

Another question is how to visually render these measurements. In order for our research work to be of consequence in the *public sphere*, we wish to address three publics: a) *researchers*, b) *jurists*, and c) the *general public*. For each, we have identified specific interests: a) to test the hypothesis of measurability of a historic process and to inspire new fields of investigation, b) to get a synthetic overview of a law, notably in terms of simultaneous or otherwise related amendments; to "identify tough cases [of complex laws] or to help resolve them" (Kades 1996) c) to navigate more easily in the realms of law and to obtain an answer to the question of rising complexity of the leas. From these interests ensue desired qualities of an optimal visualization: a) the hierarchical structure of given law must be visible; the measured quantities must be rendered; the visualization must allow for an animation reflecting the evolution of the measured quantities; b) both hierarchical and transversal (cross-references and thematic) links between elements of law must be renderable; c) the visualization must allow for an understanding of the structure of law for easier access to its elements; it should work as a metaphor of law as an "evolving organism".

Among possible visualizations, we have considered trees, partition layouts, and circle packing (Ourednik 2013). We have finally opted for a *force directed layout* as our visualization of choice (Figure 4). In effect, the hierarchical structure of a given law is apparent, without excluding the concurrent visualization of cross-references hardly introducible in tree-structured data. The visualization can easily evolve to reflect historic change (cf. Data Publica 2012). It can also be restructured in order, for instance, to give more weight to thematic, rather than cross-referential transversal links: elements of law thus related become closer in the resulting visualization. Finally from the metaphorical point of view useful for making our theoretical posture understandable to the media and to the general public, the visualization presents a strong analogy with a dendritic organism like a Physarum polycephalum or the mangrove. It thus makes more accessible the description of law in terms of "growth" or "regression".

Our second choice visualization is the circular partition tree (Figure 5). Here, transversal links cannot be introduced but the hierarchical structure of law attains better visibility. This visualization also allows for a direct appreciation of the total volume of law and its evolution over time.

At this stage of our research, we have applied quantitative measures based on textual statistics to a static situation, proceeding from an analysis of the French version of the *Loi fédérale sur le droit pénal administratif (DPA)* from March 22 nd 1974. The hypothesis of measurability of law has thus been verified and we have been able to identify our visualizations of choice. The measurement and visualization method is reproducible for any law at any given time.

Our ongoing work consists in consolidating the data on the evolution of Swiss federal law since 1947 in order to obtain distinct network structures and variable measurements for a fine-grained, week-based, observation of its evolution. Diachronic variable measurements shall be done in parallel on all 3 language versions of the legal texts, and compared, so as to control the robustness of our observations. The Swiss multilingual context offers a unique opportunity for such comparisons.

Visualizations shall be converted in an interactive interface allowing other researches to examine the results and to hypothesize, we hope, innovative explanations of the evolution of the Federal Law. They shall also be made accessible to the general public via the media as an input for the debate about legal complexity in the public sphere.

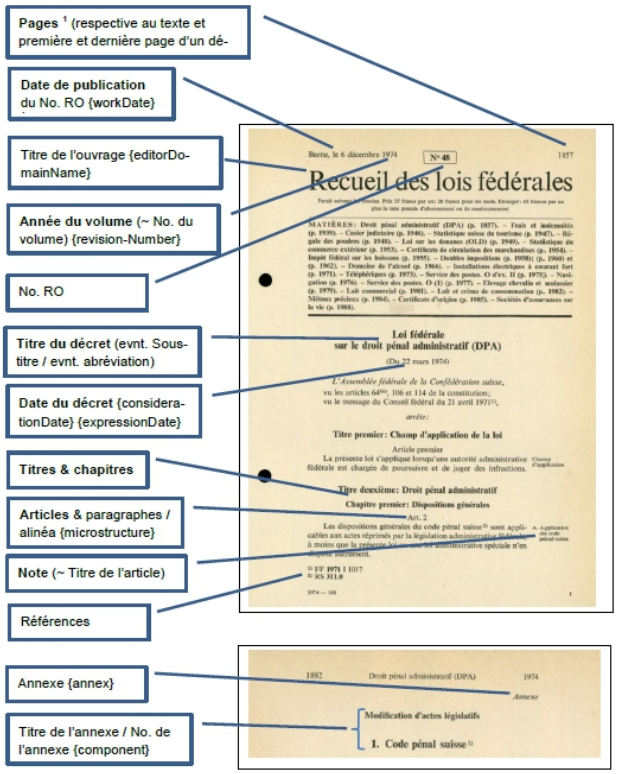


Fig. 1: Field recognition in the *DPA* from March 22nd 1974.

```
    <preface>
      <docTitle>Loi f&#233;d&#233;rale sur le droit p&#233;nal administratif</docTitle>
      <abbr>DP A</abbr>
      <date date="1974-03-22">22 mars 1974</date>
      <docketNumber>1974-0188</docketNumber>
    </preface>
    <preamble/>
    <body status="incomplete">
      <title id="tit1">
        <num>Titre premier</num>
        <heading>Champ d&#39;application de la loi</heading>
        <article id="art1">
          <num>Article premier</num>
          <heading>Champ d&#39;application</heading>
          <paragraph>La pr&#233;sente loi s&#39;applique lorsqu&#39;une autorit&#233; administrative f&#233;d&#233;rale
est charg&#233;e de poursuivre et de juger des infractions.</paragraph>
        </article>
      </title>
      <title id="tit2">
        <num>Titre deuxi&#232;me</num>
        <heading>Droit p&#233;nal administratif</heading>
        <chapter id="cha1">
          <num>Chapitre premier</num>
          <heading>Dispositions g~n&#233;rales</heading>
          <article id="art2">
            <num>Art. 2</num>
            <heading>A. Application du code p&#233;nal suisse</heading>
            <paragraph>Les dispositions g&#233;n&#233;rales du code p&#233;nal suisse 2&#62; sont appli&#592; cables aux
actes r&#233;prim&#233;s par la l&#233;gislation administrative f&#233;d&#233;rale, &#224; moins que la pr&#233;sente loi
ou une loi administrative sp&#233;ciale n&#39;en dispose autrement.</paragraph>
            <ref>RS 311.0</ref>
          </article>
          <article id="art3">
            <num>Art. 3</num>
            <heading>B. Inobservas&#183; lion de prescription d&#39;ordre</heading>
            <paragraph>Est r&#233;put&#233;e inobservation de prescription d&#39;ordre au sens de la pr&#233;sente loi la
contravention que la loi administrative sp&#233;ciale d&#233;signe sous ces termes et la contravention passible d&#39;une
amende d&#39;ordre.</paragraph>
          </article>
          <article id="art4">
            <num>Art.4</num>
            <heading>C. D&#233;rogations au code p&#233;nal suisse</heading>
            <heading>I. Enlants</heading>
            <paragraph>L&#39;enfant qui commet un acte punissable n&#39;est pas poursuivi.</paragraph>
          </article>
          <article id="art5">
```

Fig. 2: A XML Akoma Ntoso representation the *Bundesgesetz über das Verwaltungsstrafrecht* from March 22nd 1974 (excerpt).

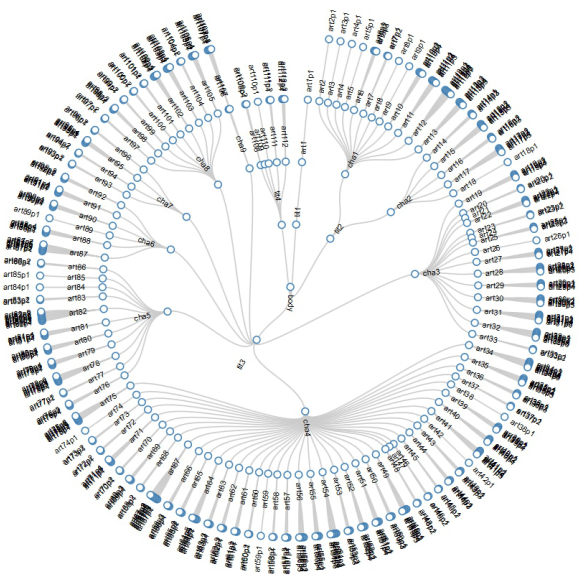


Fig. 3: Circular tree layout representation of the *DPA* from March 22nd 1974.

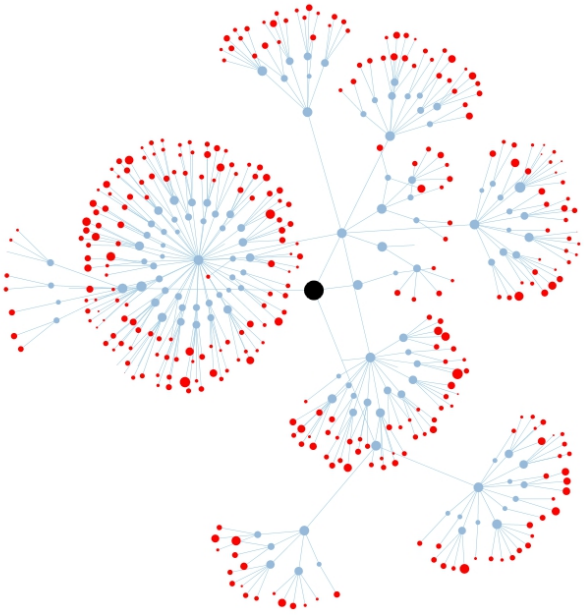


Fig. 4: Force directed layout visualization. Black: body of the *DPA* from March 22nd 1974. Blue: article with number of contained paragraphs; Red: paragraphs with word lengths.

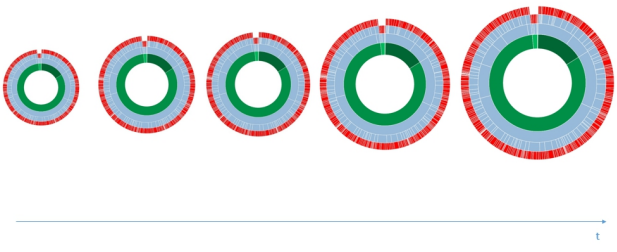


Fig. 5: Evolution of law viewed in a circle partition layout. size: total amount of text: green : titles, blue : articles and subarticles ; red : paragraphs.

Bibliographie

Akoma Ntoso Group (2013): *Akoma Ntoso - XML for parliamentary, legislative & judiciary documents* http://www.akomantoso.org/ [last seen October 15th 2015].

Callon, Michel (1986): "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fisher-men of St Brieuc Bay", in: Law, John (ed.): *Power, Action and Belief*. A New Sociology of Knowledge. London: Routledge & Kegan 196-223.

D3.js: *Data-Driven Documents* http://d3js.org/ [last seen October 15th 2015].

Data Publica (2012): *Naviguez dans le Code civil grâce à cette spectaculaire dataviz* www.data-publica.com/content/2012/07/naviguez-dans-le-code-civil-grace-a-cette-spectaculaire-dataviz/ [last seen October 15th 2015].

Epstein, Richard A. (1995): *Simple Rules for a Complex World*. Cambridge, Mass.: Harvard University Press.

ECMA International (2013): *The JSON Data Interchange Format*. ECMA.

Höfler, Stefan / Piotrowski, Michael (2011): "Building corpora for the philological study of Swiss legal texts", in *Journal for Language Technology and Computational Linguistics* 26, 2: 77-89.

Kades, Eric (1996): "Laws of Complexity and the Complexity of Laws: The Implications of Computational Complexity Theory for the Law", in *Rutgers Law Review* 49: 403.

Katz, Daniel M. / Bommarito, Michael J. (2013): "Measuring the Complexity of the Law: The United States Code", in *Social Science Research Network* http://ssrn.com/abstract=2307352 [last seen October 15th 2015].

Latour, Bruno (1984): *Les microbes, guerre et paix*. Paris: Métailié.

Latour, Bruno (1996): "On actor-network theory: a few clarifications", in: *Soziale Welt* 47, 4: 369–381.

Ourednik, André (2010): *L'habitant et la cohabitation dans les modèles de l'espace habité*. PhD, Swiss Federal Institute of Technology (EPFL).

Ourednik, André (2013): *Visualisation de flux géographiques*. Rapport pour la DATAR sous mandat du Laboratoire Cho#ros, Swiss Federal Institute of Technology (EPFL).

Palmirani, Monica / Cervone, Luca (2014): "Measuring the Complexity of the Legal Order over Time", in: Casanovas, Pompeu / Pagallo, Ugo / Palmirani, Monica / Sartor, Giovanni (eds.): *AI Approaches to the Complexity of Legal Systems*. Berlin: Springer 88-99.

Peirce, Charles S. / Hartshorne, Charles / Weiss, Paul (1932): *Collected Papers of Charles Sanders Peirce*. II: Elements of Logic. Cambridge, Mass.: Harvard University Press.

Tribou, Alex / Collins, Keith (2015): *This Is How Fast America Changes Its Mind* http://www.bloomberg.com/graphics/2015-pace-of-social-change/ [last seen October 15th 2015].

Zurlinden, Urs (2013, 10.12): "Der unbegrenzte Eifer des Gesetzgebers", in *Tagesanzeiger* http://www.tagesanzeiger.ch/schweiz/standard/Der-unbegrenzte-Eifer-des-Gesetzgebers/story/11470997 [last seen October 15th 2015].

Zurlinden, Urs (2015a, 02.26): "Die Suppe köchelt mit Gesetzen und Vorschriften", in *Handelszeitung* http://www.handelszeitung.ch/bildergalerie/die-suppe-koechelt-mit-gesetzen-und-vorschriften [last seen October 15th 2015].

Zurlinden, Urs (2015b, 03.10): "Brouet de lois et d'ordonnances légales" in : *Le Temps* http://www.letemps.ch/interactive/2015/cuisine/ [last seen April 16th 2015].

Zurlinden, Urs (2015c, 03.10): "L'inflation de lois vue à travers la cuisine d'un restaurant" in *Le Temps* http://www.letemps.ch/Page/Uuid/095a80c8-c69c-11e4-959d-74804f4bcbe7/Linflation_de_lois_vue_à_travers_la_cuisine_dun_restaurant [last seen April 16th 2015].

User-Experience von Spracharchiven: Eine Neubewertung der Interaktion von Archiv und Nutzern.

Rau, Felix
f.rau@uni-koeln.de
Universität zu Köln, Deutschland

Blumtritt, Jonathan
jonathan.blumtritt@uni-koeln.de
Universität zu Köln, Deutschland

In den letzten 15 Jahren hat sich weltweit eine aktive und diverse Landschaft digitaler Spracharchive entwickelt. Von diesen Archiven wird eine kontinuierlich wachsende Menge an digitalen Audio- und Videodaten gespeichert und zugänglich gemacht. Dabei wird immer deutlicher, dass die Nachnutzung der in den Spracharchiven archivierten Forschungsdaten noch hinter den Erwartungen zurück bleibt und im Moment die vielleicht größte Herausforderung für diese Institutionen darstellt. Eine geringe Nachnutzung von Forschungsdaten

ist allerdings ein Problem, das nicht nur digitale Spracharchive betrifft.

In unserem Beitrag präsentieren wir unsere Erfahrungen aus Betrieb und Ausgestaltung eines kürzlich neugegründeten Spracharchivs sowie Ergebnisse unserer Untersuchung der User Experience von Spracharchiven und diskutieren den Einfluss der Konzeption und Ausgestaltung der Nutzererfahrung auf die Nachnutzung der Forschungsdaten. Unsere Präsentation fokussiert auf die konzeptuelle Modellierung des Prozesses der Archivierung, Annotation und Nutzung von audio-visuellen Sprachdaten. Die Probleme und Lösungen sind aus unserer Sicht aber ebenso relevant für andere Services und Plattformen in digitalen Forschungsinfrastrukturen.

Digitale Archive und Forschungsdatenzentren sind nicht zuletzt auch Einrichtungen, in denen der Gegenwert aufwändiger Forschungsförderung gesichert und vorgehalten wird. Nationale und internationale Förderer finanzieren seit Jahren direkt oder indirekt in allen fachgebundenen Förderlinien die Erhebung von Forschungsdaten. Gleichzeitig werden Millionen in den Aufbau von Kompetenzzentren, Datenzentren und Forschungsinfrastrukturen für die Geisteswissenschaften investiert. Mit dieser Förderung ist auch die Hoffnung verbunden, dass Datenarchivierung nicht nur Vorhaltung für die Nachwelt leistet, sondern die Bereitstellung von Forschungsdaten mittel- und kurzfristig vielfältige positive Effekte entfaltet. Explizites Ziel ist es unnötige Redundanz bei der Datenerhebung zu verhindern, den Austausch unter den Forschern zu fördern und zu beschleunigen sowie Input für neue Forschungsfragen, Methoden und Verfahren zu schaffen. Jüngst hat die DFG mit der Förderlinie “Forschungsdaten in der Praxis” eine Ausschreibung veröffentlicht, die gezielt dazu anregen soll, “Forschungsfragen überwiegend durch eine Sekundär- bzw. Nachnutzung verfügbarer Forschungsdaten zu bearbeiten”. Nach einer initialen Förderzeit fragen die Geldgeber damit nun vermehrt nach dem „return on investment“, wobei die Dynamik, die aus der Verfügbarmachung von Forschungsdaten erwachsen sollte, in vielen Fällen partiell hinter den Erwartungen zurückbleibt.

Digitale Spracharchive sind Teil dieser Landschaft an Forschungsdatenzentren und haben das spezifische Ziel audio-visuelle Sprachdaten und Dokumente zu sichern und auf dieser Basis Wissensgenerierung zu ermöglichen und zu unterstützen. Ein Spracharchiv ist in diesem Sinn eine Plattform, die zwischen Produzenten und Konsumenten von Primärdaten vermittelt, so dass diese direkt oder indirekt interagieren können. Den datenproduzierenden Forschern ermöglicht das Archiv Audio- und Videoaufnahmen menschlicher Kommunikation zu archivieren und idealerweise web-basiert zugänglich zu machen. Auf der anderen Seite werden Forscher, Sprachgemeinschaften und die weitere Öffentlichkeit in die Lage versetzt, diese Daten aufzufinden, zu betrachten, herunterzuladen und weiterzuverwenden und auf dieser Grundlage neues Wissen zu generieren. Um diesen Austausch zu unterstützen, haben die verschiedenen Spracharchive komplexe Webplattformen entwickelt.

Spracharchive in den verschiedenen Länder sind aus unterschiedlichen Organisationen hervorgegangen und haben sich in sehr diversen Kontexten entwickelt. Einige digitale Spracharchive sind aus vor-digitalen Archiven entstanden, während andere Archive seit ihrer Gründung ausschließlich mit digitalen Daten umgehen. Darüberhinaus sind manche Archive relativ alleinstehende Institutionen, während andere Teil größerer Institutionen sind. Diese Einbettung in größere Netzwerke hat wiederum Einfluss auf die Ausgestaltung der Archive. In den letzten Jahren kam dazu noch eine Integration verschiedener Archive in nationale und übernationale Forschungsinfrastrukturen wie CLARIN-D hinzu. Dieser Prozess hatte Einfluss auf so diverse Aspekte wie Metadatenformate oder die Implementation von Webservices.

Unsere Erfahrungen aus Betrieb und Ausgestaltung eines kürzlich neugegründeten Spracharchivs legen nahe, die Interaktion zwischen Spracharchiven und ihren Nutzern neu zu überdenken und weiterzuentwickeln. Die vorherrschenden Konzepte der Interaktionsgestaltung sind das Produkt einer 15-jährigen Entwicklung und der Förderungskontexte, in denen diese Archive gewachsen sind. Unsere Präsentation nimmt die Webplattformen der verschiedenen Spracharchive als Ausgangspunkt unserer Diskussion. Die unterschiedliche Gestaltung der User Interfaces, Schnittstellen und Funktionalitäten ist der konkrete Ausdruck unterschiedlicher Konzepte und Schwerpunkte. Letzendlich spiegeln die teilweise sehr unterschiedlichen Ausgestaltungen der Webplattformen unterschiedliche Annahmen über die Bedürfnisse der Forscher und Forscherinnen und die Funktion von Archiven im Forschungsprozess wieder. Die Diversität der Plattformen demonstriert somit große Unterschiede in den oft impliziten Annahmen, die wir diskutieren werden.

Im Rahmen der Planung eines Zentrums für audio-visuelle Daten haben wir nun begonnen die User-Experience-Strategie und das UX-Designs unseres Spracharchivs grundlegend zu überarbeiten. Dafür haben wir das User-Interface und die User-Experience sowohl unseres Web-Front-Ends, als auch der Benutzeroberflächen anderer Spracharchive analysiert. Darüber hinaus haben wir Interviews mit Nutzern unseres Archivs sowie potentiellen neuen Nutzern durchgeführt. In diesem Vortrag berichten wir vom Prozess und den Ergebnissen dieser Arbeit. Damit geben wir sowohl direkten Einblick in die Planung eines Zentrums für audio-visuelle Daten, als auch in die Probleme und Herausforderungen, die sich uns gestellt haben.

Unsere Untersuchungen zeigen, dass die Konzeption der Interaktion bestehender digitaler Spracharchive durch die Sichtweise der Archivbetreiber bestimmt wird. Interessen der Archivnutzer werden oft erst nachträglich berücksichtigt. Dabei dominiert durchgehend die Perspektive der datenproduzierenden Forscher,

während sich die Interessen der Datenkonsumenten in der Struktur der Webportale kaum niederschlagen. Auf der anderen Seite zeigen unsere Interviews und die praktische Erfahrung aus dem Support von Archivnutzern, dass Produzenten und Konsumenten fundamental unterschiedliche Bedürfnisse haben und damit auch die Interaktion mit dem Archiv grundsätzlich anders konzeptualisieren.

Das entscheidende Ergebnis unserer Untersuchung ist somit, dass es keine allgemeine Nutzerrolle in Bezug auf Spracharchive gibt. Vielmehr muss der fundamentale Unterschied zwischen den Interessen und Erwartungen von Datenproduzenten und Datenkonsumenten anerkannt werden. Datenproduzenten möchten die Ergebnisse ihrer Arbeit präsentieren und den Zugang zu den von ihnen archivierten Daten kontrollieren. Letzendlich ist es das Ziel des Produzenten wissenschaftliche Anerkennung für die Erhebung und Kuration der Daten zu erhalten. Die Datenkonsumenten streben Zugang zu einem Datensatz an, der das Potenzial hat ihnen Informationen zu liefern, um eine bestimmte Fragestellung zu bearbeiten. Dies bedarf verlässlicher Findmechanismen, die die Identifikation entsprechender Datenmengen erlauben. Darüberhinaus müssen Datensätze möglichst offen zugänglich sein. Falls dies aus rechtlichen oder ethischen Gründen nicht möglich ist, müssen die Prozesse um Zugang zu erlangen gut dokumentiert, transparent und klar geregelt sein. Aus Sicht des Konsumenten stellen Zugangsbeschränkungen Hürden dar, die einen Datensatz weniger attraktiv machen.

Produzenten und Konsumenten stellen somit grundverschiedene User-Typen dar, deren Interaktion mit dem Archiv durch unterschiedliche Bedürfnisse und Erwartungen geleitet wird.

Die Nichtberücksichtigung der Konsumenten-Perspektive beim UX-Design führt aus unserer Sicht zu einer geringeren Rezeption der archivierten Daten und stellt damit einen der Hauptgründe für eine fehlende Nachnutzung dar. Dadurch wird wiederum der niedrigen Stellenwert von Datenpublikationen in der Sprachforschung verfestigt. Eine stärkere Berücksichtigung der User-Experience von Datenkonsumenten scheint uns wesentlich für eine erfolgreiche Weiterentwicklung digitaler Spracharchive zu sein. Verbesserungen in der Auffindbarkeit und Zugänglichkeit von Daten und bei der Referenzierung und Zitierung von Datensätzen werden zu einer größeren Relevanz von Primärdaten in der wissenschaftlichen Praxis führen. Darüber hinaus ermöglicht eine Überarbeitung der User-Experience ein Öffnung der Archive zu einer breiteren Öffentlichkeit. Letztendlich müssen Spracharchive sich von Institutionen der reinen Kuration hin zu partizipatorischen Plattformen, die den direkten und einfachen Austausch zwischen Datenproduzenten und Datenkonsumenten ermöglichen, entwickeln.

Bibliographie

Austin, Peter K. (2011): „Who uses digital language archives?“, in: *Endangered Languages and Cultures (Blog)* http://www.paradisec.org.au/blog/2011/04/who-uses-digital-language-archives/ [letzter Zugriff 13. Oktober 2015].

Beagrie, Neil / Lavoie, Brian / Woollard, Matthew (2010): *Keeping Research Data Safe 2*. Final Report http://www.webarchive.org.uk/wayback/archive/20140615221405/http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf [letzter Zugriff 12. Oktober 2015].

Bow, Catherine / Christie, Michael / Devlin, Brian (2014): „Developing a Living Archive of Aboriginal Languages“, in: *Language Documentation & Conservation* 8: 345-360.

DFG (2015): *Ausschreibung: Forschungsdaten in der Praxis*. Deutsche Forschungsgemeinschaft e.V. Bonn http://www.dfg.de/foerderung/programme/infrastruktur/lis/lis_foerderangebote/ausschreibung_forschungsdaten/index.html [letzter Zugriff 14. Oktober 2015].

Holton, Gary (2012): „Language archives: They're not just for linguists any more“, in: Seifart, Frank / Haig, Geoffrey / Himmelmann, Nikolaus P. / Jung, Dagmar / Margetts, Anna / Trilsbeek, Paul (eds.): *Potentials of Language Documentation: Methods, Analyses, and Utilization*. Honolulu: University of Hawai'i Press 111-117.

Nathan, David / Austin, Peter K. (2014) *Language Documentation and Description, 12: Special Issue on Language Documentation and Archiving*. London: SOAS 4-16.

Schwiertz, Gabriele (2012): „Online presentation and accessibility of endangered languages data: The General Portal to the DoBeS Archive“, in: Seifart, Frank / Haig, Geoffrey / Himmelmann, Nikolaus P. / Jung, Dagmar / Margetts, Anna / Trilsbeek, Paul (eds.): *Potentials of Language Documentation: Methods, Analyses, and Utilization*. Honolulu: University of Hawai'i Press 126–128.

Trilsbeek, Paul / König, Alexander (2014): „Increasing the future usage of endangered language archives“, in: Nathan, David / Austin, Peter K. (eds.): *Language Documentation and Description* 12: Special Issue on Language Documentation and Archiving. London: SOAS 151-163.

Der falsche Quijote? Autorschaftsattribution für spanische Prosa der frühen Neuzeit.

Rißler-Pipka, Nanette
nanette.rissler-pipka@gmx.de
Universität Siegen, Deutschland

Hintergrund

Das bis heute bekannteste Werk der spanischen Literatur ist weltweit als „Don Quijote“ von Miguel de Cervantes geläufig. Dass Cervantes aber, wie kurz zuvor Mateo Alemán, mit einem Fälscher zu kämpfen hatte, der den ersten Band des „Quijote“ ungefragt weiter dichtete und ein Jahr vor der eigenen Fortsetzung durch Cervantes einen zweiten Band aus eigener Feder unter dem Namen Alonso Fernández de Avellaneda heraus brachte, wissen die wenigsten Leser des „Quijote“. Es ist im strengen Sinne des Wortes auch keine Fälschung oder ein Plagiat, sondern die freie Fortsetzung eines erfolgreichen Romans unter eigenem Namen bzw. in diesem Fall unter Pseudonym. Die Identität Avellanedas ist bis heute unbekannt.

Fragestellungen / Stand der Forschung

Angesichts der historischen Publikationsbedingungen in Spanien der frühen Neuzeit (Chartier 2006), stellt sich die Frage, ob eine Autorschaftsattribution mithilfe stilometrischer Methoden auf der Grundlage aktuell zugänglicher digitalisierter Buchausgaben überhaupt möglich ist. Doch auch Patrick Juola und Christopher Coufal nahmen 2010 die Ausgabe des „Don Quijote“, die im Project Gutenberg zugänglich ist als Grundlage ihrer Analyse, die als Ergebnis hatte, dass Cervantes nicht der Autor der letzten 69 (von 74) Kapitel des 2. Bandes des „Don Quijote“ sei (Coufal / Juola 2010). Aus dieser provokanten These entwickelte sich aber keine wissenschaftliche Debatte innerhalb der internationalen Hispanistik und auch die DH-Experten Juola und Coufal erweiterten ihre Fragestellung nicht auf die sich unmittelbar anschließende Frage, wer der Autor des „falschen“ Quijotes von Avellaneda sei. Auch eine weitere statistische Untersuchung der beiden Teile des „Quijote“ von Cervantes berücksichtigt nicht Avellaneda (López Quintero 2011). Die Tatsache, dass sich der Stil Cervantes' innerhalb des zweiten Teils des „Quijote“ ändert, ist in der Hispanistik anerkannt und wird im Allgemeinen mit dem Erscheinen der apokryphen Fortsetzung durch Avellaneda in kausalen Zusammenhang gebracht (Strosetzki 1991: 93; Ehrlicher 2008: 42ff.; Blasco 2007: XVII; Gómez Canseco 2008). Bislang wurde jedoch der Stil Cervantes' gerade im Vergleich zu seinem Nachahmer Avellaneda zumeist als in jeder Hinsicht überragend dargestellt (vgl. zu einer kompakten Darstellung dieser Missachtung Avellanedas: Alvarez Roblin 2014). Erst durch die jüngste Reihe von neuen Ausgaben der Avellaneda-Fortsetzung wird dessen literarische Leistung in der Fachwelt anerkannt (Gómez Canseco 2014; Alvarez Roblin 2009; Suárez Figaredo 2014; López-Vázquez 2011). Mit dem wachsenden Interesse an Avellaneda nimmt auch die Suche nach dessen Identität mithilfe digitalisierter Korpora zu. Zum größten Teil stützen sich diese Autorschaftsattributionen auf schlichte Recherchemöglichkeiten von CORDE (Corpus diacrónica del español), CREA (Corpus de Referencia del Español Actual) und GoogleBooks (Suárez Figaredo 2011; Madrigal 2009; López-Vázquez 2011; Blasco 2005; Jiménez 2007) oder auf philologisch-historische Recherchen (Cruz Casado 2008; Sánchez Portero 2006). Insgesamt werden im Laufe der Recherche nach der wahren Identität Avellanedas 39 Namen ins Spiel gebracht. Von 18 dieser Kandidaten liegen digitalisierte Texte frei zugänglich vor. Dennoch nutzt keiner der Autorschaftsdetektive aktuelle Methoden der DH zur Autorschaftsattribution (wie z. B. JGAAP oder Stilometrie mit R; vgl. Juola 2012; Eder 2015). Neben der Autorschaftsattribution bzgl. Avellanedas apokrypher Fortsetzung des „Quijote“, stellt die Hauptfrage dieses Papers, die stilistische und stilometrische Unterscheidung zwischen Cervantes und Avellaneda dar. Es kann gezeigt werden, dass nur mithilfe stilometrischer Methoden, die festgefahrene Fachdiskussion neue Perspektiven und unerwartete Ergebnisse erhält.

Die Methode

Mithilfe des stylo-Pakets für R (Eder / Rybicki 2011), das für spanischsprachige Texte noch vergleichsweise wenig getestet wurde, sollen zum einen die vorliegenden Autorschaftsattributionen falsifiziert und die Fragen nach stilistischer Nähe zwischen Cervantes, Avellaneda und anderen zeitgenössischen Autoren spanischer Prosa geklärt werden. Somit wird auch die Methode hinlänglich ihrer Komptabilität mit spanischsprachigem Korpus überprüft. Dazu konnte die neueste Version des stylo-package (0.6.0) und die von Jannidis et al. (2015) vorgestellte Cosine Distance genutzt werden.

Zu diesem Zweck wurde zunächst ein passendes Korpus erstellt, das repräsentativ für die Zeit von 1585-1630 spanische Prosawerke enthält und sich auch aus den Kandidaten für die Autorschaft des apokryphen „Quijote“ zusammensetzt. Die Texte stammen aus digitalen Editionen von cervantesvirtual, Wikisource und Project Gutenberg. Sie wurden einheitlich in plain-text-Format abgespeichert und von Textteilen, die nicht von selben Autor stammen (wie z. B. einleitende Bemerkungen des Herausgebers) befreit.

Um zunächst ein sicheres Set zu haben, das sowohl die Autorschaft als auch Genre und Epoche betreffend vergleichbar ist, wurden zunächst nur 4 Autoren mit unterschiedlich vielen Texten ausgewählt (insgesamt 32 Texte). Eine Vergleichbarkeit die Textlänge betreffend hätte bedeutet entweder nur die Novellen oder nur die

Romane miteinander zu vergleichen. Da es auch mit unterschiedlicher Textlänge zu guten Ergebnissen kam, wurde auf diese Angleichung verzichtet (vgl. Eder 2010). Das Korpus wurde mit zwei verschiedenen und anerkannten Distanzmaßen (Eder's Delta und Cosine) und 100-5000 MFW als Cluster-Analyse ausgewertet. Schrittweise wurden dann weitere Kandidaten hinzugefügt und die Ergebnisse bewertet. Die Problematik, die sich dabei ergab, war, dass es wenig Sinn macht, Autoren mit ins Korpus zu nehmen, von denen nur ein Textbeispiel vorliegt, da diese keinem „Partner" im Dendrogramm zugeordnet werden können und somit fälschlicherweise eigentlich weiter voneinander entfernte Texte zusammen geclustert dargestellt werden. Um dieses Problem der Cluster-Analyse zu umgehen, wurde im Vergleich eine Principle Component Analysis (PCA) durchgeführt, die eine bessere Darstellung der Distanzen zwischen den einzelnen Texten im Raum zeigt.

Die Ergebnisse

Im ersten Korpus mit 32 Texten funktioniert die Zuordnung sehr gut. Die Cervantes-Texte sind trotz ihrer starken Größenunterschiede (Novellen mit ca. 7000 und Romane mit ca. 200.000 Wörtern) klar zusammen geclustert.

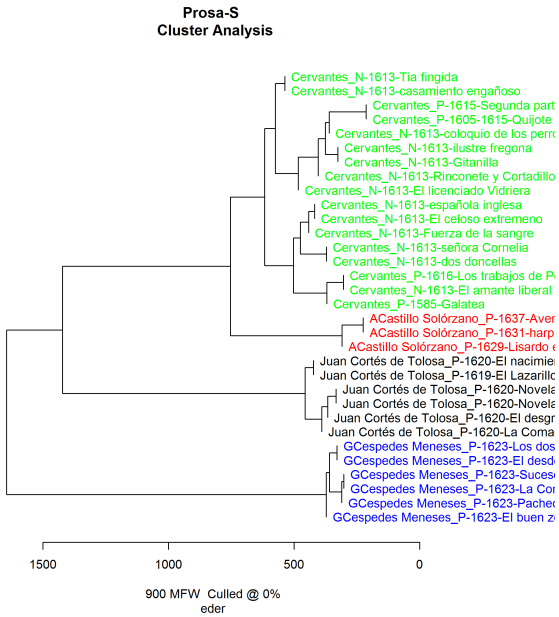


Abb.1: Cluster Analysis mit R (stylo), 900 MFW, Eder's Delta

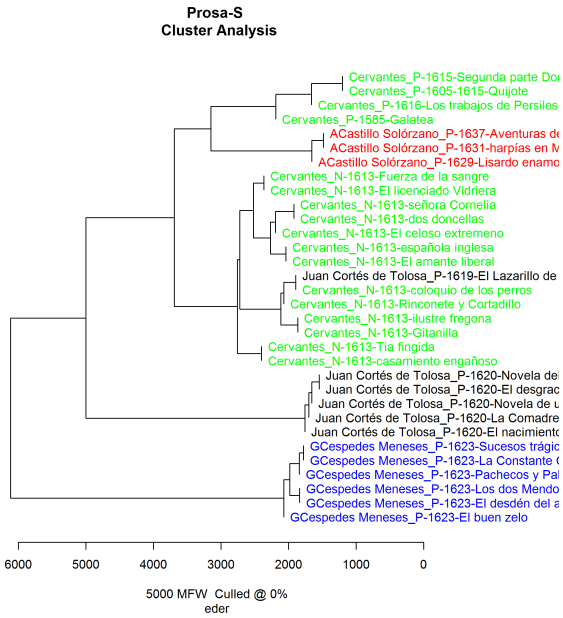


Abb. 2: Cluster Analysis mit R (stylo), 5000 MFW, Eder's Delta

Ab 900 MFW und darunter ist die Autorschaftszuordnung mit Eder's Delta einwandfrei, bei darüber liegenden Zahlen schlich sich beständig der Nachahmer-„Lazarillo" von Cortés de Tolosa und auch der Block der Werke von Castillo Slórzano zwischen die Cervantes-Werke. Zum Vergleich wurde der gleiche Versuch mit Cosine Distance durchgeführt und brachte keine größere Veränderung in der Darstellung. Weiterhin blieb der Roman von Cortés de Tolosa hartnäckig bis zum 300 MFW unter den Cervantes-Werken, jedoch konnte Castillo Solórzano früher heraus genommen werden.

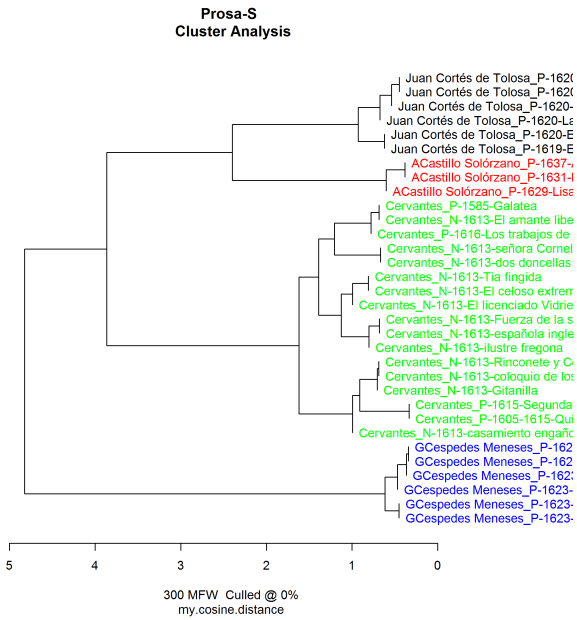


Abb. 3: Cluster Analysis mit R (stylo), 300 MFW, Cosine Distance

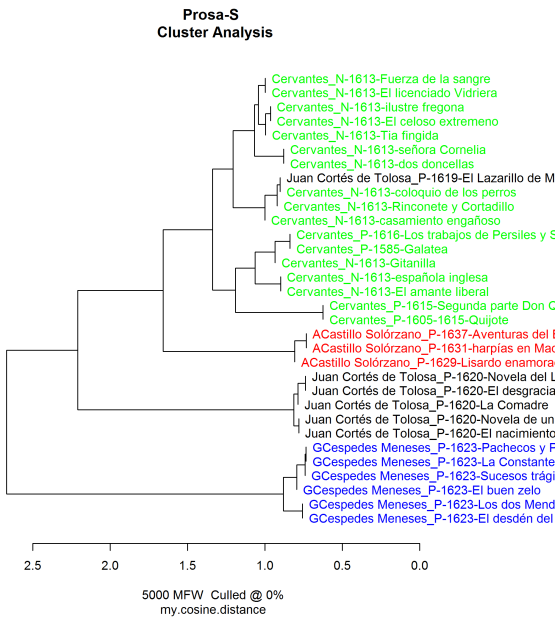


Abb. 4: Cluster Analysis mit R (stylo), 5000 MFW, Cosine Distance

Interessant ist hier, dass die PCA mit denselben Variablen zeigt, wie weit entfernt der „Lazarillo“ von Cortés de Tolosa doch von den übrigen Werken Cervantes‘ entfernt ist (und zwar bei allen 100-5000 MFW):

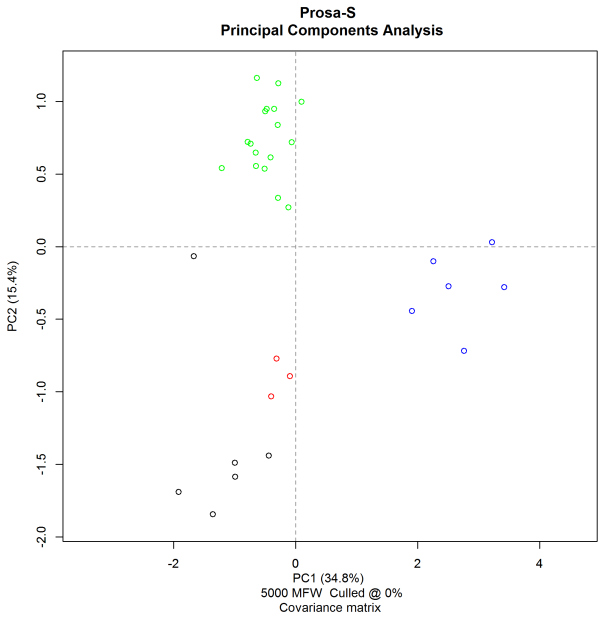


Abb. 5: PCA mit R (stylo), 5000 MFW, Cosine Distance

Deutlich wird vor allem, dass der “Lazarillo”, den Cortés Tolosa in Nachahmung des Originals (anonym 1554) geschrieben hat, sehr von seinem übrigen Werk abweicht, aber ebenso noch deutlich von den Cervantes-Werken entfernt ist. Das erklärt sich allerdings leicht durch die sprachliche Unterscheidung der vorliegenden Edition des Werkes, die eine ältere Form des Kastilischen verwendet und daher in der Vergleichbarkeit eingeschränkt bleibt.

Spannender werden die Ergebnisse, wenn man Avellaneda und weitere Kandidaten hinzunimmt:

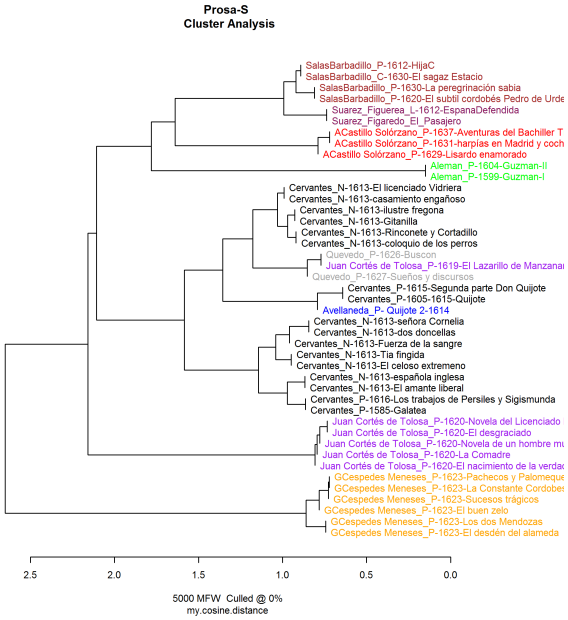


Abb. 6: Cluster Analysis mit R (stylo), 5000 MFW, Cosine Distance (Korpuserweiterung: 43 Texte)

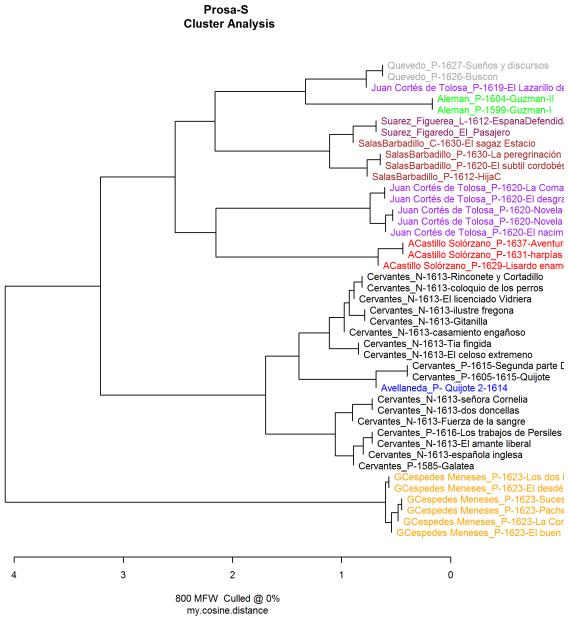


Abb. 7: Cluster Analysis mit R (stylo), 800 MFW, Cosine Distance

Auch mit der Korpuserweiterung bleiben die Zuordnungen relativ stabil (außer Cortés de Tolosa, wie zuvor). Keiner der möglichen Kandidaten wird jedoch Avellaneda zugeordnet, sondern der apokryphe „Quijote“ wird mit den anderen beiden Teilen von Cervantes zusammen geclustert – dies bleibt über 100-5000 MFW konstant. Auch mit dem einer PCA bestätigt sich dieses Ergebnis (vgl. Fig. 8: der blaue Punkt bei den schwarzen zeigt Avellaneda im Umkreis der Cervantes-Werke):

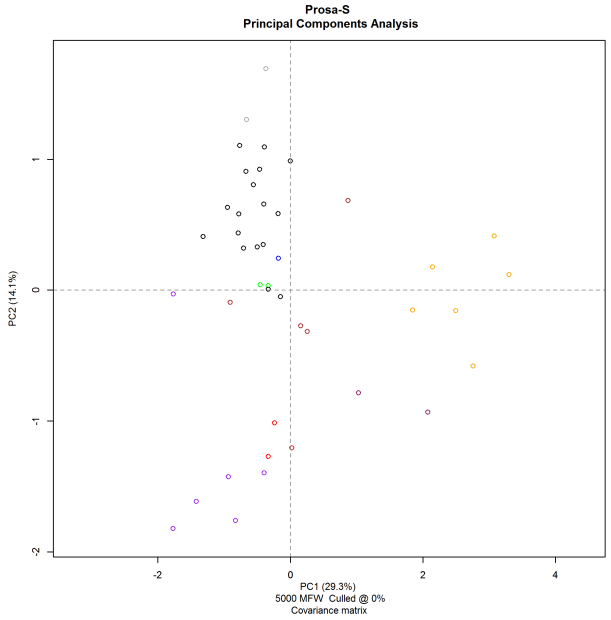


Abb. 8: PCA mit R (stylo), 5000 MFW, Cosine Distance

In einer nächsten Korpuserweiterung nehmen wir weitere Kandidaten mit Prosabeispielen hinzu, von denen nur ein Text zur Verfügung steht.

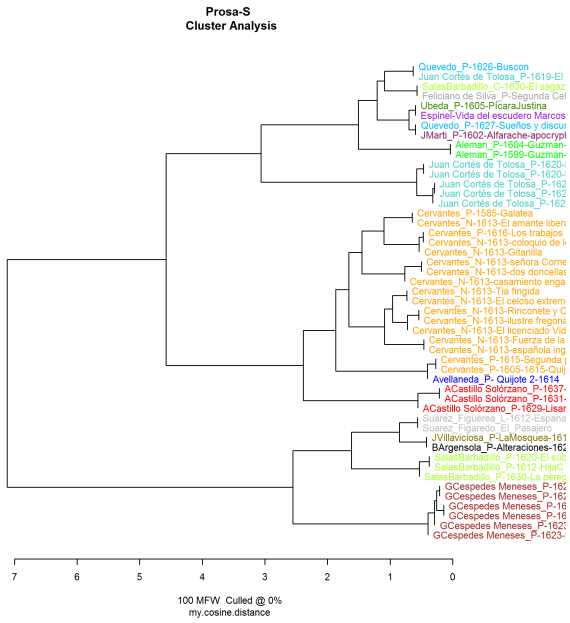


Abb. 9: Cluster Analysis mit R (stylo), 100 MFW, Cosine Distance (Korpuserweiterung: 49 Texte)

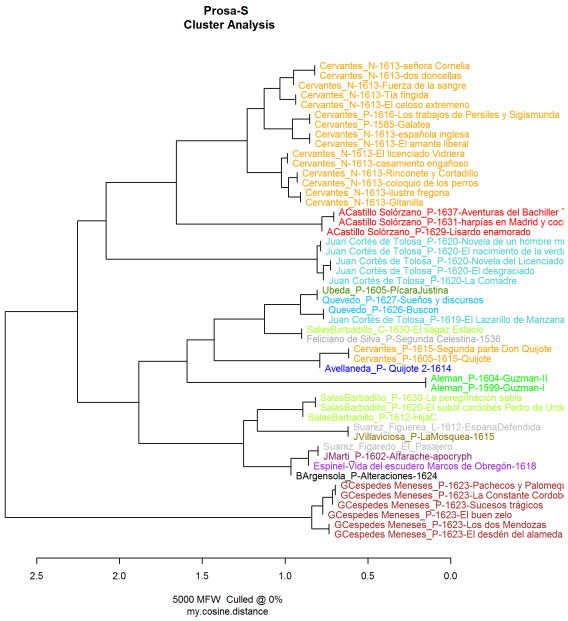


Abb. 10: Cluster Analysis mit R (stylo), 5000 MFW, Cosine Distance

Es wird schnell deutlich, dass die Einzelgänger unter den Texten, das zuvor stabile Gefüge auseinanderbringen und offensichtlich nicht für die statistische Untersuchung zu gebrauchen sind. Erstaunlich bleibt jedoch, dass selbst in diesem sehr vagen Clustering (je nach MFW schieben sich die einzelnen Texte mal dort mal dort hin), die drei „Quijote“-Texte konsistent bei 100-5000 MFW ein Cluster bilden. Kein anderer Kandidat schafft es in die Nähe von Avellaneda.

Auch die PCA bringt keine entscheidenden Vorteile gegenüber der Cluster Analysis. Espinel und Úbeda gruppieren sich zwar in die Cervantes-Gruppe, aber nicht direkt zu Avellaneda.

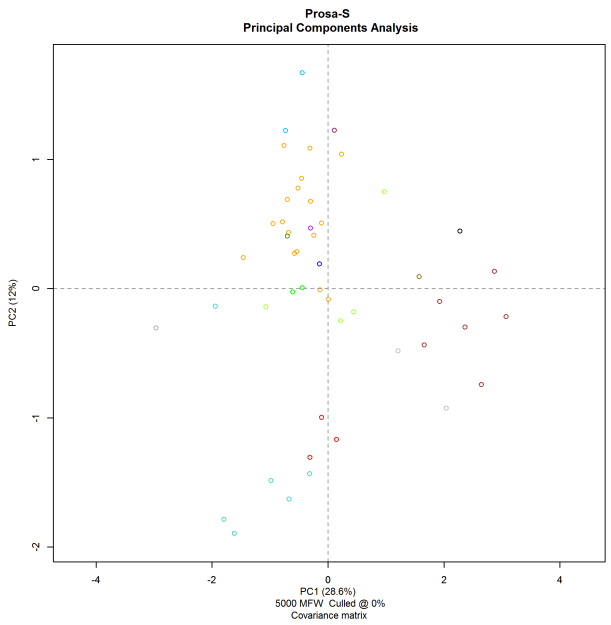


Abb. 11: PCA mit R (stylo), 5000 MFW, Cosine Distance

Schlussfolgerungen

Wenn überhaupt ein Autorname anstelle desjenigen Avellanedas genannt werden sollte, müsste es nach den vorliegenden Ergebnissen Cervantes selbst sein, der in einem gekonnten Spaß ganz im Sinne seines Quijote die Leser mit der eigenen falschen Fortsetzung an der Nase herum führt. Oder aber beide Autoren haben ihren Stil gegenseitig aufeinander abgestimmt, dass sie sich derart im Wortgebrauch ähneln. Es würde sich als folgende Untersuchung ein rolling delta anbieten, um eine kollaborative Autorschaft ausmachen zu können und um den vorgeblichen Stilwechsel im zweiten Teil von Cervantes' „Quijote“ genauer definieren und mit Avellaneda in Zusammenhang bringen zu können.

Bibliographie

Alvarez Roblin, David (2014): *De l'imposture à la création*. Le Guzmán et le Quichote apocryphes. Madrid : Casa de Velázquez.

Avellaneda, Alonso Fernández de (2014): *Segundo tomo del ingenioso hidalgo don Quijote de la Mancha*. Edición de Gómez Canseco, Luis. Madrid: Real Acad. Española, Centro para la Ed. de los Clásicos Españoles.

Avellaneda, Alonso Fernández de (2011): *El Quijote apócrifo*. Edición de López-Vázquez, Alfredo Rodríguez. Madrid: Cátedra.

Avellaneda, Alonso Fernández de (2014): *El Quijote apócrifo* (= Lemir 18: Conmemoración iv Centenario del Quijote de Avellaneda) http://parnaseo.uv.es/Lemir/Revista/Revista18/Textos/06_Quijote_Avellaneda_Figaredo.pdf [letzter Zugriff 15. Oktober 2015].

Blasco, Javier (2005): “La lengua de Avellaneda en el espejo de ‘La pícara Justina’”, in: *Boletín de la Real Academia Española* 85, 291-292: 53-109 http://uvadoc.uva.es/bitstream/10324/2436/1/LA%20LENGUA%20DE%20AVELLANEDA...%28Javier%20Blasco%29.pdfOCR.pdf [letzter Zugriff 15. Oktober 2015].

Chartier, Roger (2006): “Materialidad del texto, textualidad del libro”, in: *Orbis Tertius* 11, 12: http://www.orbistertius.unlp.edu.ar/article/view/OTv11n12a01/3774 [letzter Zugriff 8. Januar 2016].

Coufal, Christopher / Juola, Patrick (2010): “Authorship Discontinuities of El Ingenioso Hidalgo don Quijote de la Mancha as detected by Mixture-of-Experts”, in: *Digital Humanities 2010* Conference Abstracts, London http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-755.html [letzter Zugriff 15. Oktober 2015].

Cruz Casado, Antonio (2008): “Revisión de una hipótesis. Juan Valladares de Valdelomar; autor del ‘Quijote’”, in: Dotras Bravo, Alexia (ed.): *Tus obras los rincones de la tierra descubren*. Actas del VI congreso internacional de la Asociación de Cervantistas, Alcalá de Henares 29-44 http://cvc.cervantes.es/literatura/cervantistas/congresos/cg_VI/cg_VI_18.pdf [letzter Zugriff 15. Oktober 2015].

Eder, Maciej (2010): “Does size matter? Authorship attribution, short samples, big problem”, in: *Digital Humanities 2010*. Conference Abstracts, London: 132-35.

Eder, Maciej / Kestemont, Mark / Rybicki, Jan (2013): “Stylometry with R: a suite of tools”, in: *Digital Humanities 2013*. Conference Abstracts, University of Nebraska, Lincoln 487-89.

Ehrlicher, Hanno (2008): „Der andere Autor im eigenen Werk. Medialisierte Autorschaft bei Mateo Alemán und Miguel de Cervantes“, in: Dünne, Jörg / Moser, Christian (eds.): *Automedialität*. München: Fink 27-52.

Gómez Canseco, Luis (2008): “1614: Cervantes escribe otro ‘Quijote’”, in: Dotras Bravo, Alexia (ed.): *Tus obras los rincones de la tierra descubren*. Actas del VI congreso internacional de la Asociación de Cervantistas, Alcalá de Henares 29-44 http://cvc.cervantes.es/literatura/cervantistas/congresos/cg_VI/cg_VI_05.pdf [letzter Zugriff 15. Oktober 2015].

Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): “Improving Burrows’ Delta – An empirical evaluation of text distance measures”, in: *Digital Humanities 2015*. Conference Abstracts, Sydney http://dh2015.org/abstracts/index.php [letzter Zugriff 15. Oktober 2015].

Jiménez, Alfonso Martín (2007): “Cotejo por medios informáticos de la Vida de Pasamonte y el Quijote de Avellaneda”, in: *Etiópicas* 3: 69-131 http://rabida.uhu.es/dspace/bitstream/handle/10272/1616/b1551216.pdf?sequence=1 [letzter Zugriff 15. Oktober 2015].

López Quintero, Freddy (2011): "Diferencias internas en Don Quijote. Cambios en proporciones y cambios estructurales", in: *Lemir* 15: 259-270 http://parnaseo.uv.es/Lemir/Revista/Revista15/13_Lopez_Freddy.pdf [letzter Zugriff 15. Oktober 2015].

López-Vázquez, Alfredo Rodríguez (2011): "Introducción", in: Avellaneda, Alonso Fernández de: *El Quijote apócrifo*. Edición de López-Vázquez, Alfredo Rodríguez, Madrid: Cátedra 13-84.

Madrigal, José Luis (2009): „Tirso, Lope y el Quijote de Avellaneda", in: *Lemir* 13: 191-250 http://parnaseo.uv.es/lemir/revista/revista13/11_madrigal_jose.pdf [letzter Zugriff 15. Oktober 2015].

Rybicki, Jan / Eder, Maciej (2011): "Deeper Delta across genres and languages: do we really need the most frequent words?", in: *Literary and Linguistic Computing* 26, 3: 315-21.

Sánchez Portero, Antonio (2006): *El autor del 'Quijote' de Avellaneda es Pedro Liñán de Riaza, poeta de Calatayud*. Edición digital por cortesía del autor para la Biblioteca Virtual Miguel de Cervantes, Alicante http://www.cervantesvirtual.com/obra/el-autor-del-quijote-de-avellaneda-es-pedro-lin-de-riaza-poeta-de-calatayud-0/ [letzter Zugriff 15. Oktober 2015].

Strosetzki, Christoph (1991): „Der Roman im Siglo de Oro", in: Strosetzki, Christoph (ed.): *Geschichte der spanischen Literatur*. Tübingen: Niemeyer 84-118.

Suárez Figaredo, Enrique (2011): "Sobre la atribución del Quijote apócrifo a José de Villaviciosa", in: *Lemir* 15: 135-146 http://parnaseo.uv.es/Lemir/Revista/Revista15/05_Suarez_Enrique.pdf [letzter Zugriff 15. Oktober 2015].

Sonification: Vermittlungsansätze zwischen Klang und Information

Roeder, Torsten
torsten.roeder@uni-wuerzburg.de
Universität Würzburg, Deutschland

Das Verhältnis von Klang und Information

In diesem Beitrag geht um das rätselhafte Verhältnis zwischen Klang und Information. Rätselhaft deshalb, weil zwischen der meist konkreten und persistenten „Information" und dem meist unkonkreten, transitorischen „Klang" kaum Verbindungen möglich erscheinen. Wir kennen ein Verhältnis zwischen beiden aus der Beziehung zwischen erklingender Musik und lesbarer Notenschrift. Auch wenn man kaum davon sprechen kann, dass beide eindeutig voneinander ableitbar wären, stehen sie in einem nachvollziehbaren Verhältnis. Dieses ist durch Regeln bestimmt, die durch die Musiklehre festgelegt ist: Zum Beispiel wissen wir dank einer allgemeinen Konvention, dass der Ton „d" im Violinschlüssel auf der zweiten Linie von oben notiert wird, der Ton „h" hingegen auf der dritten; ebenso gibt es Übertragungskonventionen für Zeitmaße, für Lautstärke, für klangliche Parameter und vieles andere. Trotz vieler Unschärfen bei der Übertragung gelingt es in der Regel, von einer Notation ein wiedererkennbares musikalisches Abbild zu erzeugen, während abweichende Wiedergaben unterschiedlichen Interpretationen zuzuschreiben sind. In diesem Beispiel werden also Regeln angewendet, um aus Informationen Klänge zu erzeugen. Wir sind gewohnt, dies als „Kunst" oder „Unterhaltung" zu betrachten und schreiben diesem Phänomen, dem wir den Namen Musik geben, eine enorme gesellschaftliche Bedeutung zu.

Eine Analogie: Visualisierungen

Visualisierung en von Informationen sind derzeit en vogue. Dabei erscheinen Visualisierungen, die sich am Konkreten orientieren – etwa Statistiken, Landkarten, Zeitleisten – fast schon überholt. Es gilt, neue Regeln der Informationsabbildung zu entdecken, die alternative Lesarten erlauben und über die konventionellen Übertragungen auf die typischen Kategorien Globus, Kalender und Diagramm hinausgehen. Von besonderem Interesse sind dabei Beziehungsgeflechte und multidimensionale Darstellungen, um damit nicht-metrische Parameter abbilden zu können. Die aus dem kreativen Umgang mit Daten entstehenden Abbildungen sind vielfältig und die Regeln ihrer Generierung im Grunde nur durch Vorstellungskraft begrenzt; zum Teil erwachsen daraus Grafiken von fast künstlerischer Qualität. Der Erkenntniswert dieser Abbildungen ist vielerorts noch auszuloten, das „Lesen" in solchen Grafiken noch nicht kultiviert, aber zweifellos besteht ein großes Interesse daran, alternative Erkenntnismethoden zu erfinden und zu entdecken.

Hören und Sehen von Daten

Die Wahrnehmung wird allgemein durch den Sehsinn dominiert, während andere Sinne stark zurückgedrängt sind oder auf bestimmte Vorgänge limitiert sind. Als wahr gilt, was man mit eigenen Augen gesehen hat. Der Hörsinn hingegen dient zwar dem Verstehen des gesprochenen Wortes und wird für den Genuss vielschichtig arrangierter Musik eingesetzt, scheint aber für Erkenntnisvorgänge prinzipiell nicht infrage

zu kommen, da er emotional konnotiert ist und somit als völlig subjektiv ausscheidet. Ein analytisches Hören ist zwar erlernbar, jedoch stellt es sich gegen Konventionen oder ist Musikern vorbehalten. Zudem gilt die Auffassung, dass Hörbares ohnehin besser visualisiert wird. Jedoch ist das menschliche Gehör dem Sehsinn in einigen Punkten voraus: Es unterscheidet nicht nur Tonhöhen, sondern auch Lautstärken, Tempo, Klangqualität etc., und ist dadurch in der Lage, eine Vielzahl an Parametern gleichzeitig darzustellen; die plötzliche Veränderung eines Parameters kann dabei eine starke Signalwirkung hervorrufen. Außerdem ist das Gehör ein extrem granularer Sinn: Es nimmt z. B. äußerst geringe zeitliche Abstände wahr, die mit dem Auge nicht mehr nachvollziehbar sind (Hintergrund ist ein chemischer Prozess).

Wäre es denkbar, das Prinzip der Visualisierung – d. h. aus Information werden erfahrbare Bilder – auf die Welt des Klanges zu übertragen, um die Möglichkeiten des Hörsinns für die Datenexploration auszuschöpfen? Das hieße: aus Informationen werden erfahrbare Klänge. „Ausgangspunkt dafür [für Sonifikation] ist die Tatsache, dass der Hörsinn in vielen Fällen ein hohes Potenzial besitzt, zum Sehsinn komplementäre Informationen auf einfache Weise zu vermitteln." (Grond / Schubert-Minski 2009).

Einfache Übertragungen von Information in Klang sind z. B. aus dem Morsecode oder vom Geigerzähler bekannt; gut in Erinnerung dürfte außerdem das akustische Einwahlsignal eines Modems sein. Diese Klänge gehorchen keiner Ästhetik, jedoch müssen sie das auch nicht (um einen Satz von John Cage anzuwenden: „You don't have to call it music, if the term shocks you"). Klang und Informationen verhielten sich dann ähnlich, nur abstrakter, wie Musik und Notation; Klang wäre dann eine mögliche Darstellungsform von Information, nach vorher bestimmten Regeln interpretiert.

Sonifikation

Die Idee der Sonifikation wird innerhalb des Technologiezweiges „Auditory Display" ungefähr seit den 1990ern als Methode verfolgt (vgl. Flowers 2005). Eine von der *International Community for Auditory Display* (ICAD) herausgegebene Definition der Sonifikation lautet: „Sonification [is the] use of nonspeech audio to convey information; more specifically sonification is the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation." (Schoon / Volmer 2012)

Sonifikation hat inzwischen das experimentelle Stadium verlassen und wird erfolgreich eingesetzt, um komplexe Daten (Stichwort Big Data) effektiver auswerten zu können (vgl. Kramer et al. 2010) . Dabei ist es auffällig, dass die Nutzung in den Naturwissenschaften und in der Medizin bereits fortgeschritten ist, in den Digital Humanities hingegen kaum repräsentiert ist, obwohl die Methode grundsätzlich naheliegend wäre. Die bisher genutzten Verfahren (Audifikation, Auditory Graphing u. a.) werden zu sehr unterschiedlichen Zwecken eingesetzt (Schoon / Volmer 2012: 12); in der Regel werden die Sonifikationen automatisch erzeugt, für mediale Zwecke werden sie manchmal aber auch live produziert.

Eine „Einstiegsmethode" ist dabei das Pitch Coding, bei dem Tonhöhen und Codepoints einander zugeordnet werden. Dieses sehr einfache Verfahren lehnt sich an die Logik der Notation an: hohe Werte werden als hohe Töne übertragen, niedrige Werte als tiefe Töne. Als anschauliches (aber spielerisches) Beispiel ist die *Higgs Boson Sonification* (Rao 2015) zu nennen, bei der physikalische Messwerte in Tonwerte übertragen werden. Mehrdimensionales Pitch Coding, bei dem parallele Prozesse modelliert werden, wurde in *What climate change sounds like from the Amazon to the Arctic* (Reubold 2015) umgesetzt. Die Live-Umsetzung ist in diesen Fällen lediglich als Kür anzusehen; den Live-Sonifikationen geht ansonsten üblicherweise die rechnergestützte Modellierung voraus. Eine schnelle rechnergestützte Umsetzung erlaubt z. B. das frei verfügbare Tool *Sonification Sandbox* (Walker 2009), welches das Experimentieren mit mehreren klanglichen Dimensionen anhand einer Wertetabelle erlaubt und sowohl über Kommandozeile als auch GUI steuerbar ist.

Über diese grundlegenden Verfahren hinaus gehen Verfahren, die mit Nutzerinteraktivität arbeiten und neben einer visuellen Darstellung auch akustische Rückmeldungen geben; dabei kann auch Sprache zum Einsatz kommen. Diese Anwendungen zielen vor allem darauf, Personen mit eingeschränkter Sehfähigkeit einen verbesserten Zugang zu Datenabbildungen zu ermöglichen, allerdings sind die Verfahren auch für uneingeschränkt sehfähige Nutzer von grundsätzlichem Interesse. Beispielsweise werden in der *Sonification for Blind Users* (Zhao et al. 2005) Stereo-Effekte für die Umsetzung der geographischen Dimension genutzt. Von höchster Komplexität sind schließlich Soundalgorithmen, die entsprechend der Datenveränderung Tempo und Harmonie nach bestimmten Mustern verändern (Morreale et al. 2013).

Sonifikation führte tatsächlich bereits zu einigen wissenschaftlichen Erfolgen, etwa bei der Analyse von Sonnenstürmen (Alexander 2012), wobei insbesondere die akustische Darstellbarkeit der Granularität der Messdaten bei der Analyse ausschlaggebend waren.

Eine Chance für die Digital Humanities

Das Verfahren der Sonifikation bietet für die Digital Humanities eine Alternative zur Visualisierung, insbesondere im Hinblick auf die Abbildung zeitlicher,

räumlicher und paralleler Prozesse. Die Möglichkeiten der Sonifikation wurden bislang nicht in dem gleichen Maße erschlossen und ausgeschöpft, wie es für Visualisierung bereits im Gange ist. Daher besteht das dringende Desiderat, Sonifikation als in den Naturwissenschaften bereits etabliertes Verfahren endlich auch in den Digital Humanities zu erproben und ihr Potenzial zu entdecken. Der Vortrag gibt einen Überblick über die Methoden und erörtert Anwendungsmöglichkeiten an verschiedenen Beispielen.

Bibliographie

Alexander, Robert (2012): „How A Solar Storm Sounds – Particle Sonification Video“, in: *Space.com* http://www.space.com/14897-solar-storm-sounds-particle-sonification-video.html [letzter Zugriff 15. Oktober 2015].

Flowers, John H. (2005): „Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions“, in: *Proceedings of the 11th International Conference on Auditory Display (ICAD2005)* 406–409 http://www.icad.org/Proceedings/2005/Flowers2005.pdf [letzter Zugriff 15. Oktober 2015].

Grond, Florian / Schubert-Minski, Theresa (2009): „Sonifikation“ http://see-this-sound.at/kompendium/abstract/70 [letzter Zugriff 15. Oktober 2015].

Hermann, Thomas / Hunt, Andy / Neuhoff, John G. (2011): *The Sonification Handbook*. Berlin: COST / Logos.

Kramer, Gregory / Walker, Bruce / Bonebright, Terri / Cook, Perry / Flowers, John H. / Miner, Nadine / Neuhoff, John (2010):„ Sonification Report: Status of the Field and Research Agenda“, in: *Faculty Publications, Department of Psychology, University of Nebraska* http://digitalcommons.unl.edu/psychfacpub/444 [letzter Zugriff 15. Oktober 2015].

Morreale, Fabio / Masu, Raul / De Angeli, Antonella (2013): "Robin: An algorithmic Composer for Interactive Scenarios", in: *Proceedings of the Sound and Music Computing Conference 2013 (SMC 2013)* 207–212.

Rao, Achintya (2015): „What would the Higgs discovery sound like as a heavy-metal song? “ , in: *The Cylindrical Onion* http://cylindricalonion.web.cern.ch/blog/201504/what-would-higgs-discovery-sound-heavy-metal-song [letzter Zugriff 15. Oktober 2015].

Reubold, Todd (2015): „What climate change sounds like from the Amazon to the Arctic“, in: *Ensia* http://ensia.com/videos/what-climate-change-sounds-like-from-the-amazon-to-the-arctic/ [letzter Zugriff 15. Oktober 2015].

Schoon, Andi / Volmar, Axel (2012): *Das geschulte Ohr*. Eine Kulturgeschichte der Sonifikation. Bielefeld: Transcript.

Walker, Bruce N. (2009): *Sonification Sandbox*. Atlanta: Georgia Institute of Technology, http://sonify.psych.gatech.edu/research/sonification_sandbox/ [letzter Zugriff 15. Oktober 2015].

Zhao, Haixia / Plaisant, Catherine / Shneiderman, Ben (2005): *iSonic: Interactive Data Sonification for Blind Users, University of Maryland* https://www.youtube.com/watch?v=8hUIAnXtlc4 [letzter Zugriff 15. Oktober 2015].

Korpushermeneutik - Ansatz und Werkzeug zur Analyse großer Textkorpora

Rüdiger, Jan Oliver
jan.ruediger@uni-kassel.de
Universität Kassel, Deutschland

Der Vortrag fußt auf drei Säulen: Theorie, Forschungspraxis und Hochschullehre. Sie werden im Vortrag einzeln ausgeführt, dann kombiniert.

Theorie: *Korpuslinguistik* mit *Hermeneutik* zu verbinden, ist keine grundsätzlich neue Idee. Die bisherigen Vorschläge (z. B. Haß 2007; Teubert 2006) führen aber in ihrer Konsequenz zu einer einseitig gelagerten Korpuslinguistik, die entweder *corpus-driven* oder *corpus-based* orientiert ist.

Bei Haß (2007) werden wichtige Grundüberlegungen der *Korpus-Hermeneutik* diskutiert. Im Abschnitt Haß (2007: 248-258) erfolgt eine Beispielanalyse, deren Methoden fast ausschließlich dem *corpus-driven* Spektrum zuzuordnen sind. Ermittelte statistische Werte werden zwar interpretiert, jedoch führt dies nicht zu weiteren Forschungskonsequenzen. Gerade aber in der zyklischen Interpretation liegt die Stärke der Korpushermeneutik.

Bei Teubert (2006) ist der Blick auf den Sichtbereich des *corpus-based* Methodenapperats beschränkt. Korpusmaterial dient in dieser Arbeit als eine Art Steinbruch, in dem man nach Belegen schürft. *Text-Mining* ist zwar ein Aspekt der Korpushermeneutik – es darf aber nie das alleinige Merkmal sein.

Daher plädiere ich für grundlegend *neue und praktikable Korpushermeneutik*, die sowohl klassische als auch computergestützte Analyseverfahren vereint. Einen zentralen Punkt nimmt dabei die (Weiter-)Entwicklung des bestehenden Wissens ein. Annahmen, Beobachtungen und Ergebnisse werden zu Wissensmodellen korreliert und durch einen zyklisch organisierten Analyseprozess falsifiziert. Zum jetzigen Zeitpunkt ergeben sich drei grundlegende Forderungen an eine Analyse, wenn Sie unter dem Begriff *Korpushermeneutik* firmieren soll:

Die Analyse muss mehrere, abwechselnde und aufeinander aufbauende Zyklen durchlaufen.

Diese Idee lehnt sich an der bereits von Felder (2016: 124) aufgestellten Beobachtung an: „ *Für die praktische Textanalyse sind beide Herangehensweisen erkenntnisstiftend – insbesondere ein stetiger Wechsel zwischen der relativ induktiven (corpus driven) und der eher deduktiven (corpus-based oder corpus-assisted) Vorgehensweise ist vielversprechend.*“

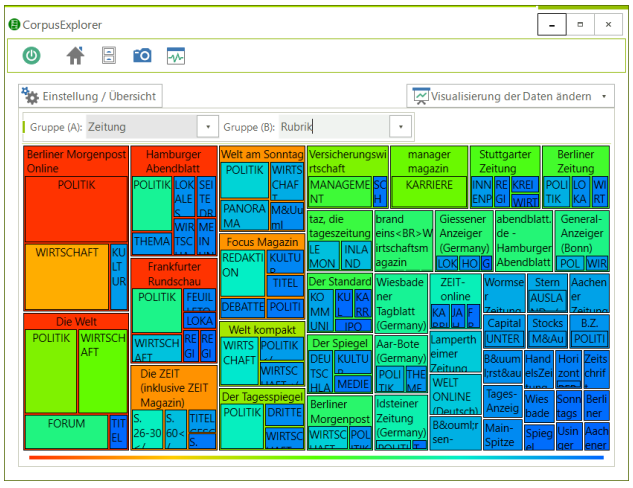
Die Entwicklung von Wissen geschieht durch Falsifikation - wie sie u. a. durch Popper (2005) und Albert (1969) gefordert wird. Daher sind Fragen, Thesen und Methoden so zu wählen, dass Vorwissen geprüft und hinterfragt wird.

Der Analyseprozess ist transparent zu gestalten - transparent in Durchführung und Fehlerbetrachtung – Die Handlungsmöglichkeiten sind aufzuzeigen. Jeder Durchführungszyklus geschieht unter der Prämisse der Falsifikation. Jede Methode ist ebenso kritisch zu hinterfragen wie die daraus resultierenden Ergebnisse. Durch einen Zyklus werden neue Handlungsmöglichkeiten offengelegt, die es zu überprüfen gilt. Daher muss transparent gemacht werden, welche Ansätze mit welchen Mitteln verfolgt werden und welche Fragen am Schluss offenbleiben.

Forschungspraxis: Gerade in den letzten fünf bis zehn Jahren ist die Möglichkeit stark gewachsen, große (linguistische) Datenmengen zu erheben und auszuwerten. Text-/Sprachdaten können fast ohne Limitierung für die Forschung erhoben werden. Die darauf aufbauenden Datenmodelle erreichen eine immer höhere Komplexität. Daher bedarf es neuer Methoden, diese zu strukturieren und teilweise auch zu reduzieren (z. B. durch Algorithmen oder Visualisierungen), damit sie (er-)fassbar für den Anwender werden. Ein Problem bei der Umsetzung der korpushermeneutischen Theorie ist die bisher existierende Softwarelandschaft der Computer-/Korpuslinguistik. Viele Programme sind notwendig, um aus einem einfachen Rohtext ein visuelles Ergebnis zu erzeugen. Die Programme sind teilweise untereinander inkompatibel [1] - oder sie folgen ausschließlich einem der beiden Paradigmen [2]. Im Vortrag wird ein von mir entwickeltes Programm vorgestellt, das diese Arbeit übernimmt und korpushermeneutische Analysen ermöglicht. Der *CorpusExplorer* ist kostenfrei verfügbar und übernimmt alle nötigen Arbeitsprozesse – angefangenen bei der Textaufbereitung, Trennung von Text und Metadaten, Annotation [3], bis hin zur Auswertung und Visualisierung (über 30 unterschiedliche Auswertungsmodule). Alles mit einem Tool, mit nur wenigen Mausklicks und vereint unter einer intuitiven Benutzeroberfläche. Der CorpusExplorer erlaubt sowohl corpus-driven als auch corpus-based Analysen und durch die zyklische Verschränkung der Werkzeuge die angestrebte korpushermeneutische Analyse. Im Vortrag wird auf konkrete Praxisbeispiele eingegangen und gezeigt, wie sich eine korpushermeneutische Analyse entwickelt. Ein exklusiver Vortragspunkt wird sein, dass neben dem Programm das CorpusExplorer-Framework erstmalig vorgestellt wird. Mit diesem werden zwei Dinge möglich. Zum einen kann der CorpusExplorer mit eigenen Funktionen erweitert werden (z. B. schreiben / anbinden neuer Tagger / Dateiformate, entwickeln eigener Analysemodule, uvm.). Zum anderen kann man den CorpusExplorer in eigene Programme integrieren. Ein Teil des Quellcodes (Import- / Export-Funktion), sowie Quellcode von An-Projekten wurde bereits veröffentlicht. Der Quellcode des Frameworks wird nach Abschluss des Promotionsprojekts freigegeben.

Hochschullehre: Eines der komplexesten Probleme, vor dem Dozenten und Institute stehen, die Korpuslinguistik in der Lehre praktizieren möchten und nicht oder nur bedingt auf Kompetenzen im Bereich Informatik bzw. Computerlinguistik zurückgreifen können, ist der immense *Toolchain*, der für einen erfolgreichen Seminarbetrieb erforderlich ist. Der *CorpusExplorer* bietet hier eine praktikable Lösung für alle, die schnelle Ergebnisse erzielen möchten. Selbst Studenten in den ersten Semestern können so in die Forschung hineinschnuppern und ihre eigenen Forschungsfragen selbstständig erkunden. Dabei stehen Forschung, empirisches Arbeiten und Auswertung/Ergebnisvisualisierung im Vordergrund, nicht aber das verwendete Programm. Der Vortrag wird Einblicke in den Seminaralltag mit dem CorpusExplorer sowie Anregungen geben, die mit den Hörern diskutiert werden können.

Bildanhang (Screenshots CorpusExplorer v2.0):
Annotationsansicht
Links: Korpora & Dokumente
Mitte: Annotiertes Dokument mit gewählter Hervorhebungen
Rechts: Gewählte Hervorhebungen (Annotationen)
Unten: Verfügbare Module des CorpusExplorers



Korpusverteilung

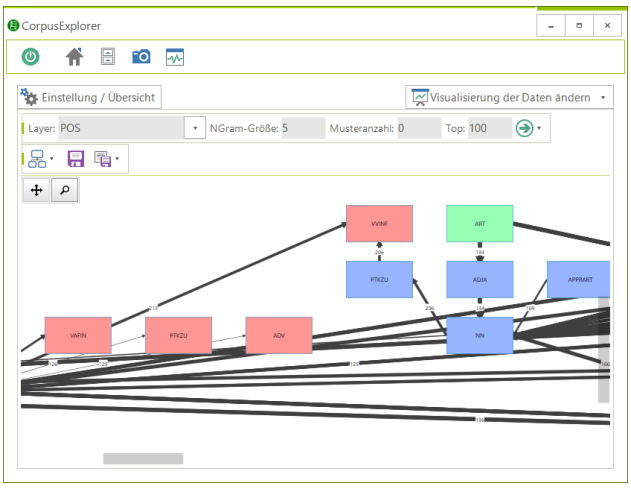
Zu sehen ist ein Kreuzvergleich von Dokumentmetadaten. Eingenommene Fläche und Farbe (warm > kalt) sind bedeutungstragend Begriffspaare / Oppositionswörter kontrastieren

Beispiel: Frau vs. Mann aus einem Zeitungskorpus (Frauenquote vs. Quotenfrau 2010-2014) via LexisNexis

Grün: Kollokatoren tendenziell Syrien

Schwarz: Gemeinsame Kollokatoren

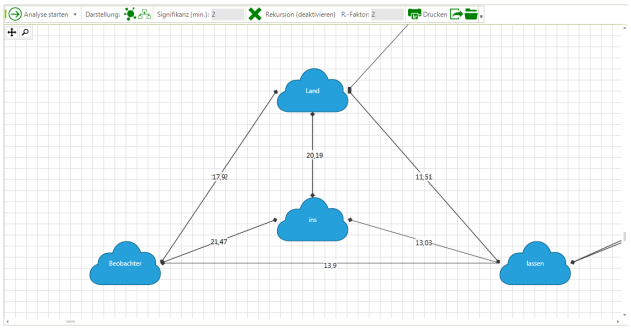
Rot: Kollokatoren tendenziell Isreal



N-Gramm-Graph

Verknüpfung von N-Grammen auf Basis von POS-Tags

Graph: Grün: N-Gramm-Kopf, Blau: N-Gramm-Zwischenteil, Rot: N-Gramm-Ende



Kookkurrenzgraph (Ausschnitt)

Das Beispiel zeigt einen per Rekursion ermittelten Teilausschnitt, der auf die Phrase: *„Beobachter / ins / Land / lassen“* rekurriert.

Notes

1. z. B. kann die Ausgabe des einen Programms nicht vollumfänglich von einem anderen eingelesen werden.
2. Gemeint sind hier die Paradigmen corpus-driven oder corpus-based.
3. Aktuell verfügbar: TreeTagger, TnT, Stanford-Tagger oder gar *Keine Annotation.*

Bibliographie

Alberrt, Hans (1969): *Traktat über kritische Vernunft*. Tübingen: J.C.B. Mohr (Paul Siebeck).

Bubenhofer, Noah (2009): *Sprachgebrauchsmuster*. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin: de Gruyter.

Dang-Anh, Mark / Rüdiger, Jan Oliver (2015): “From Frequency to Sequence: How Quantitative Methods can Inform Qualitative Analysis of Digital Media Discourse”, in: *10plus1* 1: 57–73.

Gardt, Andreas (2007): “Linguistisches Interpretieren: Konstruktivistische Theorie und realistische Praxis”, in: Hermanns, Fritz / Holly, Werner (eds.): *Linguistische Hermeneutik*. Theorie und Praxis des Verstehens und Interpretierens. Tübingen: Niemeyer 263–280.

Haß, Ulrike (2007): “Korpus-Hermeneutik: zur hermeneutischen Methodik in der lexikalischen Semantik”, in: Hermanns, Fritz / Holly, Werner (eds.): *Linguistische Hermeneutik*. Theorie und Praxis des Verstehens und Interpretierens. Tübingen: Niemyer 241–261.

Popper, Karl R. (2005): *Gesammelte Werke*. 3: Logik der Forschung Tübingen: Mohr Siebeck.

Runkler, Thomas (2010): *Data Mining*. Methoden und Algorithmen intelligenter Datenanalyse. Wiesbaden: Vieweg+Teubner.

Teubert, Wolfgang (2006): “Korpuslinguistik, Hermeneutik und die soziale Konstruktion der Wirklichkeit”, in: *Linguistik Online* 28, 3: 41–60 http://www.linguistik-online.de/28_06/teubert.html [letzter Zugriff 09. Februar 2016].

Zu virtuellen Forschungsumgebungen, einer genuin digitalen Hermeneutik sowie deren Visualisierung

Scheuermann, Leif
Leif.scheuermann@uni-graz.at
Karl-Franzens Universität Graz, Österreich

In seinen Akademie-Vorträgen vom 6. Dez. 1906 und 7. Jan. 1909, welche ihren schriftlichen Niederschlag in dem Aufsatz „Abgrenzung der Geisteswissenschaften“ (Dilthey 1970) fanden,

definiert Wilhelm Dilthey als Forschungsgebiet der Geisteswissenschaften „das Erlebnis, das Verstehen fremder Erlebnisse und Urteile und Begriffe, welche die erlebten und verstandenen Sachverhalte ausdrücken“ (Dilthey 1970: 376). Des Weiteren formuliert er programmatisch: „Alle systematischen Wissenschaften des Geistes beruhen auf der Beziehung, die zwischen dem Erlebten und Verstandenen und den Begriffen, die es ausdrücken besteht“(Dilthey 1970: 377). In Konsequenz sieht er als methodischen Ansatz der Geisteswissenschaften die Hermeneutik, deren zentrale Aufgabe es ist das Erlebte und Verstandene als „in Urteilen und Begriffen adäquat darstellbar aufzufassen“ (Dilthey 1970: 383). Nicht die Nacherzählung oder „realistische“ Rekonstruktion ist also das Ziel, sondern das Verstehen und begrifflich neu Fassen, die „Ausbildung der >analytischen Wissenschaft der einzelnen Zweckzusammenhänge, die als Kultursysteme durch die Geschichte hindurchgehen… <“ (Dilthey 1970: 384).

Nimmt man diese basalen Definitionen zur Grundlage, so stellt sich für die digitalen Geisteswissenschaften die Frage nach einer genuin digitalen Hermeneutik, also dem „in Urteilen und Begriffen adäquat darstellbar aufzufassen“ (s. o.) in den digitalen Medien. Eine solche genuin digitale Hermeneutik muss sich von einer traditionellen computerunterstützen Herangehensweise in der Weise unterscheiden, dass sie zum einen ausschließlich innerhalb der digitalen Medien zu geschehen hat (und damit auch multimedial sein muss), zum anderen aber auch darin selbstreferentiell zu dokumentieren und visualisieren ist. Um dies zu verdeutlichen und näher zu erläutern, ist das traditionelle Vorgehen eines Wissenschaftlers darauf begrenzt, Computeranwendungen dazu zu benutzen, Daten zu erheben, sie abzufragen, zu analysieren und zu visualisieren. Dies geschieht meist unter Nutzung unterschiedlichster, lokaler oder webbasierter Werkzeuge, wobei jedoch jedes für sich steht und die Ergebnisse bestenfalls durch „copy-paste“ von einer auf die andere Anwendung übertragen werden. Am Ende des Prozesses steht ein fachliches Urteil, eine Begriffsfindung, welche in einem meist textlichen Narrativ präsentiert wird – z. B. in Form eines Aufsatzes, welcher durchaus online publiziert sein mag. Der hermeneutische Prozess selbst jedoch findet nicht im digitalen Medium statt und wird auch nicht im Ergebnis dokumentiert. Es handelt sich also in dieser Form der geisteswissenschaftlichen Arbeit nicht um digitale Hermeneutik, sondern lediglich um eine Hermeneutik unter Nutzung digitaler Medien.

Wie jedoch soll nun ein genuin digitale Hermeneutik von statten gehen?

Eine erste Grundannahme ist es, dass alle potentiell zu nutzenden Forschungsdaten und Anwendungen auf einer gemeinsamen Oberfläche, einer digitalen Forschungsumgebung, zusammenzuführen und zu verknüpfen sind. Dies kann natürlich nicht bedeuten, dass sämtliche bereits bestehenden Anwendungen neu und für nur ein einziges „System“ erstellt werden, vielmehr ist es die Aufgabe der Plattform Schnittstellen zwischen den Anwendungen und Daten (wobei diese Unterscheidung in letzter Konsequenz hinfällig ist) bereitzustellen und zu verwalten, um eine freie und dynamische Kombination zu ermöglichen. Die zu integrierenden Elemente werden dabei wie Blackboxes behandelt und bleiben so in ihrer Form bestehen, was für neu zu erstellende Anwendungen ebenfalls zur Folge hat, dass sie systemunabhängig funktionieren.

Durch die Integration und freie Kombination unterschiedlichster Daten und Anwendungen kann im digitalen Medium ein hermeneutischer Prozess stattfinden, der nicht mehr implizit im Ergebnis definiert, sondern selbst Teil des Ergebnisses ist. Hierzu muss jedoch der digitale hermeneutischen Prozess dokumentiert und so nachvollziehbar gemacht werden. Dazu bedarf es einer formalen geordneten Darstellung sowohl der Fragestellungen bzw. der Argumentationen, als auch der genutzten Daten und Anwendungen sowie der Ergebnisse und für diese muss eine adäquate Visualisierungsform existieren.

Um auch dies wieder an einem praktischen Beispiel zu erläutern, kann der Nutzer einer solchen Plattform, der sich mit der Ausbreitung antiker Münzen im Mittelmeerraum beschäftigt, verschiedene in CIDOC CRM ausgezeichnete numismatische Datenbanken über generische Schnittstellen in die Plattform integrieren und den eigenen Fragestellungen entsprechend abfragen. Um diese nun in einer Ausbreitungskarte darzustellen, verbindet er die Auswahl mit einer geographischen Visualisierung (z. B. OpenStreetMap), welche ebenfalls in das System zu integrieren ist. Eine räumliche Eingrenzung auf der Karte kann nun wiederum die Abfrage der Datenbank beeinflussen. Er kann jedoch auch eine einzubettende Netzwerk-Analyse z. B. im Hinblick auf die Münzmeister hinzuziehen und diese wiederum für weitere Analysen nutzen. Von zentraler Bedeutung ist es nun, diesen frei definierten Workflow zu protokollieren und darzustellen, so dass ein weiterer Nutzer diesen nach nicht nur vollziehen sondern sich einer anderen Fragestellung auch Teile der Argumentation zu Eigen machen und in den eigenen Workflow integrieren – im Fallbeispiel z. B. zu Mittelalterlichen Münzen.

Ansätze und erste Schritte zu einer Umsetzung eines solchen Systems möchte dieser Vortrag m Fallbeispiel einer digitalen Forschungsumgebung zur Raumwahrnehmung der Stadt Rom in der späten Republik präsentieren. Dabei soll der Fokus auf den theoretischen Grundlagen wie auf der technologischen Umsetzung liegen.

Bibliographie

Dilthey, Wilhelm (1970): *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften*. Frankfurt a.M.: Suhrkamp 365-393.

Scheuermann, Leif (in Vorbereitung): "Die Abgrenzung der digitalen Geisteswissenschaften", in: *Digital Classics* 2.

Scheuermann, Leif (2014): "On co-productive web-based digital mapmaking. Preconditions, risks and opportunities", in: Rau, Susanne / Schönherr, Ekkehard (eds.): *Mapping Spatial Relations, their Perceptions and Dynamics* (= Lecture Notes in Geoinformation and Cartography). Switzerland: Springer International Publishing 17-23.

Scheuermann, Leif (2006): "Ontologien in den historischen Wissenschaften", in: *Historical Social Research* 31, 3: 308-316.

Darstellung heterogenen und dynamischen Wissens mit CIDOC CRM und WissKI

Scholz, Martin
martin.scholz@fau.de
Friedrich-Alexander Universität Erlangen-Nürnberg, Deutschland

Goerz, Guenther
guenther.goerz@fau.de
Friedrich-Alexander Universität Erlangen-Nürnberg, Deutschland

Wagner, Sarah
s.wagner@gnm.de
Germanisches Nationalmuseum Nürnberg

Fichtner, Mark
m.fichtner@wiss-ki.eu
Germanisches Nationalmuseum Nürnberg

Dieser Vortrag stellt die praktische Arbeit mit WissKI und die ontologische Modellierung mit dem CIDOC-CRM anhand zweier Use Cases vor. Dabei steht der dynamische Umgang mit heterogenen Daten im Fokus.

Die WissKI-Software

Die virtuelle Forschungsumgebung “WissKI” (Wissenschaftliche Kommunikations-Infrastruktur, http://wiss-ki.eu/) entstand aus Anforderungen an die kooperative Forschung im Bereich des Kulturerbes und seiner Dokumentation im digitalen Medium. Im Rahmen des DFG-geförderten, gleichnamigen Projekts WissKI wurde die Softwareplattform auf der Basis des Open-Source Content Management Systems Drupal (http://drupal.org) in Zusammenarbeit zwischen dem Germanischen Nationalmuseum Nürnberg (GNM), dem Zoologischen Forschungsmuseum Alexander Koenig in Bonn (ZFMK) und der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) entwickelt. Fokus des Systems sind neben der einfachen Bereitstellung und offenen Verfügbarkeit von Quellmaterialien – strukturierten Texten, Grafiken, Bildern, Video, Audio – und Metadaten in digitaler Form, auch das interaktive und vernetzte Arbeiten auf der Basis semantischer Tiefenerschließung. Indem die typischen Eigenschaften von gängigen Content Management Systemen unberührt bleiben, verfügt das System über eine detaillierte Nutzersteuerung mit Rechteverwaltung und ist in der Lage Web-Inhalte wie Webseiten, Foren und Wikis zu verwalten und online zu präsentieren.

Die Software ist als Open-Source kostenfrei unter http://github.com/WissKI (Görz et al. 2009-*) veröffentlicht und kann dementsprechend nachgenutzt und erweitert werden.

Semantische Modellierung mit dem CIDOC-CRM

Bei der semantischen Modellierung kommt dem Conceptual Reference Model (CIDOC-CRM, ISO 21127) von ICOM-CIDOC als formaler Referenzontologie eine Schlüsselrolle zu. Unsere Implementation in der Web Ontology Language bildet in Verbindung mit verschiedenen Werkzeugen des Semantic Web die Grundlage des WissKI-Systems. Auch diese Softwarekomponente ist als Open-Source kostenfrei verfügbar und kann unter http://erlangen-crm.org heruntergeladen werden.

Die Datenakquisition im WissKI-System erfolgt primär über karteikartenähnliche Formulare oder Freitextfelder, beides tradierte Formen der Datenerfassung in den vorliegenden Anwendungsbereichen. Die Daten werden vom System im Hintergrund jedoch durch das CIDOC-CRM semantisch erschlossen, d. h. neben den Daten wird auch ihre ontologiebezogene Bedeutung für Mensch und Maschine lesbar gespeichert. Hierzu bedient sich WissKI konsequent der Techniken des Semantic Web wie RDF und OWL-DL und zielt auf Veröffentlichung der Daten als Linked Open Data. Die eingegebenen Daten bilden einen Wissensgraphen, der auf einfache Weise weltweit und (potentiell) transdisziplinär vernetzt werden kann. Die formale Ontologie, die über eine hierarchische Struktur von Konzepten und Eigenschaften und einer standardisierten logischen Sprache ein System von Fachbegriffen bildet, dient der Erfassung und semantischen Erschließung der Daten. Je nach Fachgebiet und Sammlungsschwerpunkt kann das CIDOC-CRM als grundlegende Ontologie um fachspezifische Begriffe mit Anwendungsontologien erweitert werden. Diese

flexible Art der Wissensspeicherung lässt schnell und unproblematisch Änderungen des Datenmodells am laufenden System zu. Durch die Einbindung einer standardisierten logischen Sprache und die Nutzung von Open-Source-Software ermöglicht WissKI eine hohe Anschlussfähigkeit, leichte Zugänglichkeit und Langzeit-Interpretierbarkeit der Daten.

Exemplarische Anwendungsszenarien

EDEN

Die Epigraphische Datenbank Erlangen-Nürnberg (EDEN, Dreyer et al. 2014-*) ist eine Online-Datenbank für antike Inschriften aus Kleinasien. Momentan beschränkt sie sich auf die drei antiken griechischen Siedlungen Metropolis in Ionien, Magnesia am Mäander und Apollonia am Rhyndakos. Inhaltlich wird die Datenbank seit 2012 von Boris Dreyer (Professur für Alte Geschichte, FAU) in enger Zusammenarbeit mit Archäologen der FAU und dem Lehrstuhl für Graphische Datenverarbeitung sowie archäologischen Kollegen in der Türkei gepflegt und weiterentwickelt. Ziel ist die Schaffung eines effektiven Werkzeugs zur interdisziplinären und internationalen Forschung, das Forscher, Studenten und die breite Öffentlichkeit im Spannungsbereich zwischen Alter Geschichte und Archäologie nutzen können. Während sich Althistoriker primär mit den immateriellen Eigenschaften der Inschrift, also Textinhalt und -form, beschäftigen, ist für Archäologen der materielle Träger der Inschrift von größerem Interesse. Die Verknüpfung und gemeinsame Präsentation von Daten für beide Disziplinen bildet ein Novum in diesem Bereich. Daher waren von Anfang an web-basiertes, kollaboratives Arbeiten sowie flexible Datenhaltung wichtige Voraussetzungen der technischen Infrastruktur.

Der Fokus der Datenbank ist (noch) nicht das Bereitstellen einer großen Anzahl an Inschriften, sondern die Angabe detaillierter Metadaten aus verschiedenen Disziplinen: zu den edierten Inschriften kommen zahlreiche Metadaten in textueller und tabellarischer Form, u. a. Funddaten, wissenschaftliche Kommentare, Übersetzungen und hochauflösende Bilder. Diese geben wertvolles Wissen für Historiker und Archäologen wieder. Entsprechend den Entwicklungen durch Kooperationen und neue Forschungsschwerpunkte rücken frühere Randinformationen schnell in den Fokus der Datenbank und werden sukzessive verfeinert und ergänzt. Ebenso werden weiterführende Informationen zu wiederkehrenden Themen wie bestimmten Herrschern, Genres oder Orten laufend hinzugefügt. Die Einbindung von 3D-Modellen soll die Datenbank mittelfristig zu virtuellen Ausgrabungsstätten erweitern.

Sammlung Rück

Seit 2015 widmen sich Forscher am GNM einem einzigartigen Bestand an Musikinstrumenten. Die Sammlung Rück war die größte deutsche Sammlung historischer Musikinstrumente in Privatbesitz und wurde 1880 vom Pianisten und Lehrer Wilhelm Rück gegründet. Nach seinem Tod setzten sich seine Söhne den systematischen Ausbau der Sammlung und die Dokumentation der Entwicklung abendländischer Musikinstrumente zum Ziel. Eine seit 1929 akribisch geführte Dokumentation aller Sammlungsaktivitäten ist Ulrich Rück, dem Sohn des Sammlungsgründers, zu verdanken. Diese Korrespondenz mit über 1000 Partnern umfasst 17.200 Briefe und Postkarten, die zusammen mit 1.500 Musikinstrumenten und weiterem Material 1962 vom GNM erworben wurden. Alle Dokumente werden seither im Historischen Archiv, die Objekte selbst im Sammlungsreferat für Musikinstrumente aufbewahrt.

Die Korrespondenz zum Aufbau der Sammlung Rück bietet heute eine einzigartige Quelle, Einblicke in das Handeln mit Musikinstrumenten und in die Entstehung einer Sammlung von Objekten des Kulturlebens in der Zeit zwischen Weltwirtschaftskrise und Wirtschaftswunder gewinnen zu können. Besondere Bedeutung kommt dem Schriftverkehr mit Gutachtern bezüglich Qualität und Marktwert der Instrumente zu, da hieraus ermöglicht wird, einen Preisspiegel für den Handel mit historischen Musikinstrumenten der damaligen Zeit zu erstellen. Dieser soll zum Abschluss des Projekts gemeinsam mit Recherche-Ergebnissen zu den Erwerbsumständen und unmittelbaren Vorbesitzern online zugänglich gemacht werden. Weitere wichtige Aspekte der Untersuchung sind Sammlungs-, Kommunikations- und Marketingstrategien, die zu einer europaweiten Vernetzung der Sammlung führten, und die in einer Monographie dargestellt werden.

Zentrale Herausforderung des Nachlasses Rück ist die Heterogentität der Materialien. So sind neben den Archivgütern, die inhaltlich tiefenerschlossen werden sollen, auch Objekt-, Personen- und Literaturdaten mit entsprechenden Kreuzverweisen zu erfassen. Auf der Basis dieser Verweise soll die Abfolge der Kommunikation mit den verschiedenen Kommunikationspartnern, die geschäftlichen Reisen der Familie Rück, Erwerbungen in ihren zeitlichen und räumlichen Beziehungen und der generelle Kontext zur damaligen Zeit dargestellt werden.

Der Großteil der Archivmaterialien aus dem Nachlass Rück ist mit Schreibmaschine geschrieben und deshalb gut lesbar. Eine händische Transkription der Materialien ist nicht notwendig. Die Erschließung erfolgt auf Basis von hochqualitativen Digitalisaten der Briefe, die anschließend die Wissenschaftler inhaltlich zusammenfassen. In diesen Zusammenfassungen werden entsprechende Referenzierungen zu den anderen Materialien des Projektes vorgenommen. Hierfür ist eine

Verlinkung und idealerweise eine Auszeichnung aus dem Fließtext heraus notwendig. Gleichzeitig wird ein System benötigt, das die verknüpften Materialien geeignet präsentieren kann.

Praktische Umsetzung

EDEN

Durch die Einbindung von Drupal bringt WissKI einerseits die nötigen Voraussetzungen für web-basiertes, kollaboratives Arbeiten mit. Der generische Aufbau ermöglicht andererseits die Anpassung an die Erfordernisse der beteiligten Disziplinen. Weiterhin erleichtert der Einsatz von Ontologien zur Datenspeicherung einen oben angesprochenen Fokuswechsel, da Daten nicht neu erfasst werden müssen, sondern lediglich die bereits bestehenden Randinformationen angereichert werden und somit der Detailgrad auf neue Anforderungen angehoben werden kann. In EDEN wurden beispielsweise die rudimentären, zunächst aufgrund von Nennungen im Text erfassten Personendaten zu Datensätzen ersten Ranges mit eigenem wissenschaftlichen Kommentar ausgebaut.Ebenso könnten die geographischen Informationen durch eine Kooperation mit Geographen ausgeweitet werden und EDEN zu einer für (kultur-)geographische Forschungen nutzbaren Quelle machen. Diese Änderungen des Datenmodells konnten mit WissKI problemlos parallel zum Einpflegen der Daten erfolgen. Somit kann die Datenbank nicht nur hinsichtlich der Datenmenge, sondern auch hinsichtlich der Fülle der Metadaten bei Bedarf stetig wachsen.

Sammlung Rück

Das im Archivbereich des GNM bisher benutzte Dokumentationssystem “Faust” (http://www.land-software.de/) ist für die Erfassung, Erschließung und Abbildung der Projektdaten in vielerlei Hinsicht unzureichend. Zum einen würde eine Anpassung des Archivsystems auf die Bedürfnisse des Projekts Rück dazu führen, dass in allen anderen Anwendungsbereichen entsprechende Felder auftauchen. Eine inhaltliche Tiefenerschließung in der Qualität des Projekts Rück ist für die sonstigen Archivmaterialien eher untypisch, da für die Erfassung letzterer insbesondere die Quantität im Vordergrund steht.

Faust hat bislang keine Komponente zur Text-Annotierung in Fließtexten, was für die semantische Tiefenerschließung der Archivmaterialien essenziell ist. Auch ist Faust nur bedingt in der Lage, die Verknüpfungsdichte an Daten darzustellen. Unter diesem Aspekt kann eines der Kernziele, das Netzwerk agierender Personen im zeitlichen und räumlichen Kontext abzubilden, kaum realisiert werden.

WissKI erfüllt die Bedürfnisse des Projekts, da es über einen Text-Annotierer für Fließtext verfügt und auf der Basis von Semantic Web-Techniken Verknüpfungen flexibel darstellt und auswertet. Auch konnte mit WissKI eine direkte Schnittstelle am Objektdatensatz zu MIMO (2009-*) (musical instrument museums) eingerichtet werden.

In WissKI können die Forschungsprimärdaten inklusive der Digitalisate in voller TIFF-Qualität und allen Annotierungen der Fachöffentlichkeit zur Verfügung gestellt und eine Präsentationsoberfläche für die breite Öffentlichkeit geschaffen werden.

Fazit und Ausblick

Obgleich in ihren Disziplinen und Objektgattungen verschieden, haben beide Anwendungsfälle gemeinsame Herausforderungen: Zum einen sind sie mit heterogenen, stark vernetzten Daten und anwendungsspezifischen Anforderungen konfrontiert. Zum anderen entwickeln sich die Anforderungen an die Software hinsichtlich Umfang und Vernetzung der Metadaten im Projektverlauf. WissKI ist jedoch aufgrund seiner Systemarchitektur darauf bestens vorbereitet.

Im derzeit noch laufenden Projekt WissKI² wird die Entwicklung der WissKI-Software konsequent weitergeführt. Neben den bereits bewährten Formularfeldern und Bildern sollen nun verstärkt auch interaktive Landkarten, Zeitstrahlen, 3D-Animationen und weitere Medientypen eingebunden und dargestellt werden. Damit ergeben sich weitere Möglichkeiten der Datenpräsentation und der Wissensvernetzung für die vorgestellten Anwendungsfälle.

Bibliographie

Dreyer, Boris / Holdenried, Marvin / Scholz, Martin (2014-*): *EDEN — Epigraphische Datenbank Erlangen-Nürnberg*. WissKi: Friedrich-Alexander-University, Erlangen-Nuremberg (FAU) http://wisski.cs.fau.de/eden/ [letzter Zugriff 08. Januar 2016].

Görz, Günther (2011): "WissKI: Semantische Annotation, Wissensverarbeitung und Wissenschaftskommunikation in einer virtuellen Forschungsumgebung", in: *Kunstgeschichte. Open Peer Reviewed Journal* http://www.kunstgeschichte-ejournal.net/167/ [letzter Zugriff 08. Januar 2016].

Görz, Günther / Scholz, Martin (2012): "WissKI: A Virtual Research Environment for Cultural Heritage", in: De Raedt, Luc, Luc De Raedt, Bessiere, Christian / Dubois, Didier / Doherty, Patrick / Frasconi, Paolo / Heintz, Fredrik / Lucas, Peter (eds.): *20th European Conference on Artificial Intelligence, ECAI 2012,*

Proceedings. IOS Press http://www2.lirmm.fr/ecai2012/ [letzter Zugriff 08. Januar 2016].

Görz, Günther / Scholz, Martin / Merz, Dorian / Krause, Siegfried / Fichtner, Mark / Reinfandt, Kerstin / Grobe, Peter / Pfeifer, Maria Anna (2009-*): *WissKi: Scientific Communication Infrastructure*. Friedrich-Alexander-University, Erlangen-Nuremberg (FAU); Digital Humanities Research Group, Department of Computer Science, Germanisches Nationalmuseum (GNM) Nuremberg; Biodiversity Informatics Group, Department of Museum Informatics, Zoologisches Forschungsmuseum Alexander Koenig (ZFMK) Bonn http://wiss-ki.eu/ [letzter Zugriff: 08. Januar 2016].

Hohmann, Georg / Fichtner, Mark (2015): "Chancen und Herausforderungen in der praktischen Anwendung von Ontologien für das Kulturerbe", in: Robertson-von Trotha, Caroline Y. / Schneider, Ralf M. (eds.): *Digitales Kulturerbe*. Bewahrung und Zugänglichkeit in der wissenschaftlichen Praxis (= Kulturelle Überlieferung – digital 2). Karlsruhe: Karlsruhe Scientific Publishing 115-128.

MIMO (2009-*): *Musical Instrument Museums Online*. Philharmonie de Paris: Cité de la musique, The University of Edinburgh, Germanisches Nationalmuseum, Muziekinstrumentenmuseum, Association "Amici del Museo degli Strumenti Musicali" http://www.mimo-international.com/MIMO/accueil-ermes.aspx [letzter Zugriff 08. Januar 2016].

Scholz, Martin / Holdenried, Marvin / Dreyer, Boris / Meyer-Wegener, Klaus / Görz, Günther (2014): "Und Semantik wuchs in EDEN - Eine Vorstellung und ein Erfahrungsbericht", in: *Magazin für digitale Editionswissenschaften* 1, 1: 22-30 https://www.mde.fau.de/files/2015/03/MdE-2015-01_3_Scholz_et_al.pdf [letzter Zugriff 08. Januar 2016].

Geisteswissenschaftliche Fachdatenrepositorien im Semantic Web. Modellierung, Vernetzung, Visualisierung.

Schrade, Torsten
Torsten.Schrade@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland

Implizite und explizite Semantik TEI-basierter Fachdatenrepositorien

Zahlreiche geisteswissenschaftliche Fachdatenrepositorien setzen zur Modellierung ihrer Forschungsdaten auf die Richtlinien der Text Encoding Initiative (TEI) und somit auf XML als primäres Datenformat. XML eignet sich sehr gut zur Lösung editorisch-philologischer Aufgabenstellungen und entspricht den geforderten Kriterien der Interoperabilität und Nachhaltigkeit von Forschungsdaten. Durch die standardkonforme Auszeichnung der Forschungsgegenstände in TEI werden diese formal und inhaltlich erschlossen. TEI-kodierte Daten beinhalten in jeder Hinsicht semantische Bezüge (bspw. Raumbezüge, Personenbezüge, begriffliche und konzeptuelle Bezüge etc.). Aus der Perspektive des *Semantic Web* sind diese Bezüge jedoch zunächst nur implizit und nicht explizit in den Daten vorhanden (Abbildung 1). Im Gegenzug gründen sich *Semantic Web*-Technologien auf das *Resource Description Framework* (RDF) zur Formulierung semantischer Aussagen (*statements*) in Form von Subjekt – Prädikat – Objektbeziehungen (*triples*). Die besondere Stärke von RDF liegt in der automatisiert möglichen Vernetzung (*interlinking*), Zusammenführung (*merging*) und Analyse (*reasoning*) eigentlich separater Datenbestände. RDF ist modellierungstechnisch auf einer höheren Abstraktionsebene anzusiedeln als TEI-kodierte XML-Daten (Abbildung 2; vgl. auch Polleres u. a. 2009).

Abb. 1

Abb. 2

Während die formale Erschließung geisteswissenschaftlicher Forschungsgegenstände mittels XML-basierter Annotationsmethoden mittlerweile als weit fortgeschritten gelten kann, bleibt die semantische Erschließung häufig noch weit zurück. Zwar wird in den Daten oft das Auftreten bestimmter Ortsnamen, Personennamen, Werktitel etc. annotiert. Dennoch gehen diese Annotationen meist nicht darüber hinaus, anzuzeigen, dass eine bestimmte Entität an einer spezifischen Stelle erwähnt ist. Damit bleibt die Semantik der Fachdaten weit hinter den Möglichkeiten zurück, die aktuelle Technologien – insbesondere die des *Semantic Web* und des *Linked Open Data* (LOD) – bieten könnten. Der durch LOD mögliche Zugang auf vernetzte Forschungsdaten eröffnet neuartige Perspektiven der Nutzung bisher isoliert stehender Fachdatenrepositorien. Wesentlich ist dabei, dass LOD und RDF bestehende Standards der *Digital Humanities* (wie bspw. TEI / XML) um die Verwendung gemeinsamer Terminologien und Metadatenschemata erweitern (vgl. Iglesia et al. 2015). Dadurch wird es möglich, auch Bestände verteilter Provenienz und unterschiedlicher Struktur gemeinsam inhaltlich zu beschreiben und zu analysieren.

Insgesamt existiert momentan also noch eine Kluft: Auf der einen Seite die zahlreichen geisteswissenschaftlichen Fachdatenrepositorien mit implizit semantischem Potential, auf der anderen Seite die Technologien und Datenmodelle des *Semantic Web*, die neue Sichten und Analysemethoden auf die Daten eröffnen könnten. Zwar existieren einige Sprachkonzepte, Methoden und Tools zur Übersetzung zwischen TEI / XML und RDF. Diese sind jedoch ausnahmslos komplex, teilweise technisch veraltet, verfügen nur über prototypische Implementierungen oder sind hochgradig spezialisiert auf einen bestimmten Datenbestand. [1] Während die dem *Semantic Web* zugrunde liegenden Technologien aus informatischer Sicht als erschlossen und anwendbar angesehen werden können (vgl. Lanthaler 2014: 11–35), besteht zum jetzigen Zeitpunkt also ein Bedarf an exemplarischen Bearbeitungen repräsentativer Forschungsdatenbestände aus den Geisteswissenschaften, um die Tragfähigkeit dieser Technologien auch für die geisteswissenschaftliche Forschung zu demonstrieren.

Semantische Aussagen aus XML mit Hilfe des *XTriples*-Webservices

An dieser Stelle setzt der *XTriples* -Webservice der *Digitalen Akademie* der Mainzer Akademie der Wissenschaften und der Literatur an. Grundgedanke des generischen Dienstes ist das Crawling beliebiger XML-Datenbestände und die anschließende Generierung semantischer Aussagen aus den XML-Daten auf Basis definierter Aussagemuster. Das Prinzip der Explizierung semantischer Aussagen aus XML ist dabei nicht sonderlich komplex: Wird die URI einer XML-Ressource oder eine Dateneinheit in dieser Ressource als das Subjekt einer semantischen Aussage begriffen, können diesem Subjekt über Prädikate aus kontrollierten Vokabularen weitere Werte aus den XML-Daten bzw. URIs zu weiteren Datenressourcen als Objekte zugeordnet werden. Im Übersetzungsvorgang zwischen XML und RDF geht es also vor allem um die Bestimmung semantischer Aussagemuster, die sich gesamthaft auf alle Ressourcen eines XML-Datenbestandes anwenden lassen.

Die Aussagemuster werden in Form einer einfachen, XPATH-basierten Konfiguration an den Dienst übermittelt. Dabei ist es auch möglich, über die Bestände eines spezifischen XML-Repositoriums hinauszugehen und externe Ressourcen oder Dateneinheiten in die Transformation mit einzubeziehen (bspw. aus der GND, der *Dbpedia,* aus *Geonames* u. a *)*. Die technische Realisierung als Webservice hat den Vorteil, dass AnwenderInnen keine weitere Software zur semantischen Übersetzung von Forschungsdaten benötigen. Gleichzeitig kann der Webservice auch als eine Art „externe“ RDF-Schnittstelle (im Sinne eines Proxy) für ein oder mehrere XML-Repositorien eingesetzt werden. Grundvoraussetzung hierfür ist lediglich, dass die jeweiligen Repositorien über HTTP erreichbar sein müssen.

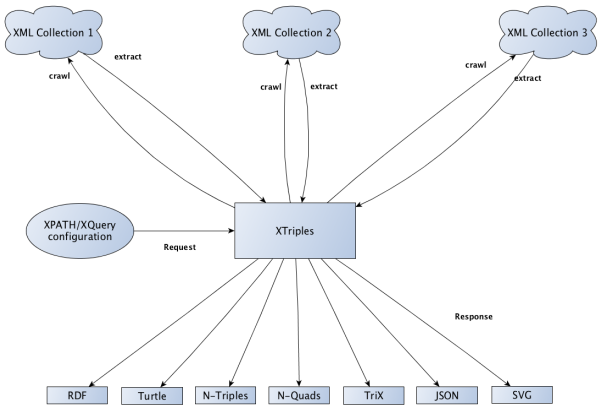


Abb. 3

Das Ergebnis einer *XTriples*-Extraktion steht in einer Vielzahl gängiger RDF-Serialisierungen zur Verfügung (Abbildung 3). Neben rein RDF-basierten Formaten ist es auch möglich, die semantischen Bezüge eines Repositoriums mittels SVG darzustellen oder das Extraktionsergebnis zur weiteren Analyse und Visualisierung an *Semantic Web*-Tools weiterzureichen. [2]

Anwendungsbeispiele

XTriples wurde vom Autor im Kontext des Akademievorhabens *Deutsche Inschriften Online* in Verbindung mit dem BMBF-Projekt *Inschriften im Bezugssystem des Raumes* entwickelt und steht der DH-Community in einer stabilen Version unter Open Source

Lizenz (MIT) zur Verfügung. Das zugrunde liegende Softwarepaket ist vollständig dokumentiert und auf GitHub veröffentlicht.

Neben den *Deutschen Inschriften* wird *XTriples* aktuell auch in den Akademievorhaben *Regesta Imperii* und *Die Schule von Salamanca* verwendet. Das Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte nutzt *XTriples* für eine CIDOC-CRM basierte, semantischen Modellierung von EpiDoc-Daten im Rahmen des BMBF-Projektes *Relationen im Raum*. Gemeinsam mit der Berlin-Brandenburgischen Akademie der Wissenschaften wird gerade eine Schnittstelle zwischen *XTriples* und *correspSearch* , dem Webservice der BBAW zur dezentralen Aggregation digitaler Briefeditionen, implementiert.

Folgende Beispiele geben einen ersten Überblick über die unterschiedlichen Anwendungsgebiete von *XTriples*:

- Semantische Extraktion und nachfolgende Visualisierung von Familienbeziehungen aus dem *Epidat* Grabstein Corpus für den jüdischen Friedhof in Hamburg-Altona (Ausgangsdaten EpiDoc/TEI): http://xtriples.spatialhumanities.de/examples/dh/epidat/index.html
- Semantische Extraktion und SVG-Visualisierung eines Briefnetzwerks (Teilbestand der Korrespondenz Goethes aus den in *correspSearch* aggregierten CMI-Daten bei gleichzeitiger *on-the-fly* Einbeziehung der RDF-Schnittstellen von *GND* und *Geonames*): http://bit.ly/1LKK1dv
- Beispielhafte semantische Extraktion und Visualisierung europäischer Kommunikationsnetzwerke auf Basis der *correspSearch* TEI / CMI-Daten: http://metacontext.github.io/presentation-correspsearch-xtriples/viz/map.html

Weitere Beispiele zu den einzelnen Funktionalitäten finden sich auf der *XTriples*-Website unter http://xtriples.spatialhumanities.de/examples.html . Einen schnellen Überblick über die Funktionsweise des Dienstes gibt folgende Präsentation: http://metacontext.github.io/presentation-correspsearch-xtriples .

Ziel des Vortrags ist eine Veranschaulichung der Methoden und Potentiale, die sich aus der semantischen Extraktion, Modellierung, Vernetzung und Visualisierung XML-basierter, geisteswissenschaftlicher Fachdaten ergeben. Neben einer Darstellung der technischen Hintergründe des *XTriples*-Webservices werden auch Fragen der semantischen Modellierung geisteswissenschaftlicher Fachdaten mittels bestimmter Ontologien (bspw. FOAF, CIDOC-CRM u.a.) in den Blick genommen. Weiterhin werden auch beispielhafte Analyse- und Visualisierungsmöglichkeiten für semantisch modellierte geisteswissenschaftliche Fachdaten vorgestellt.

Notes

1. Projekte wie bspw. SPQR oder das Textual Encoding Framework sind veraltet oder technisch nicht generalisiert. Einen interessanten Ansatz bietet die XSPARQL Language Specification des DERI, die 2009 in Form einer W3C Member Submission niedergelegt wurde. Hier fehlen jedoch praktische Implementierungen. Die Benutzung von RDFa innerhalb von XML-Daten stellt eine weitere Möglichkeit dar, doch verfolgen die wenigsten geisteswissenschaftlichen Fachdatenrepositorien eine so ausgerichtete semantische Markup-Strategie. Auch das bereits 2007 in Form einer W3C Recommendation grundgelegte GRDDL-Framework (Gleaning Resource Descriptions from Dialects of Languages) ist bis heute eine theoretische Spezifikation geblieben. Der *OxGarage* Transformations-Webservice der *Text Encoding Initiative* bietet zwar eine Routine für die Konvertierung von TEI kodierten Daten nach RDF an, legt sich für die Transformation aber auf das CIDOC-CRM als Ontologie fest. Mit *OxGarage* können *out-of-the-box* also keine anderen Ontologien für eine semantische Modellierung benutzt werden. Zudem ist der Webservice nicht darauf ausgelegt, auch weitere, externe Datenrepositorien in eine Transformation mit einzubeziehen oder andere RDF-Serialisierungen jenseits von RDF/XML zurückzugeben.
2. Beispielsweise an den RDF zu SVG Transformations-Webservice oder an die RDF Visualisierungsbibliothek d3sparql .

Bibliographie

Akademie der Wissenschaften und der Literatur Mainz (o. J.a): *Digitale Akademie* http://www.digitale-akademie.de [letzter Zugriff 16. Februar 2016].

Akademie der Wissenschaften und der Literatur Mainz (o. J.b): *Regesta Imperii* http://www.regesta-imperii.de/startseite.html [letzter Zugriff 16. Februar 2016].

BBAW (o. J.): *Berlin-Brandenburgische Akademie der Wissenschaften* http://www.bbaw.de/ [letzter Zugriff 16. Februar 2016].

correspSearch (o. J.): *correspSearch*. Search diverse letter editions http://correspsearch.bbaw.de/index.xql [letzter Zugriff 16. Februar 2016].

Deutsche Inschriften Online (o. J.): http://www.inschriften.net [letzter Zugriff 16. Februar 2016].

Haft, Michael (2013): "RDF als Verknüpfungsmethode zwischen geisteswissenschaftlichen Forschungsdaten und Geometrien am Beispiel des Projektes 'Inschriften im Bezugssystem des Raumes'",in: *Skriptum* 2,3 http://nbn-resolving.de/urn:nbn:de:0289-2013120622 [letzter Zugriff 14. Oktober 2015].

i3Mainz / Akademie Mainz (o. J.): *Inschriften im Bezugssystem des Raumes* http://www.spatialhumanities.de/ibr/startseite.html [letzter Zugriff 16. Februar 2016].

IBR (Inscriptions in their spatial context) / Academy of Sciences and Literature, Mainz / Institute for Spatial Information and Surveying Technology i3Mainz (o. J.): *XTriples* http://xtriples.spatialhumanities.de/index.html [letzter Zugriff 16. Februar 2016].

Iglesia, Martin de la / Moretto, Nicolas / Brodhun, Maximilian (2015): "Metadaten, LOD und der Mehrwert standardisierter und vernetzter Daten." In: Neuroth, Heike / Rapp, Andrea / Söring, Sibylle (eds.): *TextGrid: Von der Community – für die Community*. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften. Göttingen: Universitätsverlag Göttingen 91–102 http://dx.doi.org/10.3249/webdoc-3947 [letzter Zugriff 14. Oktober 2015].

Lange, Felix, Martin Unold (2015): Semantisch angereicherte 3D-Messdaten von Kirchenräumen als Quellen für die geschichtswissenschaftliche Forschung, in: *Zeitschrift für digitale Geisteswissenschaften* 1 http://dx.doi.org/10.17175/sb001_015 [letzter Zugriff 14. Oktober 2015].

Lanthaler, Markus (2014): *Third Generation Web APIs*. Bridging the Gap between REST and Linked Data. Diss. Institute of Information Systems and Computer Media. Technische Universität Graz http://www.markus-lanthaler.com/research/third-generation-web-apis-bridging-the-gap-between-rest-and-linked-data.pdf [letzter Zugriff 14. Oktober 2015].

Polleres, Axel u.a. (2009): *XSPARQL Language Specification* http://www.w3.org/Submission/xsparql-language-specification letzter Zugriff 14. Oktober 2015].

Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte (o. J.): http://www.steinheim-institut.de [letzter Zugriff 16. Februar 2016].

Schrade, Torsten (2013): "Datenstrukturierung", in: *Über die Praxis des kulturwissenschaftlichen Arbeitens*. Ein Handwörterbuch. Bielefeld: transcript 91–97.

Topic, Genre, Text
Topics im Textverlauf von Untergattungen des spanischen und hispanoamerikanischen Romans (1880-1930)

Schöch, Christof
christof.schoech@uni-wuerzburg.de
Universität Würzburg, Deutschland

Henny, Ulrike
ulrike.henny@uni-wuerzburg.de
Universität Würzburg, Deutschland

Calvo, José
jose.calvo@uni-wuerzburg.de
Universität Würzburg, Deutschland

Schlör, Daniel
daniel.schloer@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Popp, Stefanie
stefanie.popp@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Der Beitrag möchte zeigen, wie die Berücksichtigung detaillierter, gattungsbezogener Metadaten auf produktive Weise mit dem Verfahren des Topic Modeling verbunden werden kann, um bisher nicht bekannte thematische Strukturen im Textverlauf in einer Sammlung spanischer und hispanoamerikanischer Romane zu entdecken. Ausgangshypothese ist, dass die Wichtigkeit bestimmter Topics nicht nur im Textverlauf variiert, sondern dies auch in verschiedenen Untergattungen auf unterschiedliche Weise tut. Eine Pilotstudie wurde im März 2015 beim Workshop zu Computational Narratology bei der DHd-Tagung in Graz vorgestellt. Im Rahmen der interdisziplinären Würzburger eHumanities-Nachwuchsgruppe "Computergestützte literarische Gattungsstilistik (CLiGS) wurde dieser Fragestellung nun mit weiter entwickelten Methodik sowie einer neu erstellten Sammlung spanischsprachiger Romane aus Spanien und Hispanoamerika nachgegangen.

Stand der Forschung und Fragestellung

Die Frage nach dem Text- oder Handlungsverlauf in narrativen literarischen Texten hat jüngst zunehmende Aufmerksamkeit in der digitalen Literaturwissenschaft erhalten. Matthew Jockers kam durch Sentiment Analysis im Verlauf zahlreicher Romane zu dem (kontrovers diskutierten) Ergebnis, es gäbe sechs oder sieben grundlegende Plotstrukturen (Jockers 2015). Ben Schmidt hat unter anderem den Verlauf von Topic-Wahrscheinlichkeiten in der "screen time" amerikanischer Fernsehserien verfolgt (Schmidt 2014). Der vorliegende Beitrag verbindet die Frage nach dem Textverlauf mit der

nach den Untergattungen, seine zentrale Fragestellung lautet: Können wir nach Untergattung unterschiedliche Verlaufsmuster für bestimmte Topics über den Textverlauf hinweg feststellen?

Daten

Die Textsammlung enthält 150 spanische und hispanoamerikanische Romantexte aus der Zeit von 1880 bis 1930 (für den spanischen Roman: Altisent 2008; de Nora 1963, für den hispanoamerikanischen Roman: Gallo 1981; Williams 2009). Die Texte sind in TEI aufbereitet und mit detaillierten Metadaten versehen worden. Es wurden vier weit gefasste Untergattungen gewählt, um die Romane miteinander vergleichen zu können: *novela sentimental*, *novela histórica*, *novela político-social* und *novela de tendencia subjetiva*. Die Auswahl der Texte ist auch von der Verfügbarkeit als digitaler Volltext beeinflusst und daher nicht unbedingt repräsentativ. Abbildung 1 zeigt die Verteilung der Romane nach ausgewählten Metadaten.

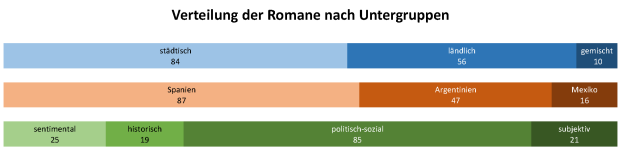


Abb. 1: Verteilung der Romane nach Metadaten

Methode

Topic Modeling ist eine unüberwachte, nicht-deterministische Methode aus dem Bereich des *Natural Language Processing*, die auf Annahmen aus der distributionellen Semantik basiert und verborgene semantische Strukturen in großen Textsammlungen aufdeckt (einführend Blei 2011, grundlegend Blei 2003). Gruppen semantisch verwandter Wörter werden insbesondere aufgrund ihres häufigen gemeinsamen Auftretens in den untersuchten Dokumenten entdeckt. Ein Topic ist eine Wahrscheinlichkeitsverteilung von Wörtern; ein Dokument wird als Wahrscheinlichkeitsverteilung von Topics beschrieben. Topic Modeling ist eine in den DH äußerst beliebte Methode (Anwendungsbeispiele: Blevins 2010; Rhody 2012; Jockers 2013; Schöch 2015).

Hier wurde Topic Modeling als Teil eines umfassenden, weitgehend automatischen Arbeitsablaufes als Serie von Python-Skripten implementiert: Präprozessieren der Texte (Segmentierung, Binning, Lemmatisierung, POS-Tagging), das eigentliche Topic Modeling (mit Mallet, siehe McCallum 2002), Aufbereitung des Mallet-Outputs, zahlreiche Visualisierungen als Perspektiven auf die Ergebnisse. Die wichtigsten Parameter: Berücksichtigung ausschließlich der Substantive, Weglassung der 70 häufigsten Substantive, Romansegmente von ca. 600 Wörtern (unter Berücksichtigung von Absatzgrenzen), Anzahl von 70 Topics. Die Python-Skripte sind frei verfügbar und ausführlich dokumentiert, Begleitmaterialien (Skripte, Parameterdatei, Metadaten, Abbildungen) sind unter https://github.com/cligs/projects/tree/master/2016/dhd einsehbar.

Ergebnisse und Diskussion

Es werden zunächst die Topics selbst dargestellt, dann Unterschiede in den Topic-Verteilungen nach Untergattungen, über den Textverlauf hinweg und schließlich über den Textverlauf in Abhängigkeit der Untergattung.

Topics

Die Mehrheit der erhobenen Topics beinhaltet konkrete typische Themen und Motive des spanischsprachigen Romans der Epoche. Man erkennt eine klare semantische Beziehung der Wörter: ein konkreter Bereich menschlicher Tätigkeiten, wie in Topic 19 (maestro-colegi o-escuela, dt. "Lehrer-Schule-Schule") oder Topic 23 (sangre-golpe-arma, dt. "Blut-Schlag-Waffe"); oder abstrakte Begriffe und Gefühle, wie bei Topic 69 (conciencia-honor-crimen, dt. "Gewissen-Ehre-Verbrechen"). Weniger kohärent ist Topic 45 (marido-rato-chico, dt. "Ehegatte-Weile-Junge"). Die folgenden Wordclouds (Abbildung 2) veranschaulichen die erwähnten Topics.

Abb. 2: Wordclouds für ausgewählte Topics.

Untergattungen und Topics

Die folgende Heatmap (Abbildung 3) zeigt die Verteilung der durchschnittlichen Topic-Wahrscheinlichkeiten in den vier Untergattungen für diejenigen 20 Topics, deren Werte zwischen den Untergattungen besonders stark schwanken (nach

Standardabweichung). Besonders distinktive Topics existieren für die *novela de tendencia subjetiva* (Topic 11: mirada-huerto-silencio, dt. "Blick-Garten-Stille") und die *novela sentimental* (Topic 45). Wenig überraschend auch, dass die *novela histórica* als distinktiven Topic unter anderem Topic 57 hat (rey-caballero-príncipe, dt. "König-Ritter-Prinz"). Für die *novela histórico-social*, für die aufgrund der großen Zahl von Beispielen eine größere Bandbreite an Topic-Verteilungen zu erwarten ist, gibt es keinen vergleichbar stark distinktiven Topic. Dennoch sind die Untergattungen ein wichtiger Faktor für die Verteilung der Topics in der Sammlung und die thematische Komponente spielt für die Definition der Untergattungen tatsächlich eine wesentliche Rolle.

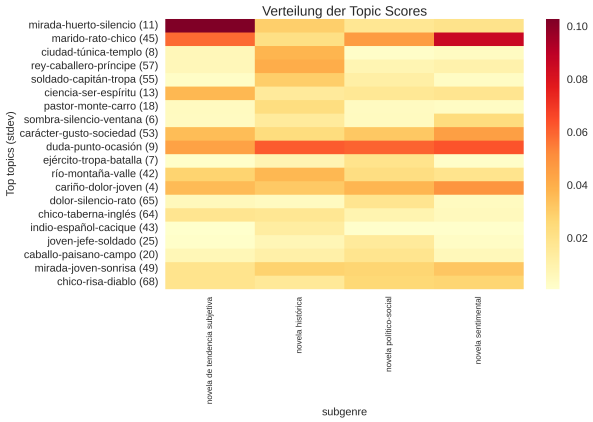


Abb. 3: Verteilung von Topic-Scores nach Untergattungen.

Topics im Textverlauf

Die Ausprägung der Topics variiert nicht nur hinsichtlich der Untergattungen, sondern auch über den Textverlauf hinweg. So gibt es einige Topics, deren Vorkommen am Anfang der Romane besonders wahrscheinlich ist (Abbildung 4a). Dazu zählen Topic 10 (vino-plato-pan, dt. "Wein-Teller-Brot"), Topic 17 (sombrero-ropa-bota, dt. "Hut-Kleidung-Stiefel") und Topic 19, welche auf die Beschreibung von Ambiente, Situation und Personen hindeuten. Gegen Ende der Romane sind andere Topics wahrscheinlicher (Abbildung 4b), z. B. Topic 2 (pecado-caridad-conciencia, dt. "Sünde-Wohltätigkeit-Gewissen"), Topic 23 und Topic 69, also abstraktere Themen oder solche, die sich auf Wertvorstellungen beziehen. Dies deutet darauf hin, dass in den Romanen am Ende Bilanz gezogen wird, die Handlung einen drastischen Ausgang nimmt oder das im Textverlauf Behandelte in gesellschaftliche oder religiöse Diskurse eingebunden wird.

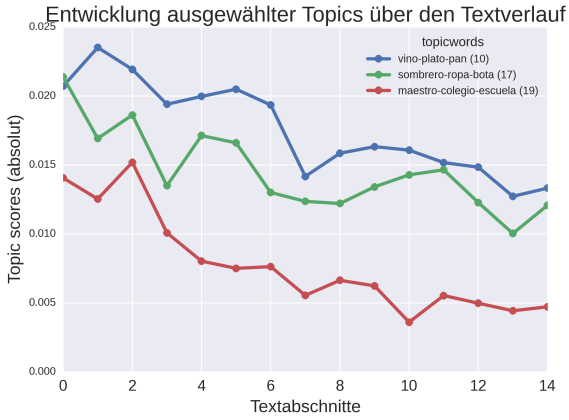


Abb. 4a: Verteilung von Topics im Textverlauf (fallend).

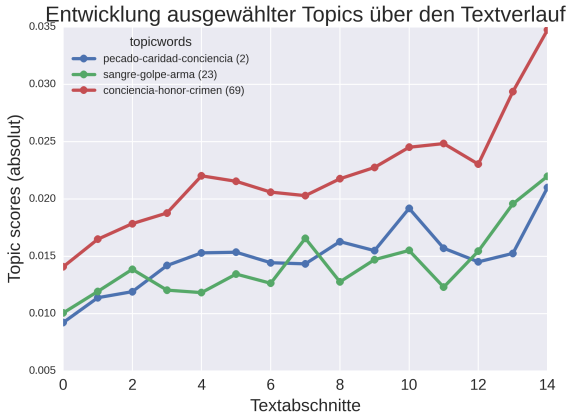


Abb. 4b: Verteilung von Topics im Textverlauf (steigend).

Textverlauf abhängig von den Untergattungen

Für einige der genannten Topics, die in bestimmten Bereichen des Textverlaufs wahrscheinlicher sind, kann die Tendenz über alle Untergattungen hinweg bestätigt werden (bspw. bei Topic 10 und 17, siehe oben). Es gibt aber auch Themen, bei denen sich durch die Betrachtung des Verlaufs in den einzelnen Untergattungen ein differenzierteres Bild ergibt. Die Wahrscheinlichkeit von Topic 23 beispielsweise nimmt nur für die *novela político-social* zum Ende hin zu (Abbidung 5a):

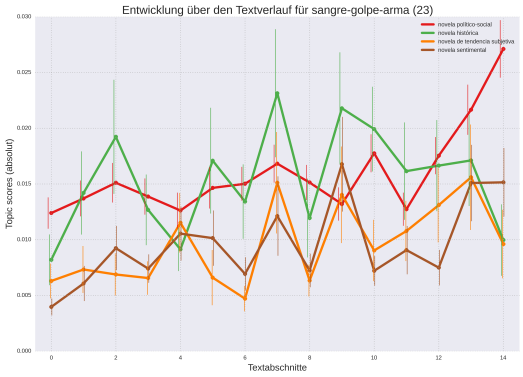


Abb. 5a: Topic 23 nach Textverlauf und Untergattung.

Das kann so interpretiert werden, dass die *novela político-social* im Gegensatz zu den anderen Untergattungen dazu tendiert, am Ende des Textes mit einer gewalttätigen Szene und einem Umbruch zu schließen. Topic 19 ist nicht in allen Untergattungen zu Beginn des Textverlaufes stark ausgeprägt, sondern nur bei der *novela de tendencia subjetiva*. Dies erklärt sich, weil bei diesen Romanen das Schulthema als Teil einer autofiktionalen Erzählung zu Beginn erscheint (Abbildung 5b):

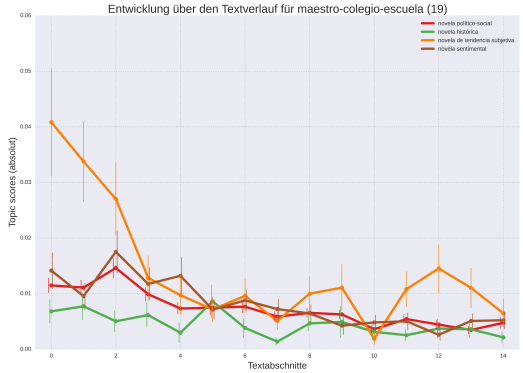


Abb. 5b: Topic 19 nach Textverlauf und Untergattung

Allgemein gilt, dass die Untergattungen sich in ihrer Topicverteilung im Textverlauf auch dann deutlich unterscheiden können, wenn dies für alle Untergattungen zusammengenommen nicht der Fall ist und so leicht übersehen werden könnte.

Für die Berechnung wurden die Romansegmente von 600 Wörtern bezüglich des Textverlaufs auf 15 Romanabschnitte (Bins) verteilt, um die unterschiedliche Romanlänge zu berücksichtigen. Diese Bins wurden hinsichtlich der Untergattung gruppiert und jeweils das arithmetische Mittel bestimmt. Die in den Plots eingezeichneten Kurven entsprechen der linearen Interpolation dieser gemittelten Werte. Zusätzlich wurde der Standardfehler vertikal um den jeweiligen Kurvenpunkt eingezeichnet, der deutlich macht, wie sehr die jeweiligen dem Mittelwert zugrunde liegenden Werte streuen, also wie gut der Mittelwert die Gesamtheit der Segmentwerte repräsentiert.

Die Ergebnisse im literaturgeschichtlichen Kontext

Insgesamt zeigen sich verschiedene Zusammenhänge: Zwischen bestimmten Topics und einzelnen Roman-Untergattungen, zwischen Topics und dem Textverlauf, und dies zum Teil dann auch wieder in Abhängigkeit von den Untergattungen. Aus literaturgeschichtlicher Perspektive betrachtet erweisen sich die in die Untersuchung einbezogenen Metadaten für eine Einordnung der Topic-Resultate als nützlich. Topics sind für die Romangattungen im vorliegenden Korpus ein wichtiger Faktor, ähnlich wie dies für Gattungen wie die klassische Komödie und Tragödie bereits gezeigt werden konnte (Schöch 2015).

Ein detaillierterer Blick zeigt beispielsweise Folgendes: Topic 11, welches typisch für die *novela de tendencia subjetiva* ist, ist vor allem in den 1910er- und 1920er-Jahren wichtig sowie für bestimmte Autoren. Interessanterweise ist dieses bei spanischen und hispanoamerikanischen Modernisten vorkommende Thema auch bei der früher wirkenden Schriftstellerin Juana Manuela Gorriti schon wichtig, die offenbar thematische Präferenzen späterer Autoren vorweggenommen hat. Außerdem kommt Topic 11 bei Larreta in einem (modernistischen) historischen Roman vor, obwohl es ansonsten vor allem für die Romane subjektiver Tendenz typisch ist. Es ist anzunehmen, dass für dieses spezielle Thema eher die literarische Strömung bestimmend ist als die Untergattung. Der Topic enthält einige für die modernistische Strömung typische Wörter, etwa zu Sinneseindrücken (azul, dt. "blau", olor, dt. "Geruch") und Zurückgezogenheit (huerto, silencio, campo, soledad, dt. "Garten, Ruhe, Land, Einsamkeit").

Fazit und Ausblick

Die Nutzung von Topic Modeling als Methode kann für die digitale Literaturwissenschaft verbessert werden, wenn spezifisch literaturwissenschaftliche Metadaten in die Betrachtungen einbezogen werden und die Textstruktur - hier als Sequenz von Textverlaufseinheiten - berücksichtigt wird. Verschiedene Visualisierungsstrategien erweisen sich als entscheidende "Interfaces" zu den Daten (im Sinne von Doueihi 2012), die Muster sichtbar machen und den Blick lenken. Die Ergebnisse des Topic Modelings können differenzierter und aus verschiedenen Perspektiven betrachtet und mit literaturhistorischem Wissen in Verbindung gebracht werden. Die Ergebnisse ergänzen und erweitern etablierte hermeneutische Lektürestrategien, insofern

sie einen synthetisierenden Blick auf sehr umfangreiche Textsammlungen erlauben.

Nächste Schritte betreffen insbesondere die weitere Auseinandersetzung mit der Signifikanz von Unterschieden in den Topic-Wahrscheinlichkeiten im Textverlauf, deren Berechnung u. a. durch die mangelnde Normalverteilung der Werte nicht trivial ist. Zusätzlich zu den Untergattungen sollen auch Kategorien wie das Setting modelliert werden. Zudem sollen die Textverlaufs-Daten für die automatische Klassifikation von Romanen nach Untergattungen genutzt werden. Schließlich wird bereits an der Erweiterung der Textsammlung gearbeitet, insbesondere mit Blick auf den Umfang und ein ausgeglicheneres Verhältnis der Untergattungen.

Bibliography

Altisent, Marta E. (2008): *A Companion to the Twentieth-Century Spanish Novel*. Woodbridge: Tamesis.

Blei, David M. (2011): "Introduction to Probabilistic Topic Models," in: *Communication of the ACM*.

Blei, David M. / Ng, Andrew Y. / Jordan, Michael I. (2003): "Latent Dirichlet Allocation," in: *Journal of Machine Learning Research* 3: 993–1022.

Blevins, Cameron (2010): "Topic Modeling Martha Ballard's Diary," in: *Historying* http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/ [letzter Zugriff 16. Februar 2016].

Doueihi, Milad (2012): *Pour un humanisme numérique (2011)*. Paris: Seuil.

Gallo, Marta (1981): *La Novela Hispanoamericana En El Siglo XIX*. Madrid: La Muralla.

García de Nora, Eugenio (1963): *La Novela Española Contemporánea*. Madrid: Gredos.

Jockers, Matthew L. (2013): *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Jockers, Matthew L. (2015): "Revealing Sentiment and Plot Arcs with the Syuzhet Package" in: *Matthew. L. Jockers* http://www.matthewjockers.net/2015/02/02/syuzhet/ [letzter Zugriff 09. Februar 2016].

McCallum, Andrew K. (2002): *MALLET: A Machine Learning for Language Toolkit* http://mallet.cs.umass.edu [letzter Zugriff 09. Februar 2016].

Nachwuchsgruppe CLiGS (o.J.): *Computergestützte literarische Gattungsstilistik* http://cligs.hypotheses.org/ [letzter Zugriff 16. Februar 2016].

Rhody, Lisa M. (2012): "Topic Modeling and Figurative Language," in: *Journal of Digital Humanities* 2 http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody [letzter Zugriff 09. Februar 2016].

Schmidt, Benjamin M. (2014): "Typical TV Episodes: Visualizing Topics in Screen Time," in: *Sapping Attention* http://sappingattention.blogspot.de/2014/12/typical-tv-episodes-visualizing-topics.html [letzter Zugriff 09. Februar 2016].

Schöch, Christof (2015): "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama [submitted]", in: *Digital Humanities Quarterly*.

Williams, Raymond L. (2009): *The Twentieth-Century Spanish American Novel*. Austin, Texas: University of Texas Press.

LERA - Explorative Analyse komplexer Textvarianten in Editionsphilologie und Diskursanalyse

Schütz, Susanne
susanne.schuetz@romanistik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg

Pöckelmann, Marcus
marcus.poeckelmann@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg

Ausgangssituation und Projektrahmen

Der Vergleich von Textfassungen ist eine zentrale Aufgabe der Editionsphilologie und Grundlage für die Erstellung von Variantenapparaten in kritischen Editionen. Zudem lässt sich über den Textvergleich die inhaltliche Evolution von Texten nachvollziehen. Je umfänglicher oder komplexer die zu edierenden Werke sind, desto schwieriger ist es für die Editoren und die Nutzer von Editionen den Überblick über die Textfassungen zu behalten. Mit dem hier vorgestellten, in die Arbeitsumgebung LERA (Bremer et al. 2015) integrierten Ansatz, wird diesem Problem mit einer Kombination von graphischen Werkzeugen begegnet, deren Zusammenwirken den Überblick über große Textmengen und ihre inhaltliche Auswertung erleichtert und zudem die Möglichkeit zum ergbnisoffenen Erkunden der Textvarianten bietet.

LERA ist eine interaktive, webbasierte Arbeitsumgebung zur Untersuchung mehrerer Fassungen eines Textes. Sie wurde im Rahmen des vom Bundesministeriums für Bildung und Forschung geförderten Projekts: *Semiautomatische Differenzanalyse von komplexen Textvarianten (SaDA)* entwickelt (Medek et al. 2015). Als Fallbeispiel für die Entwicklung von LERA wurde ein Buch aus der 'Histoire philosophique et politique des établissements et du commerce des Européens dans les deux Indes' von Thomas-Guillaume Raynal, einem der gesamteuropäisch einflussreichsten Erfolgs- und Skandalbücher des 18. Jahrhunderts,

gewählt. Das Werk ist eine kritische Auseinandersetzung mit der europäischen Kolonialexpansion in den ‚beiden' Indien und liegt in vier stark überarbeiteten Fassungen aus den Jahren 1770, 1774, 1780 und 1820 vor (Schütz / Pöckelmann 2014).

Für die Kollationierung verschiedener Textfassung gibt es bereits eine Reihe digitaler Werkzeuge. Drei der bekanntesten, CollateX , Juxta und TUSTEP , liefern für viele Fragestellungen gute Ergebnisse, unterscheiden sich in ihren Anwendungsszenarien aber erkennbar von LERA. Während LERA einen zweistufigen Ansatz verfolgt, bei dem vor dem detaillierten Textvergleich zunächst eine Alignierung größerer Textpassagen berechnet wird, was denEinsatz komplexer Signaturwerte [1] für den Vergleich notwendig macht, liegt der Fokus von CollateX auf der Alignierung (normalisierter) Token, für den auf einfache Zeichenketten als Signaturwerte zurückgegriffen wird. Juxta zeigt mit seinen hilfreichen Visualisierungen die Unterschiede zu einer ausgewählten Leithandschrift auf, während LERA für den direkten Vergleich mehrerer Textfassungen konzipiert wurde und stets die Textänderungen zwischen allen Fassungen darstellt. Der wohl vielseitigste digitale Werkzeugkasten aus dem Bereich der Geisteswissenschaften, TUSTEP, bietet ebenfalls Möglichkeiten zum Vergleich verschiedener Textfassungen. Allerdings erfordert der Umgang mit dem textbasierten Interface eine längere Einarbeitungszeit, die sich zumindest für einige Anwender durch die neu entwickelte XML-basierte Variante TXSTEP verringert. LERA wurde hingegen von Beginn an durch interaktive graphische Elemente als intuitiv bedienbare Arbeitsumgebung entwickelt. Veränderte Vergleichsparameter sollen dabei durch schnelle Neuberechnung eine direkte Präsentation des Ergebnisses ermöglichen und so zum Experimentieren einladen.

Explorative Analyse mit LERA

Die Arbeitsumgebung erzeugt eine synoptische Gegenüberstellung von größeren Textsegmenten, welche die Grundlage für den Textvergleich bilden. Für die „Histoire des deux Indes", die in vielen Drucken mit variierendem Zeilenfall vorliegt, wurden Absätze als Vergleichsebene gewählt. Unterschiede zwischen den Textfassungen werden durch Farbmarkierungen in der Synopse und in einem gemeinsamen Variantenapparat dargestellt. Das Vergleichsergebnis kann durch Filtereinstellungen beeinflusst und so an verschiedene Fragestellungen angepasst werden. Beispielweise lassen sich orthographische Varianten ausblenden, so dass dem Nutzer hauptsächlich die inhaltlichen Unterschiede präsentiert werden.

Aufbauend auf dieser Grundfunktionalität wurden die drei im Folgenden beschriebenen Komponenten in LERA integriert, die das Überblicken und Auswerten der Textunterschiede erleichtern sollen.

Integrierte Suche

Als erster wichtiger Baustein wurde eine Suchfunktion eingebunden, die das schnelle Auffinden einzelner Schlagwörter ermöglicht. Das Suchfenster zeigt an, wie häufig der – intern normalisierte – Begriff in der gesamten Synopse vorkommt. Die Treffer der Suche sind dabei in den Texten farbig unterlegt und können über Navigationspfeile angesteuert werden.

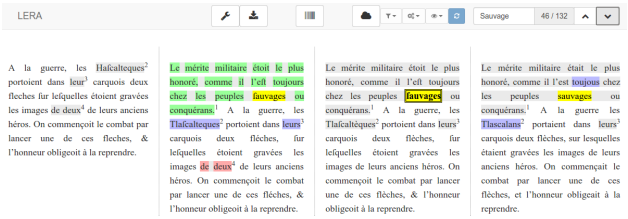


Abb. 1: Suchfunktion in LERA: Die gelb hervorgehobenen Suchtreffer können durch die Schaltflächen der Suchmaske (oben rechts) angesteuert werden.

CATview: Strukturelle Übersicht

Der oben beschriebene Textvergleich bildet die Datenbasis der integrierten interaktiven Übersichts- und Navigationsleiste CATview (Pöckelmann et al. 2015). Die Synopse wird schematisch in CATview dargestellt, indem jedes Segment der Vergleichstexte als Rechteck abgebildet und entsprechend seiner Position in der Synopse angeordnet wird. Die Farbe der Rechtecke zeigt die Intensität der Überarbeitung an. Je stärker dabei ist die Veränderung des Textsegments ausfällt, desto dunkler wird das Farbfeld dargestellt. So verschafft CATview einen Überblick über die Struktur der Texte und die Verteilung der Unterschiede. Zudem erleichtern weitere Funktionen wie die Verlinkung der Rechtecke mit den entsprechenden Textsegmenten die Navigation in der Synopse.

Abb. 2: Die in LERA integrierte CATview stellt Struktur und Unterschiede zwischen den Textfassungen schematisch dar und erleichtert die Navigation durch Verlinkungen und einen Bildlauf-Indikator (orange).

Wortwolken: Thematische Übersicht

Zusätzlich zum synoptischen Textvergleich und dessen Visualisierung in CATview kann die Arbeitsumgebung LERA für jede Textfassung eine interaktive Wortwolke generieren, die einen Vergleich des Auftretens und der Häufigkeit von Schlagwörtern ermöglicht und so einen Überblick über die Inhalte der jeweiligen Textfassung gibt. Welche Wörter dabei angezeigt werden, hängt von den Nutzervorgaben ab. So kann beispielsweise die Auswahl der Wörter auf einzelne Wortarten beschränkt

2. Die Bestimmung der Wortarten für die Wortwolken erfolgt automatisch mit Hilfe des TreeTaggers . Siehe „TreeTagger - a language independent part-of-speech tagger“.
oder eine Mindestlänge festgelegt werden. Zudem sind das Bewertungskriterium (Häufigkeit, Tf-idf-Maß) und verschiedene Darstellungsformen wählbar. Durch die Gegenüberstellung der Wortwolken aller Textfassungen wird das Erkennen einer veränderte Wortwahl und Themensetzung zwischen den Textfassungen erleichtert.

Abb. 3: Von LERA generierte Wortwolken für vier Textfassungen: Entsprechend der Nutzereinstellungen wurden als eine der wählbaren Darstellungsformen dreidimensionale Kugeln mit den jeweils 20 häufigsten Wörtern generiert und die Wortgröße über alle Fassungen hinweg ermittelt. So wird beispielsweise die veränderte Häufigkeit des Begriffs „espagnol“ kenntlich (67, 77, 63 und 81).

Das Zusammenspiel der Werkzeuge

Die innovative Kombination der beschriebenen Ansätze zur Analyse und Darstellung von Varianten ermöglicht das effiziente Erkunden umfangreicher Texte und die wirkungsvolle Visualisierung von Suchergebnissen. Die Arbeitsumgebung bietet so verschiedene Zugänge zur explorativen Analyse an, von denen im Folgenden drei Kombinationsmöglichkeiten erläutert werden.

Kombination von CATview und Suchfunktion

Die Übersichtsleiste CATview zeigt die Überarbeitung der Textfassungen an und vereinfacht die Navigation in umfangreichen Synopsen. Kombiniert man dieses Werkzeug mit der Suchfunktion, werden die Treffer der Suche nicht nur farbig im Text unterlegt, sondern ebenfalls in den Rechtecken der Übersichtsleiste angezeigt. So kann die Verteilung eines Themas über den gesamten Text auf einen Blick erfasst werden. Mit dieser Funktion lassen sich zudem korrespondierende Textstellen, die nicht direkt aligniert sind, leichter auffinden.

Kombination von Suche und Wortwolken

Die Ergebnisse der Suche werden in den interaktiven Wortwolken ebenfalls farbig hervorgehoben, was das Auffinden der Schlagworte und damit den Vergleich der Relevanz bzw. Häufigkeit in den verglichenen Textfassungen erleichtert. Neben dieser intendierten Suche kann die Kombination der beiden Komponenten den Nutzern Anregungen für die weitere Textexploration geben. In den Wortwolken sind Begriffe enthalten, die auf Grund ihrer Häufigkeit eine Relevanz für den Textinhalt vermuten lassen, ein Klick auf die angezeigten Begriffe in den Wortwolken löst die Suche aus.

CATview in Kombination mit Wortwolken und Suchfunktion

CATview bietet ein Auswahlwerkzeug an, mit dem der Textbereich für die Analyse eingeschränkt werden kann. Dieses Vorgehen bietet sich an, wenn eine Passage des Textes wegen ihrer Überarbeitung oder der Verteilung von Suchbegriffen besonders interessant erscheint. Durch die Beschränkung der Textauswahl verändert sich sofort die Zusammensetzung der Wortwolken, da auch ihre Datengrundlage eingeschränkt wird, indem nur noch die Begriffe aus der Auswahl betrachtet werden. Die Wortwolken werden für jede neue Auswahl umgehend aktualisiert. Die Auswahlbox kann durch drag-and-drop in der Übersichtsleiste bewegt werden. Durch diese Bewegung verändern sich folglich auch die Wortwolken, sodass stets die Themen der aktuell gewählten Textpassage angezeigt werden. Dadurch erzeugt die Bewegung eine Art interaktive Animation an Hand derer die inhaltliche Entwicklung der Texte veranschaulicht wird. Dieses Vorgehen kann zu neuen interessanten Fragestellungen führen.

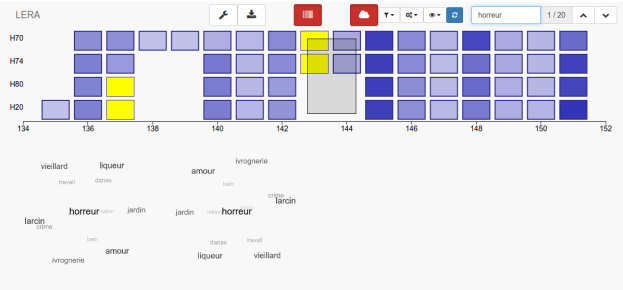


Abb. 4: Beispiel für das Zusammenspiel von Suche, CATview und den Wortwolken: Mit Hilfe des Auswahlwerkzeugs von CATview wurden die Absätze 143 und 144 markiert, was sogleich die Wortwolken aktualisiert und so eine grobe inhaltliche Übersicht dieser aus H80 entfernten Textpassage erzeugt. Mit einem Klick auf den Begriff „horreur“ in den Wolken wird eine Suche gestartet, die wiederum in CATview durch die Markierung der Suchtreffer einen Hinweis dafür liefert, dass Teile der Passage bei der Überarbeitung in den Absatz 137 eingeflossen sind.

Zusammenfassung

Neben der Grundfunktionalität zum Vergleich mehrerer Fassungen eines Textes entsprechend getroffener Nutzereinstellungen wurden in die Arbeitsumgebung LERA mit der Suchfunktion, der Übersichts- und Navigationsleiste CATview sowie den interaktiven Wortwolken drei Komponenten zum Analysieren der gefundenen Unterschiede integriert, deren Kombination das traditionelle Vorgehen der systematischen Textauswertung um die ergebnisoffene Exploration erweitert. So führt das willkürliches Experimentieren mit diesen Komponenten gegebenenfalls zu unerwarteten Erkenntnissen und neuen Hypothesen. Das ist insbesondere für Nutzer von digitalen Editionen interessant, die oft andere Fragestellungen verfolgen als die ursprünglichen Editoren. So werden das Nachvollziehen von Argumentationsketten und die historische Veränderung von Diskursen im Zuge der Überarbeitung von Texten effektiv unterstützt.

Anmerkungen

Diese Arbeit wurde durch das Bundesministerium für Bildung und Forschung (BMBF) [Projektkürzel: 01UG1247 / human-325-010 / SaDA] im Rahmen des Projekts „Semi-automatische Differenzanalyse von komplexen Textvarianten“ unter Leitung von Prof. Dr. Thomas Bremer, Prof. Dr. Paul Molitor, Dr. Jörg Ritter und Prof. Dr. Hans-Joachim Solms gefördert.

Weitere Informationen samt Demonstratoren zu LERA finden Sie auf der Projektseite von SaDA: http://sada.uzi.uni-halle.de

Notes

1. Für einen Absatz fließen beispielsweise dessen Position und enthaltene signifikante Wörter in die Bestimmung des Signaturwerts ein. Ein Wort gilt dabei als signifikant, wenn es in mindestens zwei der zu vergleichenden Texte vorkommt und in der allgemeinen Sprachverwendung relativ selten ist.

Bibliographie

ARP (2012): *Juxta*. Compare - Collate - Discover. http://www.juxtasoftware.org/ [letzter Zugriff 15. Oktober 2015].

Bremer, Thomas / Molitor, Paul / Pöckelmann, Marcus / Ritter, Jörg / Schütz, Susanne (2015): "Zum Einsatz digitaler Methoden bei der Erstellung und Nutzung genetischer Editionen gedruckter Texte mit verschiedenen Fassungen - Das Fallbeispiel der Histoire philosphique des deux Indes von Guillaume Thomas Raynal" in: Nutt-Kofoth, Rüdiger / Plachta, Bodo / Woesler, Winfried (eds.) *Editio*. Internationales Jahrbuch für Editionswissenschaften 29, 1: 29–51.

Bremer, Thomas / Molitor, Paul / Ritter, Jörg / Solms, Hans-Joachim(2012-2015): *Semi-automatische Differenzanalyse von komplexen Textvarianten* http://sada.uzi.uni-halle.de [letzter Zugriff 15. Oktober 2015].

Medek (*Gießler), André / Pöckelmann, Marcus / Bremer, Thomas / Solms, Hans-Joachim / Molitor, Paul / Ritter, Jörg (2015): "Differenzanalyse komplexer Textvarianten - Diskussion und Werkzeuge", in: *Datenbank-Spektrum* 15, 1: 25-31.

Pöckelmann, Marcus / Medek (*Gießler), André / Molitor, Paul / Ritter, Jörg (2015): "CATview - Supporting The Investigation Of Text Genesis Of Large Manuscripts By An Overall Interactive Visualization Tool", in: *Digital Humanities, DH2015, Sydney, Australia, 29.06.-03.07.2015*.

Pöckelmann, Marcus / Molitor, Paul / Ritter, Jörg (2015): *CATview*. The Colored and Aligned Texts view http://catview.uzi.uni-halle.de/ [letzter Zugriff 15. Oktober 2015].

Schmid, Helmut(1994-): *TreeTagger*. A language independent part-of-speech tagger http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [letzter Zugriff 15. Oktober 2015].

Schütz, Susanne / Pöckelmann, Marcus (2014): "IT-Werkzeuge zur Unterstützung elektronischer Edition am Beispiel eines französischen Textes aus dem 18. Jahrhundert", in: *9. Kongress des Frankoromanistenverbands, Münster, 24.-27.09.2014*.

The Interedition Development Group (2010-2013): *CollateX*. Software for Collating Textual Sources http://collatex.net/ [letzter Zugriff 15. Oktober 2015].

Zentrum für Datenverarbeitung der Universität Tübingen (1978-): *TUSTEP*. Tuebingen System of Text

Processing tools http://www.tustep.uni-tuebingen.de/ [letzter Zugriff 15. Oktober 2015].

Zentrum für Datenverarbeitung der Universität Tübingen / pagina GmbH Publikationstechnologien / Hochschule der Medien Stuttgart (2010-): *TXSTEP*. Die XML-Version von TUSTEP http://www.txstep.de/ [letzter Zugriff 15. Oktober 2015].

Tambora: Erkenntnisgewinne durch kartographische Visualisierungen von Forschungsdaten aus der historischen Klimatologie

Specht, Sebastian
s_specht@ifl-leipzig.de
Leibniz-Institut für Länderkunde, Deutschland

Christian, Hanewinkel
c_hanewinkel@ifl-leipzig.de
Leibniz-Institut für Länderkunde, Deutschland

Sebastian, Koslitz
s_koslitz@ifl-leipzig.de
Leibniz-Institut für Länderkunde, Deutschland

Heike, Steller
h_steller@ifl-leipzig.de
Leibniz-Institut für Länderkunde, Deutschland

Mit der virtuellen Forschungsumgebung Tambora.org wird Wissenschaftlern eine IT-Infrastruktur zur Verfügung gestellt, die den ganzen Prozess historisch-klimatologischer Quellenarbeit abbildet. Nach der Recherche, Katalogisierung und Erfassung von Texten mit klimarelevanten Hinweisen bildet die räumliche, zeitliche und umweltgeschichtlich-klimatologische Kodierung der erfassten historischen Quellentexte ein zentrales Element der Forschungsumgebung (Borel / Steller 2012). Karten und Visualisierungen spielen in diesem Prozess eine große Rolle (Specht / Hanewinkel 2013).

Die erfassten Forschungsdaten variieren stark hinsichtlich ihres Detailgrades und der Datendichte entlang der Dimensionen Zeit und Raum. Die Benutzung heute gebräuchlicher Referenzsysteme zur Kodierung (zum Beispiel das räumliche WGS84-Referenzsystem bzw. mittelbar über geographische Namensdatenbanken oder der gregorianische Kalender in seiner in ISO8601 niedergelegten Spezifizierung) sind dabei als gemeinsame Basis aller Forschungsprojekte vorgegeben. Spezielle Zusatzkodierungen wie zum Beispiel Jahreszeiten oder Sonnenauf- und untergangszeiten werden vom System bereitgestellt. Wie bei Raum und Zeit kann auch der Gehalt an historisch-klimatologischer Informationen mehr oder weniger spezifisch sein und von einer einfachen Verschlagwortung über ordinale Intensitätsskalen bis hin zu Messwertreihen kodiert werden. Referenzsysteme dieser Art sind für die Kodierung grundsätzlich nötig, um eine Verarbeitbarkeit der Daten zu ermöglichen und diese sinnvoll recherchierbar zu machen. Weiterhin können in allen Dimensionen Kodierfehler oder systematische Fehler nicht ausgeschlossen werden.

Die Arbeit an der Forschungsumgebung zeigt: Die gespeicherten Texte und kodierten Daten erfahren auch ein Interesse aus der allgemeinen Öffentlichkeit außerhalb der historischen Klimatologie. Gleichzeitig zeigen Menschen außerhalb der Forschergemeinde ein Interesse, ihnen zur Verfügung stehende historische Quellen in die Forschungsumgebung einzuarbeiten. Ob Popularisierung von Forschungsdaten oder „Citizen Science“: Mit der Erweiterung der Nutzerkreise stellen sich auch in Bezug auf die Kartennutzung eine Reihe von Fragen. Sind sich die Nutzer außerhalb der Fachwissenschaften der komplexen Eigenschaften historischer Texte wie Lückenhaftigkeit, Subjektivität und Glaubwürdigkeit bewusst? Wie ausgeprägt ist das Verständnis für klimatologische Phänomene und deren räumliche und zeitliche Ausdehnung und Wirkung? Können diese komplexen Qualitätseigenschaften der Inhalte mit Hilfe der Kartographie in den Visualisierungen kommuniziert werden? Werden die Auswirkungen des Kodierungsprozesses auf die in den Visualisierungen enthaltenen Informationen nachvollzogen?

Zentrales Ziel der Visualisierung ist neben dem der Kartographie innewohnenden Anliegen der Kommunikation räumlichen Wissens vor allem die gleichzeitige Herausarbeitung der räumlichen und zeitlichen Begrenzung dieses Wissens. Eine tragende Rolle kommt dabei der Basiskarte zu: Sie sollte frei von Anachronismen sein und in einem weiten Maßstabsbereich eine angemessene Projektion besitzen. Ebenso sollte neben der räumlichen Orientierungsfunktion auch eine inhaltsorientierte räumliche und zeitliche Begrenzung der gespeicherten Daten kommuniziert werden, um so zum Beispiel durch „weiße Flecken“ den entscheidenden Unterschied zwischen dem Nicht-Vorhanden-Sein von historischem Wissen und dem Nicht-Auftreten von klimatologischen Ereignissen zu verdeutlichen. Der „weiße Fleck“ ist dabei nicht nur ein Ausdruck von Leere, sondern ein in der kartographischen Tradition stehender, „auf das engste mit dem Medium der Karte [verbundener] Forschungsauftrag.“ (Schelhaas 2009). Aus historisch-klimatologischer Sicht handelt es sich dort um eine „terra incognita“ und die Karte ist, zumindest in der Raumdimension, die Matrix, auf der

entsprechende Orte mit Bedeutung konstruiert werden können (Paez i Blanch 2015).

Die kartographische Symbolisierung von hunderttausenden, auf dem oben beschriebenen Kodierprozess beruhenden raum-zeitlichen Ereignissen wird auf dieser Kartengrundlage durch Aggregationsmechanismen ermöglicht. Dabei darf der Charakter dieser Symbolisierung keine unangemessene Exaktheit vortäuschen. Verbliebene Restunsicherheiten in den Daten sollten durch die Art der visuellen Kommunikation vermittelt werden.

Durch die vielfältigen Tools innerhalb der Forschungsumgebung werden sehr unterschiedliche Sichten auf die Daten ermöglicht, die neben der gewünschten neuen Forschungserkenntnis im Speziellen vielleicht auch „citizen scientists“ zum weiteren Forschen in bisher unbekannten Räumen anregen soll. Erfahrungen im OpenStreetMap-Projekt bestätigen die motivierende Rolle von Visualisierungen: Leere Orte ziehen Aufmerksamkeit auf sich und das Füllen dieser Lücken wird von Nutzern als befriedigend empfunden (Budhathoki / Haythornthwaite 2013).

Diese Arbeit stellt den Aufbau und die Systemarchitektur der virtuellen Forschungsumgebung sowie deren Vernetzung in die Community und zu externen Diensten vor. Im Speziellen werden die im Projekt entstandenen Visualisierungen vorgestellt und zu ihrer Tauglichkeit in Bezug auf die formulierten Anforderungen und Ansprüche diskutiert. Die Vorstellung der offenen Datenschnittstelle des Projektes (Tambora.org REST API) soll unter anderem Kartographen oder auch Informatiker zu einem „map hacking“ einladen und zu alternativen Visualisierungen anregen. Diese Schnittstelle wird damit als integraler Bestandteil des kartographischen Kommunikationsprozesses betrachtet.

Bibliographie

Borel, Franck / Steller, Heike (2012): "Tambora – Die Entstehung einer virtuellen Forschungsumgebung", in: *B.I.T.online* 5: 423-430.

Budhathoki, Nama R. / Haythornthwaite, Caroline (2013): "Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap", in: *American Behavioral Scientist* 57, 5: 548-575.

Paez i Blanch, Roger (2015): "Mapping emptiness: cartographic activations of empty space", in: *Cuadernos De Proyectos Arquitectó Nicos* 5: 155-157 http://polired.upm.es/index.php/proyectos_arquitectonicos/article/view/3067 [letzter Zugriff 01. September 2015].

Schelhaas, Bruno (2009): "Das „Wiederkehren des Fragezeichens in der Karte“. Gothaer Kartenproduktion im 19. Jahrhundert", in: *Geographische Zeitschrift* 97, 4: 227-242.

Specht, Sebastian / Hanewinkel, Christian (2013): "Maps as Research Tools Within a Virtual Research Environment", in: Buchroithner, Manfred F. / Prechtel, Nikolas / Burghardt, Dirk / Pippig, Karsten / Schröter, Benjamin (eds.)*Proceedings of the 26th International Cartographic Conference*, Dresden, Germany http://icaci.org/files/documents/ICC_proceedings/ICC2013/ [letzter Zugriff 08. Januar 2016].

Usability in den Digital Humanities am Beispiel des LAUDATIO-Repositoriums

Stiller, Juliane
jstiller@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Deutschland

Thoden, Klaus
kthoden@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Deutschland

Zielke, Dennis
zielkede@cms.hu-berlin.de
Humboldt-Universität zu Berlin

Einleitung

Unter Usability verstehen wir die Gebrauchstauglichkeit einer Software, die ein Qualitätsmerkmal der Benutzeroberfläche darstellt und als Teilaspekt der gesamten User Experience anzusehen ist. Sie bildet die Grundlage positiver Interaktionen und ist ein wichtiger Faktor für den Erfolg eines Tools. Es spielt dabei eine entscheidende Rolle, inwieweit sich Nutzer_innen innerhalb der Anwendung durch die verständliche Bedienbarkeit und den zu erwartenden Workflow der einzelnen Funktionen zurechtfinden. Dabei ist es hilfreich, sie von Anfang an mit in den Entwicklungsprozess einzubinden und ihre Arbeitsmethoden kennenzulernen. Dazu dienen auch Mockups wie das Beispiel in Abbildung 1 illustriert. In diesem Vortrag wollen wir darstellen, wie Usability in den Digital Humanities evaluiert werden kann und welche Probleme sich generell bei der Softwareentwicklung in der Forschung auftun. Dafür werden wir grundlegende Überlegungen zur Usability-Problematik anstellen. Anhand von Nutzungsszenarien am Beispiel des im Projekt LAUDATIO [1] (weiter-)entwickelten Forschungsdatenrepositoriums für historische Textdaten demonstrieren wir, wie eine kostengünstige und einfache Usability-Evaluierung stattfinden kann. Als Verfahren wurden hierfür der Thinking-Aloud Test (Lewis /

Rieman 1994), eine der wichtigsten Methoden für die nutzerorientierte Evaluation, sowie eine heuristische Evaluation angewandt.

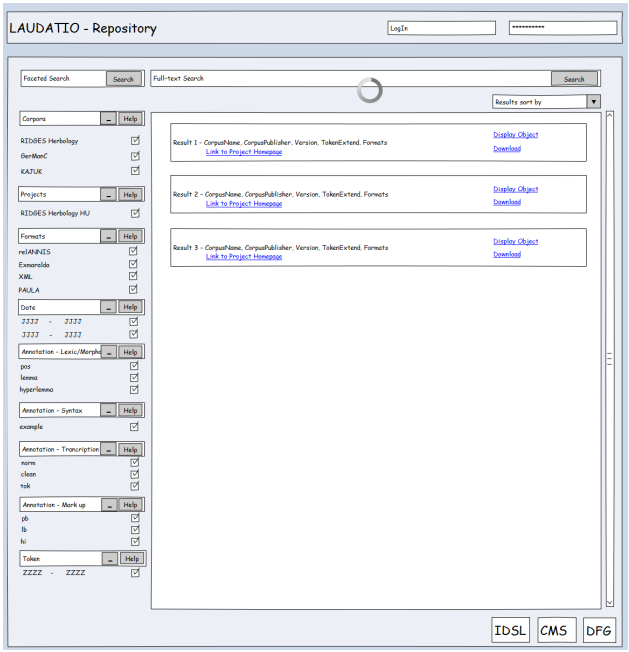


Abb. 1: Mockup der facettierten Suche mit Suchergebnissen

Motivation und Hintergrund

Die Usability von Tools und deren grafischen Benutzeroberflächen ist essentiell, um Nutzer_innen ein effektives Arbeiten zu ermöglichen und Bedienfehler zu minimieren. Dabei muss man sich in der großen Auswahl bereits existierender, einschlägiger DH-Tools [2] entscheiden, welche Anwendung für ein jeweiliges Szenario geeignet ist.

Was in großen Teilen der Software-Entwicklung Standard ist, findet nur langsam Einzug in die Entwicklung von Tools für die geisteswissenschaftliche Forschung. So stellt Burghardt (2012) eine Kluft zwischen den Tool-Entwickler_innen und den die Tools nutzenden Fachwissenschaftler_innen heraus, die sich beispielsweise in fehlendem Usability-Engineering für linguistische Annotationswerkzeuge niederschlägt. Anhand eines Heuristic Walkthroughs [3] erstellt er Design-Empfehlungen für Annotationswerkzeuge. Er unterscheidet allgemeine und domainspezifische Usability-Probleme. Der Grund für nutzerunfreundliche Funktionen sind häufig mangelnde Ressourcen, aber auch fehlende Expertise auf dem Gebiet – dadurch können entwickelte Tools fehleranfälliger sein, besonders wenn sie nicht ausreichend getestet wurden. Gibbs und Brown (2012) sehen die fehlende Investition in Nutzeroberflächen als Grund für die Ablehnung von Tools durch frustrierte Nutzer_innen. Oftmals sind Software oder Tools von enthusiastischen EntwicklerInnen nur für ein Nischenpublikum gebaut, und die Usability hat bei der Entwicklung keinen hohen Stellenwert.

Bisher wurde der Bedienbarkeit und Benutzbarkeit von digitalen Tools und Anwendungen im Bereich der Digital Humanities wenig Beachtung geschenkt. Aber gerade hier kann eine größere Akzeptanz und Nutzung des Tools/Services erreicht werden. Dabei muss eine Balance gefunden werden zwischen dem finanziellen Mehraufwand durch Umsetzung von Usability-Kriterien und der Spezifität des Tools, die eventuell nur einer kleinen Nutzergruppe zu Gute kommt.

Die Kriterien nach Nielsen (1995) sollten auch bei der Entwicklung geisteswissenschaftlicher Tools/Services Anwendung finden, da sie enorm wichtig sind, um die Validität und die Nachprüfbarkeit der Ergebnisse, die durch ein solches Tool zustande kommen, zu fördern. Ein gut nutzbares Tool hilft dabei, Fehler zu vermeiden und Daten nachnutzbar zu machen – somit wird auch das Vertrauen gefördert.

Usability am Beispiel des LAUDATIO Repositoriums

Das evaluierte Repositorium, in dem die Daten gespeichert werden, basiert auf Open-Source-Technologien [4]. Dessen Software-Module entsprechen allgemein den Anwendungsfunktionen Präsentation, Speicherung, Importieren und Suche. Basierend darauf wurde zuerst eine heuristische Evaluation [5] durchgeführt, bei der zwei Experten anhand der Usability-Heuristiken von Nielsen das Repositorium begutachteten. Dabei lag der Fokus auf diesen näher spezifizierten Funktionen von LAUDATIO:

- Präsentation von Korpora,
- Suche nach Korpora und in deren Dokumenten und Annotationen,
- Herunterladen von Korpora,
- Importieren von neuen und erweiterten Korpora,
- Auseinandersetzung mit den verschiedenen Beschreibungen der einzelnen Korpora.

Hierbei wurden folgende Problematiken ausgemacht, die einer Nutzbarkeit des Systems abträglich sind :

1. Inkonsistenz in der Benennung und Verwendung von mehrdeutigem Vokabular und Missachtung graphischer Konventionen
2. Intransparenz des Systemstatus
3. Fehlende oder schwer auffindbare Dokumentation
4. Fehlende Strategien zur Fehlerverhütung
5. Missachtung von Konventionen in der Suche

Im nachfolgenden Schritt wurden von Seiten der EntwicklerInnen 10 Aufgaben für einen Usability-Test erstellt, die sich aus Nutzeranfragen und dem alltäglichen

Arbeiten mit dem Repositorium ergaben. Diese Aufgaben wurden mit einer Testperson durchgeführt, wobei zwei Personen die Aktionen und die Kommentare der Probandin protokollierten, sowie Screencasts der einzelnen Sessions aufzeichneten. Hierfür wurde seitens der Betreiber ein Testkorpus mit Metadaten zur Verfügung gestellt, wofür XML-Kenntnisse vorausgesetzt wurden. Die Testperson führte den Test zum einen als nicht eingeloggte Nutzerin und zum anderen als eingeloggte Nutzerin mit erweiterten Rechten durch. Im zweiten Fall konnte sie über den Importprozess ein selbständig verändertes Korpus hochladen und anzeigen lassen. Obwohl die Testerin nicht aus dem Bereich der Linguistik kam und sich deswegen auch erst in die Begrifflichkeiten und Funktionen des Portal einarbeiten musste, wurden erhebliche Usability-Probleme eminent. In einem gemeinsamen Gespräch mit den Betreibern des LAUDATIO Repositoriums wurden die Problematiken benannt und eine Priorisierung der Usability-Probleme vorgenommen. Dabei wurden folgende Punkte als Kriterium für die Priorisierung herangezogen:

- Schwere der Problems / Störfaktor bei der Benutzung des Repositoriums
- Häufigkeit, mit der die Nutzer mit dem Problem konfrontiert sind
- Menge der Ressourcen, die für die Behebung des Problems nötig wären

Anhand der Priorisierungsliste wurden Verbesserungen in einem noch unveröffentlichten Prototypen vorgenommen und Massnahmen zur verbesserten Nutzerführung ergriffen. Diese sollen auch vorgestellt werden.

Schlussfolgerung

Damit digitale Tools in den Geisteswissenschaften kontinuierlich Verwendung finden, müssen sie auch einfach und effizient zu bedienen sein. Eine einfache Methode um die Usability kostengünstig und schnell testen zu können, bieten die von Nielsen (2012) aufgestellten Merkmale für eine nutzerfreundliche Software: Erlernbarkeit, Einprägsamkeit, Effizienz, Fehlertoleranz und Nutzerzufriedenheit. Spezifische Tools in den Geisteswissenschaften haben jedoch oft nicht die Möglichkeiten, eine grafische Benutzungsoberfläche anzubieten, die all diesen Kriterien vollends gerecht werden. Durch unsere Erfahrungen mit dem LAUDATIO Repositorium schlagen wir deshalb folgende Interpretation der Merkmale vor:

Erlernbarkeit: Die meisten geisteswissenschaftlichen Tools sind sehr speziell und setzen eine fachspezifische Grundausbildung voraus, die sich oft auch schon in dem verwendeten Vokabular niederschlägt. Die Erlernbarkeit sollte also darauf zielen, den NutzerInnen erklärende Tutorials und Dokumentationen anzubieten, die eine Einarbeitung verkürzen. Der Aufwand des Erlernens der Software sollte im Einklang mit dem Nutzen der Software stehen.

Einprägsamkeit: Hier geht es darum, dass NutzerInnen sich auch nach längerer Pause wieder in die Funktionalitäten des Tools einarbeiten können. Wie in Punkt 1 ist hier Dokumentation wichtig, aber auch eindeutiges Vokabular und prägnantes Design. Wenn es viele Parameter zur Einstellung gibt, sollten Möglichkeiten vorhanden sein, diese zu speichern oder zu exportieren, um sie später nachnutzen zu können.

Effizienz: Hier geht es um die Schnelligkeit, mit der NutzerInnen zum Ziel kommen. Dies beinhaltet neben schnellen Ladezeiten auch das Liefern von Systemfeedback für Nutzerinteraktionen.

Fehler: In diesem Punkt geht es um die Toleranz, mit der das System auch fehlerhafte Eingaben von NutzerInnen verarbeitet und zurückmelden kann und dieser trotzdem sein Ziel erreicht. Dies hat aber für geisteswissenschaftliche Tools noch eine andere Dimension: Verfahren und algorithmische Berechnungen müssen so transparent sein, dass NutzerInnen nicht durch falsche Bedienung Resultate, die einen Einfluss auf die Beantwortung der Forschungsfrage haben, zurückgeliefert bekommen. Es sollte immer klar sein, was das Tool mit welchen Daten macht und wie die Resultate zu interpretieren sind und warum welche Ergebnisse erzielt wurden. Es ist natürlich immer die Aufgabe der WissenschaftlerInnen, die sich des Tools bedienen, auf die Verlässlichkeit und Nachvollziehbarkeit ihrer Daten zu achten. Doch spielen hier die graphischen Benutzeroberflächen eine unterstützende Rolle und sollten gerade im wissenschaftlichen Bereich nicht in die Irre führen.

Nutzerzufriedenheit: Natürlich spielt auch die Nutzerzufriedenheit eine Rolle, die sich meist aus dem Vorhandensein der vorangegangen Qualitätsmerkmale ergibt.

In diesem Vortrag wollen wir unsere Erfahrung aus dem Usability-Test mit dem LAUDATIO Repositorium schildern und auf die daraufhin eingeleiteten Massnahmen zur Verbesserung der Nutzerführung eingehen.

Notes

1. Long-term Access and Usage of Deeply Annotated Information
2. Digital Research Tools (DiRT) Wiki mit umfangreicher Sammlung digitaler Tools für Geisteswissenschaftler_innen: http://dirtdirectory.org
3. Eine von Sears (1997) entwickelte Technik, die eine Kombination aus heuristic evaluations, cognitive walkthroughs und usability walkthroughs ist.
4. Die Software ist unter Apache Licence 2.0 frei verfügbar und kann unter https://github.com/DZielke/laudatio aufgerufen werden.

5. Sears (1997) beschreibt die Vor- und Nachteile verschiedener Evaluationstechniken – auch der heuristischen Evaluation. Er stellt heraus, dass nicht jeder Verstoß gegen eine Heuristik auch immer ein Usability-Problem darstellt. Gerade ungeübte GutachterInnen laufen Gefahr, die Heuristiken auf jeder einzelnen Seite starr umgesetzt sehen zu wollen und verlieren dabei aus dem Blick, ob Verletzungen der Heuristiken wirklich einen Einfluss auf NutzerInnen haben.

Bibliographie

Burghardt, Manuel (2012): "Annotationsergonomie: Design-Empfehlungen Für Linguistische Annotationswerkzeuge", in: *Information - Wissenschaft & Praxis* 63, 5: 300-304.

Department of Corpus Linguistics (Humboldt-University Berlin) / Department of Historical Linguistics (Humboldt-University Berlin) / Computer and Media Service (CMS) (Humboldt-University Berlin) / National Institute for Research in Computer Science and Control (INRIA France) / Berlin School of Library and Information Science (BSLIS) (2011-): *LAUDATIO*. Long-term Access and Usage of Deeply Annotated Information http://www.laudatio-repository.org/repository/ [letzter Zugriff 13. Oktober 2015].

Gibbs, Fred / Owens, Trevor (2012): "Building Better Digital Humanities Tools: Toward Broader Audiences and User-Centered Designs", in: *Digital Humanities Quarterly* 6, 2 http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html [letzter Zugriff 10. Februar 2016].

Nielsen, Jakob (1995): "10 Usability Heuristics for User Interface Design", in: *Nielson Norman Group*. Evidence-Based User Experience Research, Training, and Consulting http://www.nngroup.com/articles/ten-usability-heuristics/ [letzter Zugriff 14. Oktober 2015].

Nielsen, Jakob (2012): "Usability 101: Introduction to Usability", in: *Nielson Norman Group*. Evidence-Based User Experience Research, Training, and Consulting https://www.nngroup.com/articles/usability-101-introduction-to-usability/ [letzter Zugriff 14. Oktober 2015].

Schreibman, Susan / Siemens, Ray / Unsworth, John (2004): *Companion to Digital Humanities* (= Blackwell Companions to Literature and Culture). Oxford: Blackwell Publishing Professional.

Sears, Andrew (1997): "Heuristic Walkthroughs: Finding the Problems Without the Noise", in: *International Journal of Human-Computer Interaction* 9, 3: 213-234.

Lewis, Clayton / Rieman, John (1994): *Task-Centered User Interface Design*. A Practical Introduction. http://grouplab.cpsc.ucalgary.ca/saul/hci_topics/tcsd-book/contents.html [letzter Zugriff 14. Oktober 2015].

Ein Facebook der anderen Art: Digitalisierte Epigraphiken als Quelle der Kulturforschung

Streiter, Oliver
ostreiter@nuk.edu.tw
Staatliche Universität Kaohisung, Taiwan

Gräber, Netzwerke und Kulturforschung

Friedhöfe, Friedhofssektionen, Grabmäler, Grabsteine, Inschriften, Wörter, Symbole, Motive und Formen bilden komplexe Strukturen durch die sich einzelne Personen, Familien, religiöse oder ethnische Gruppen in ihrer Zeit, in ihrem Raum, in ihrer sozialen Umwelt und in ihren metaphysischen Vorstellungen lokalisieren. Durch gemeinsame Referenzen, Ähnlichkeiten oder räumliche Nähe entstehen Netzwerke von Vorfahren, Kameraden, Kollegen, Namensvettern, Leidensgenossen, Mitgläubigern, Mittätern, Selbstmördern oder nur zufällig Zusammengelegten, die durch die Systematik ihrer Bestattungsform oder ihrer Grabinschriften einen Aufschluß über ihre Rolle in einer Zeit oder in einem geographischen oder sozialen Raum zulassen.

Die meisten dieser Netzwerke erzählen jeweils nur wenige Geschichten. Die große Anzahl an möglichen Netzwerken, die Zugehörigkeit einer Person zu vielen Netzwerken sowie der Vergleich ähnlicher Netzwerke aus verschiedenen Zeiten oder Regionen ermöglichen jedoch vielschichtige Einblicke in vergangene Kulturen, Lebensweisen oder Machtstrukturen. Darüber hinaus können diese Netzwerke auch im Rahmen einer abstrakten Betrachtung der Entwicklung solcher sozialen Strukturen herangezogen werden. Aspekte einer solchen theoretischen Betrachtungsweise sind die Entstehung einer Tradition aus einer erfundenen Praxis, die Transformation einer Tradition durch Migration (Goudin et. al. 2011), die Adaption einer Tradition an neue politische oder soziale Kontexte oder die Konsolidierung von Machtstrukturen durch die Neudefinierung von semantischen Feldern (Bourdieu 1979).

Feldforschung, Sozialer Kontext und Datensatz

Voraussetzungen für solche Forschungsarbeiten ist eine flächendeckende und diachronische Dokumentation von Sepulkralkulturen, die bewußt

viele mögliche Gegensätze einschließt, wie zum Beispiel Heldenfriedhöfe, Gräber nationaler Eliten, Heidenfriedhöfe, Selbstmördergräber, Klosterfriedhöfe, Fabriksfriedhöfe, Friedhöfe in dörflichen und städtischen Gemeinden, anonyme Gräber, Grabstätten ethnischer, sprachlicher, religiöser oder andersartiger Minderheiten etc. Eine solche Dokumentation der Unterschiedlichkeit, der Macht, der Unterdrückung und des Vergessens hilft, nationalistische, nationale oder auch gut gemeinte Schulbuchgeschichtsschreibung zu hinterfragen: Wer waren diese Leute, wo kamen sie her und was haben sie gemacht und gedacht? Wie war ihre Stellung in der Gesellschaft und wie hat sich diese im Laufe der Zeit verändert?

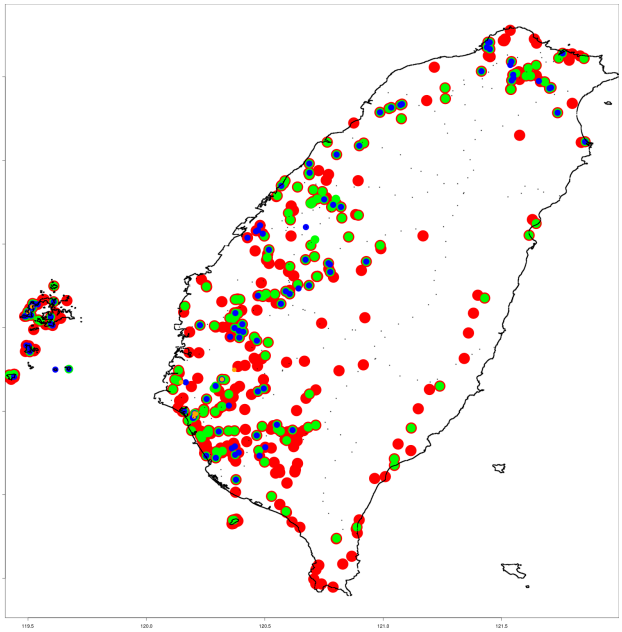


Abb. 1: Auf Taiwan und den vorgelagerten Penghu-Inseln dokumentierte Friedhöfe.

Dies ist der Forschungsansatz den wir seit 2007 in dem Projekt 'ThakBong' für Taiwan und seine umliegenden Inseln verfolgen. Die etwa 400-jährige Geschichte dieser Inseln ist geprägt von Migration und einer sich wiederholenden Abfolge von Kolonialherrschaften. Geschrieben wurden diese Geschichten aus der Sicht der jeweiligen Besatzer. Holländer, Spanier, Handel treibende Haudegen, Mandschuren, Japaner und Chinesen zwangen die Bevölkerung in ihre Kulturen, Religionen, Ideologien und Sprachen, wobei die „eigene“ Kultur meistens stark von der Kultur der vorhergehenden Besatzer geprägt war. Erst nach Aufhebung des Kriegsrechtes 1987 wurde es weiten Schichten der Bevölkerung möglich, aktiv ihre Identität außerhalb der Kultur der letzten Besatzer zu suchen. Dies führte zu einer Rückbesinnung auf Zeiten und Kulturen, die aus heutiger Sicht einen romantischen Rahmen für die moderne Daseinsform bieten. Diese Rückbesinnung reicht von einer vorgeschichtlichen, indigenen Basis, über eine an Japan angelegte Lebensform bis zur Identifizierung als Chinese. Diese unterschiedlichen Selbstbetrachtungen schließen sich nicht aus und können zeitgleich in einer Person oder einer Gruppe in unterschiedlichen Formen des Diskurs wiedergefunden werden.

In den acht Jahren des Projekts wurden in ca. 500 Tagen Feldforschung 700 Friedhöfe mit 59.000 Gräbern durch 212.000 digitale, georeferenzierte Fotos dokumentiert. Fotos werden zu verschachtelten Modellen von Friedhöfen, Gräbern, Grabsteinen und Inschriften in einer relationalen Datenbank zusammengeführt. Diese Modelle werden dann durch Annotationen und Transkriptionen angereichert und klassifiziert. Der daraus resultierende Datensatz wird allmonatlich aktualisiert und der Forschung zur Verfügung gestellt. Begleitet werden diese Daten von einer Batterie an Hilfsfunktionen, geschrieben in der Programmiersprache R, die einen einfachen Zugriff auf Gräber einer speziellen Region, einer speziellen Periode oder einer bestimmten Eigenschaft erleichtert. Einfache Analysen und Grafiken können hiermit in Programmen erstellt werden, die in der Regel zwischen fünf und zehn Zeilen umfassen (Streiter / Morris 2015).

Focus, Ortsnamen und Soziale Identität

Ein zentrales Element unserer Analyse von Inschriften ist der sogenannte Fokus. Der Fokus ist das Element der Inschrift, das durch seine Größe und seine Anordnung auf dem Grabstein hervorspringt und, semantisch betrachtet, nicht die Unterschiedlichkeit der Verstorbenen unterstreicht, sondern deren Einbettung in eine wirkliche oder imaginäre Gemeinschaft. Mögliche Realisierungen des Fokus sind ein Kreuz, eine arabische Sure, eine Swastika, ein Ortsname, ein politisches Symbol oder der Name einer politischen Ära.

Abb. 2: Ein Fokus mit einem Verweis auf die Ming Dynastie (##) auf einem Grabstein auf den Penghu-Inseln von 1669, während in China schon die Qing Dynastie herrscht.

Der Ortsname, der am häufigsten vorkommende Fokus, kann in drei Untergruppen unterteilt werden. Dies sind a) lokale Ortsnamen, die auf einen Ort auf Taiwan verweisen, b) Ortsnamen in China, 'Jiguan' genannt, die den Ausgangspunkt einer Migration aus den südchinesischen Provinzen Guangdong and Fujian benennen und c) Ortsnamen in Nordchina, genannt 'Tanghao', die auf die historischen aber nicht unbedingt heute existierenden oder lokalisierbaren Orte verweisen, in deren Zusammenhang ein chinesischer Familienname erstmals schriftlich erwähnt wurde. Diese Kopplung von Familienname und Ortsnamen wurde seit der Periode der südlichen Song im Buch der Einhundert Familiennamen zusammengefaßt (Theobald 2011). Obwohl die meisten Familien im Prinzip auf jeden dieser drei Typen zugreifen könnten, wird, geschichtlich und regional betrachtet, jeweils einer dieser drei Ortsnamentypen bevorzugt (Streiter / Goudin 2013, 2014).

Abb. 3: Die erste Seite des ältesten jemals gefundenen Baijiaxing (###), geschrieben um 1100. Es zeigt die kanonisierte Beziehung zwischen Familiennamen, z. B. Zhao (#), und deren Tanghao, z. B. Tianshui (##).

Alle drei Untergruppen könnten oberflächlich gesehen sehr leicht als Ausdruck einer Identität interpretiert werden. Der Taiwanische Ortsname stünde demnach für eine lokale Identität, die entweder aus der indigenen Geschichte Taiwans oder aus einer Abkehr von China zu erklären wäre. Das 'Jiguan' unterstrich die Verbundenheit mit Südchina, den gemeinsamen südchinesischen Sprachen und Gebräuchen und der Migration die hauptsächlich in der Qing Periode stattfand. Das 'Tanghao' verwiese, der gleichen Logik folgend, auf die Wiege der chinesischen Kultur in Nordchina.

Aber wie der Fall des 'Tanghao' zeigt, können ideologische Bestrebungen der nationalistischen chinesischen Regierung in Taiwan die Präferenz für das Tanghao besser erklären als es die zirkuläre Interpretation eines Identitätsausdrucks vermag, denn durch das Tanghao verschmelzen die imaginären Ursprungsorte der lokalen Bevölkerung, die über den Familiennamen

identifiziert werden, mit den Herkunftsregionen der Chinesen, die mit Chiang Kai-shek nach dem verlorenen Bürgerkrieg 1949 als Besatzer nach Taiwan kamen. "Obwohl der Weg ein anderer war, ist der Ursprung der gleiche." Dies ist ein Mantra, das bis heute von der regierenden nationalistischen chinesischen Partei immer wieder in unterschiedlichen Formen bemüht wird (Office of the President, Republic of China 2012).

Ebenso verschwand das 'Jiguan', die Referenz auf die Migration von Südchina, von den Grabsteinen Taiwans in dem Maße, wie die Kolonialmacht Japan bestrebt war, die Bürger Taiwans zu seinen Bürgern zu machen. Die Hoffnung, die lokale Bevölkerung an dem Krieg der Besatzer teilnehmen zu sehen, dürfte bei diesem Bestreben ebenso eine Rolle gespielt haben, wie wenige Jahre später, als eine Chinesische Identität lokale Taiwanesen motivieren sollte, China von der Kommunistischen Partei zurückzuerobern.

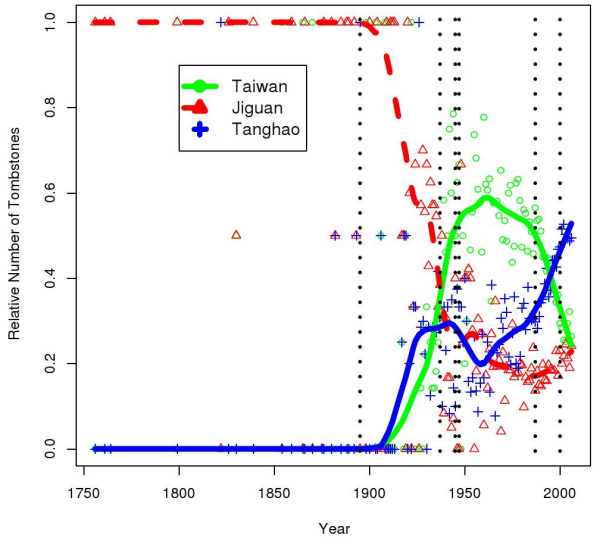


Abb. 4: Ortsnamentypen im Wandel: Taiwan unter den Qing (bis 1895), unter dem Japanischen Kaiserreich (bis 1945) und unter der Chinesischen Republik.

Paradigmen, Macht und Mediation

Ein Verständnis der Bedeutung der Ortsnamenwahl kann also nicht über eine oberflächliche Kongruenz von Ausdrucksform und möglichen Inhalten ermittelt werden. Vielmehr gilt es zu ermitteln, welche Ausdrucksformen überhaupt in welcher Zeit und in welcher Region mit welchen syntaktischen Funktionen zur Verfügung standen und welche Ausdrucksformen in welcher Weise den jeweiligen Besatzern entgegenkamen.

Unsere Analysen unterstreichen in diesem Zusammenhang die Rolle von professionellen Steinmetzen, im Gegensatz zu handwerklich geschickten, aber semantisch ungeschulten Kräften, die, involviert in den Bau des Grabes, gleich die Beschaffung und Beschriftung des Grabsteins mit übernahmen. Die Professionalität, mit der dieser den Grabstein bearbeitete, läßt sich mathematisch einfach ermitteln, da, für das ungeübte Auge unsichtbar, der Grabstein in seiner Höhe, Breite und in der Anzahl der Schriftzeichen vom Fachmann durch glücksbringende Zahlen strukturiert wird, die durch Zufall vom Laien kaum reproduziert werden können. In unserer Analyse zeigen wir, wie die Inschriften der ungeschulten Handwerker, spontan und unsystematisch, selten zu einer langfristigen Entwicklung führten, während die Präferenzen der Steinmetze strategischen Entscheidungen im Rahmen der Traditionen einer Steinmetzschule unterlagen. Solche strategischen Entscheidungen wurden, wie unsere Analysen zeigen, zu Zeiten einer Krise getroffen, zum Beispiel bei der Eroberung durch ein anderes Regime, und wurden dann weitgehend beibehalten bis eine neue Krise zu neuen Entwicklungen führte.

Hierdurch ergibt sich eine Umkehrung der Beziehung von Ausdrucksform und Inhalt, zumindest in den Zeiten, im wesentlichen nach 1750 und in den urbanen Zentren, in denen der Einfluß von professionellen Steinmetzen nachgewiesen werden kann. Hier entwickelte sich die soziale Identität wahrscheinlich im Zusammenspiel mit einer historischen Interpretation der professionellen Inschriften. Aufgabe der Wissenschaft ist es, hier Elemente eines Diskurses in geographischer und zeitlicher Nähe nachzuweisen, die eine solche Annäherung von Inschriften und Identitäten ermöglichten. Es ist dieser integrierte Ansatz der Dokumentation und Analyse von Diskurs und materieller Kultur, die wir als Digitale Anthropologie bezeichnen.

Bibliographie

Bourdieu, Pierre (1979): *La distinction: Critique sociale du jugement* (= Collection Le sens Commun). Paris: Éditions de Minuit.

Goudin, Yoann / Streiter, Oliver / Huang, Chun (Jimmy) / Mei-Fang, Lin Ann (2011): " Digital Anthropology and the Renewal of Waishengren Studies: From Digitized Tombs to Identity Claims ", in: *African and Asian Studies* 15, 2: 21–45.

Office of the President, Republic of China (2012): *Der Präsident besucht den "Miaoli Hakka Kulturpark" und wohnt der "Nacht der Hakka Welthelden" bei* [Übersetzung aus dem Chinesischen] http://www.president.gov.tw/Default.aspx?tabid=131&itemid=28020 [letzter Zugriff 15. Oktober 2015].

Streiter, Oliver / Goudin, Yoann (2013): "The Tanghao on the Tombstones of Taiwan and Penghu: The Statal Recuperation of Tactics for the Creation of a National Space", in: *Oriental Archive* 81: 459-494.

Streiter, Oliver / Goudin, Yoann (2014): "Extracting Family Genealogis from Taiwan's Tombstones for a Study of Historical Chang e s in Tombstone Inscriptions", in: *International Journal of Humanities and Arts Computing* 8, 1: 49–83.

Streiter, Oliver / Morris, James X. (2015): "Researching Taiwan's Gravesites with ThakBong and R", in: *PNC 2015 Annual Conference and Joint Meetings*. Macau.

Theobald, Ulrich (2011): "Chinese Literature: Baijiaxing ### 'The Hundred Family Names'" in: *CHINAKNOWLEDGE*. A universal guide for China studies http://www.chinaknowledge.de/Literature/Classics/baijiaxing.html [letzter Zugriff 15. Oktober 2015].

"Alles ist Wechselwirkung" – auch in den Digital Humanities: Von 'D' nach 'H' und zurück durch Humboldts Kosmos-Vorträge (1827/28)

Thomas, Christian
christian.thomas.1@staff.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland; Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Projekthintergrund: A. v. Humboldts „Kosmos-Vorträge" (1827/28)

Im Projekt Hidden Kosmos: Reconstructing A. v. Humboldt's »Kosmos-Lectures« , das an der Berliner Humboldt-Universität (HU) aus Mitteln der Exzellenzinitiative über den Zeitraum von zwei Jahren (Juni 2014–Mai 2016) gefördert wird, werden sämtliche derzeit bekannte Nachschriften von Besuchern der weltberühmten *Vorlesungen über physikalische Geographie* Alexander von Humboldts erschlossen und als vernetztes Forschungskorpus ediert.

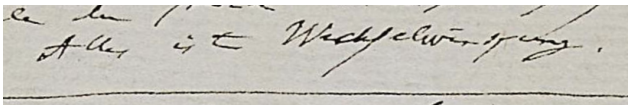


Abb. 1: „Alles ist Wechselwirkung." Humboldt [1803/04]: 27r, http://resolver.staatsbibliothek-berlin.de/SBB0001527C00000000 [1]

Humboldts heute so genannten Kosmos-Vorträge fanden im Wintersemester 1827 / 28 in zwei unabhängig voneinander verlaufenden Zyklen statt: Er absolvierte insgesamt 62 Vortragsstunden vor etwa 400 Studierenden und Lehrbefugten in der Berliner Universität und parallel dazu 16 Vorträge vor einer bis dahin unerreichten Zahl von etwa 1000 Zuhörern im großen Saal der Berliner Singakademie. Humboldt verband in diesen Vorträgen seine eigenen Forschungen, dabei aus dem reichen Erfahrungsschatz seiner fünfjährigen Amerikareise schöpfend, mit dem damals aktuellen Erkenntnisstand auf faktisch jedem Gebiet der aufstrebenden Naturwissenschaften. Beide Zyklen unterscheiden sich wesentlich hinsichtlich der Abfolge der besprochenen Themen und – aufgrund ihres sehr ungleichen Umfangs – auch im Hinblick auf die dabei jeweils erreichte Tiefe und Ausführlichkeit der Darstellung. Gemeinsam ist beiden Vortragsreihen jedoch der Anspruch, einen in sich abgeschlossenen Überblick zu geben, d. h. die astronomischen und tellurischen Phänomene, die Gestalt der Erdoberfläche und das organische Leben auf ihr, die kulturelle Entwicklung der Menschheit in den für Humboldt allgegenwärtigen „Wechselwirkungen" (Abbildung 1) darzustellen.

Forschungsstand zu den Kosmos-Vorträgen und Ziele des Projekts Hidden Kosmos

Die Kosmos-Vorträge können nach wie vor als ‚blinder Fleck' der Humboldt-Forschung gelten (vgl. Erdmann / Thomas 2014: 35f.). Zum einen wohl deshalb, weil aufgrund einer (nachweislich falschen) Behauptung Humboldts im Kosmos (Humboldt 1845: X) die Manuskripte des Vortragenden als nicht existent galten, zum anderen weil auch der größte Teil der Nachschriften seiner HörerInnen de facto unbekannt blieb. Bis zur laufenden Veröffentlichung mehrerer Manuskripte durch das Hidden Kosmos-Projekt waren nur zwei solcher Nachschriften publiziert worden – beide jedoch in kaum wissenschaftstauglichen Editionen. Die übrigen zehn bisher bekannten Nachschriften lagerten unberührt in verschiedenen Bibliotheken in Deutschland und Polen bzw. in Privatbesitz.

Die digitale Edition dieser unikalen Manuskripte schafft überhaupt erst eine solide Materialbasis, um die intensive Erforschung der Vortragsreihen zu ermöglichen. [2] Das Projekt Hidden Kosmos arbeitet dabei eng mit dem Deutschen Textarchiv (DTA) der BBAW zusammen, wo die Nachschriften im Kontext des weltweit umfangreichsten digitalen ‚Alexander-von-Humboldt-Korpus' (s. Thomas 2015) veröffentlicht werden. Die weitere Dissemination und langfristige Bereitstellung der Daten erfolgt über das web- und zentrenbasierte Infrastrukturprojekt CLARIN-D . [3]

Derzeit (15.10.2015) stehen fünf Nachschriften mit mehr als 2 100 handschriftlichen Seiten im DTA bzw. über CLARIN-D bereit; bis zur DHd 2016 werden zehn Nachschriften mit 3 760 Seiten zur Verfügung stehen.

Abb. 2: DTA-Präsentation der Nachschrift eines anonym gebliebenen Zuhörers der Kosmos-Vorträge an der Berliner Universität: [N. N.]: Die physikalische Geographie von Herrn Alexander v. Humboldt, vorgetragen im Semestre 1827 / 28. [Berlin], [1827/28].

Gliederung des Vortrags

Überblick über die edierten Materialien und erste Forschungsergebnisse

Auf der DHd 2016 wird zunächst ein konzentrierter Überblick über die bis dahin publizierten Materialien gegeben und werden erste, auf dieser Grundlage gewonnene Forschungsergebnisse vorgestellt. Der Fokus wird dabei im Sinne des Konferenzthemas auf der **Modellierung** der Daten und deren Annotation in TEI-XML, der **Vernetzung** der Dokumente sowie verschiedenen **Visualisierungen** der annotierten Forschungsdaten liegen. Im Vergleich der tief strukturierten und annotierten Online-Volltexte mit früheren Printeditionen zweier Nachschriften werden die Vorzüge digitaler Editionen sichtbar: Auf der Makroebene, d. h. hier: mit Blick auf die Gesamtheit der vielstimmigen Überlieferung, liegen diese Vorzüge zum einen in deren Perfektibilität und permanenten Erweiterbarkeit, zum anderen in der Vernetzung *aller* bekannten Nachschriften untereinander und mit weiteren elektronischen Ressourcen. In der Mikroperspektive, d. h. hier: mit Blick auf die einzelne, jeweils unikale Quelle ermöglicht die digitale Edition eine überlieferungsadäquate Repräsentation jedes einzelnen Manuskripts und der Besonderheiten seiner handschriftlichen Verfasstheit. Die digitale Edition erreicht dabei eine tiefere Granularität und eine größere Flexibilität der Nutzungsmöglichkeiten als die Print-Edition.

Anschließend sollen die Auswirkungen diskutiert werden, die der Einsatz von Methoden und Verfahren aus dem Bereich der ‚Digital Humanities' sowohl auf die Rezeptions- als auch auf die Produktionsseite einer (Manuskript-)Edition haben. Ganz im oben zitierten Sinne Humboldts sollen diese als ‚Wechselwirkungen' zwischen Projektdesign und Produzent bzw. Rezipient erfasst werden.

Wechselwirkung zwischen Projektdesign und Nutzerperspektive

Anschließend an die Präsentation im vergangenen Jahr auf der DHd 2015 in Graz (Thomas 2014 / 15) kamen zwei grundsätzliche Fragen aus dem Publikum, die vor dem Hintergrund des zum Zeitpunkt der DHd 2016 fast abgeschlossenen Projekts beantwortet werden sollen. Die erste Frage stellte sich aus der Perspektive der/s Rezipientin/en, der/die – so die an den Herausgeber der Edition herangetragene Forderung – an die Quellen ‚herangeführt' werden müsse. Denn im Unterschied zum überkommenen Prinzip der ‚Leithandschrift' bzw. der Konstruktion eines ‚idealen' Textes durch den Editor, das den Beschränkungen gedruckter Editionen im ‚typographischen Paradigma' (Sahle 2013: 88) geschuldet ist, stehen im DH-Projekt Hidden Kosmos zehn parallele, teils einander ergänzenden, teils miteinander konkurrierende Nachschriften prinzipiell gleichwertig nebeneinander. Die komplexe und vielschichtige Überlieferungslage soll für den / die Nutzer_in der edierten Texte transparent werden, anstatt sie wie im typographischen Paradigma einzuebnen. Der/die NutzerIn soll ermutigt und befähigt werden, sich diese Komplexität und Vielschichtigkeit durch eine parallele Lektüre verschiedener Quellen und durch die angebotenen, explorativen Zugänge zu erschließen. [4]

Um dieses Ziel zu erreichen, wurden im Hidden Kosmos-Projekt mehrere Dokument-übergreifende Zugänge geschaffen. In der vergleichsweise kurzen zweijährigen Projektlaufzeit mit sehr begrenzten personellen Ressourcen konnte diese Aufgabe überhaupt nur durch den konsequenten Einsatz digitaler Methoden bewältigt werden. Da mir das Problem begrenzter Zeit- und Personalressourcen typisch für den heutigen Forschungsalltag erscheint, soll auf die dabei verwendeten Ansätze näher eingegangen werden.

Automatisch erstellte, korpusübergreifende Zugänge als Orientierungshilfe

Anhand der -Ebenen der TEI-strukturierten Volltexte wurden je eine thematische Gliederung für den Universitätszyklus und für die Singakademie-Vorträge extrahiert und diese wiederum mit der chronologischen Gliederung nach Vortragsstunden des jeweiligen Zyklus kombiniert. Diese abstrahierte, alle Nachschriften verbindende Orientierungshilfe konnte in dieser Vollständigkeit nur durch eine Kombination aller vorliegenden Quellen erstellt werden, da die jeweiligen Schreiber ihre Hefte entweder nur grob thematisch strukturierten oder nur die Daten der jeweiligen Vortragsstunde notierten. Das Ergebnis ist eine sehr viel detaillierte Übersicht über die Kurse als die von Humboldt selbst überlieferte (Abbildung 3).

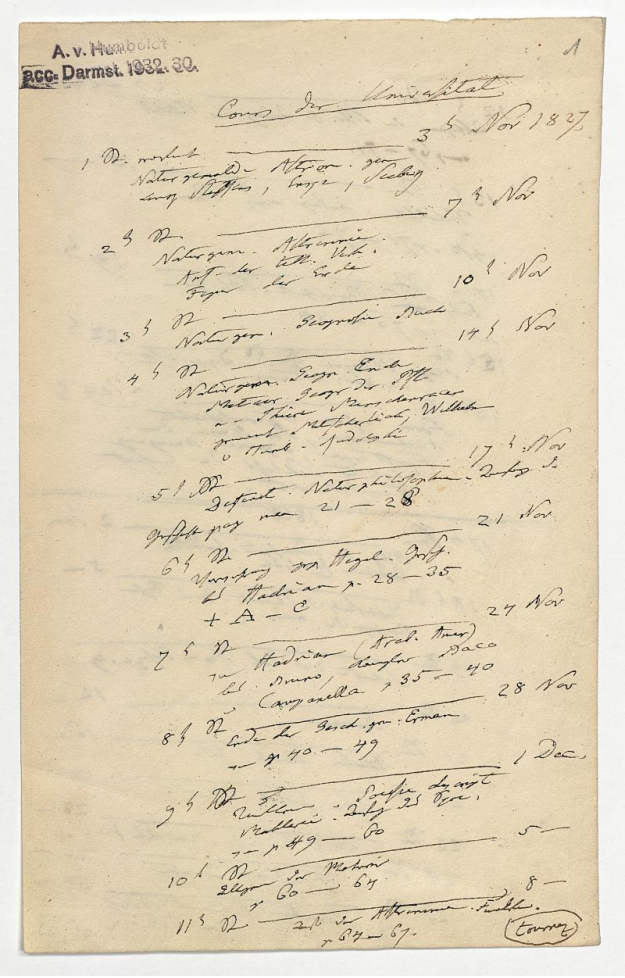


Abb. 3: A. v. Humboldt: Vorlesungsverzeichnis für die Berliner Universität, Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Nachl. Alexander von Humboldt, gr. Kasten 8, Nr. 5a, Bl. 1r (http://resolver.staatsbibliothek-berlin.de/SBB0001676C00000000)

NutzerInnen der Edition können sich anhand der extrapolierten Gliederung leicht einen Überblick über die beeindruckende Themenfülle der Vortragsreihen verschaffen und zwischen dem Einstieg in ein beliebiges Thema oder eine Vortragsstunde wählen. Zudem wird die unterschiedliche Abfolge der Themen in beiden Vortragszyklen anschaulich und ein gezielter Einstieg in deren vergleichende Lektüre ermöglicht. Zugleich bilden die Gliederungen einen Anknüpfungspunkt für die weitere Kontextualisierung der Kosmos-Vorträge im übrigen Humboldt'schen Œuvre. Beispielseise übernahm Humboldt für seinen monumentalen *Kosmos* (1845–62) im Wesentlichen die Anordnung der Singakademie-Vorträge (vgl. Erdmann / Thomas 2014: 37), wodurch sich die Vorträge nun auf Grundlage der aus den Nachschriften extrapolierten Gliederung mit den entsprechenden -Ebenen der im DTA verfügbaren XML-Volltexte des *Kosmos* verknüpfen lassen.

Ebenso automatisch wurde aus den annotierten Daten, die mit dem TEI-Element <persName> und einem @ref-Attribut mit Link auf verfügbare Normdaten versehen wurden, ein übergreifendes Personenverzeichnis extrahiert, das derzeit mehr als 2000 Einträge enthält. Durch die Verknüpfung mit Normdaten aus der GND, VIAF o.Ä. werden die zwischen den Nachschriften teilweise erheblich voneinander abweichenden Vorlageformen der von Humboldt erwähnten Personen vereinheitlicht. Jeder Eintrag im Gesamtregister führt per Klick zum Kontext derjenigen Nachschrift, in der der gewählte Personenname getaggt wurde. Eine Verbindung des Personenregisters mit weiterführenden Informationsangeboten z. B. über eine BEACON-Datei bietet NutzerInnen – ohne nennenswerten Mehraufwand für das Projekt – direkten Zugang zu weiteren Informationen. Neben seiner Funktion als übergreifende Orientierungshilfe eignet sich das Personenregister auch als Impulsgeber für die Beforschung der Nachschriften: Fehlt beispielsweise ein Personenname in einem Dokument, der in einem anderen an der entsprechenden Stelle referenziert wurde, erlaubt dies schon erste Rückschlüsse auf die Zuverlässigkeit und Vollständigkeit der Nachschrift.

Weitere dokumentübergreifende Ordnungshilfen listen beispielsweise die von Humboldt im Zusammenhang mit naturwissenschaftlichen Untersuchungen eingesetzten Instrumente, die von ihm erwähnten Himmelskörper sowie die mineralischen und chemischen Elemente auf, die im Laufe der Vorträge eine Rolle spielten. Diese und weitere ‚Inventarlisten' konnten direkt aus den Volltexten extrahiert werden, dank der Verbindung mit der computerlinguistischen Erschließung der Volltexte im Deutschen Textarchiv. Die im DTA implementierte Kombination der linguistischen Suchmaschine DDC mit dem Wortnetz GermaNet (Henrich / Hinrichs 2010; Hamp / Feldweg 1997) ermöglicht es beispielsweise, alle Nachschriften gezielt nach denjenigen Begriffen zu durchsuchen, die in GermaNet als „Element" bzw. übergeordnet als „Grundstoff; Urstoff" klassifiziert wurden (Abbildung 4). Eine umfassende Liste mit knapp 3000 Treffern und mehr als 300 verschiedenen Mineralien, Gesteinen und Substanzen ist das Ergebnis,

das der/dem NutzerIn unmittelbar zur zielgerichteten Navigation angeboten werden kann.

27:	[dta:nn_mgermqu2345_1827:156]	Der Physiker Ritter meinte es seien viele	Meteorsteine	gefallen wenn das Nordlicht erschienen doch dies und...
28:	[dta:nn_oktavgfeo79_1828:338]	Wollaston hat gefunden, daß das	Meerwaller	außerdem auch salzsaures und schwefelsaures Kali, jedoch...
29:	[dta:nn_n0171u1_1828:15]	Dieſes conſtante Zuſammenſein einzelner	Foſſilien	|: Stoffe :| bildet Gebirgsarten,...
30:	[dta:parthey_mgermqu1711_1828:549]	Die Trokkenheit der	Luft	nimt auf den Bergen nach oben hin zu...
31:	[dta:nn_mgermqu2345_1827:10]	... 3½ Jahr zurückkehrt und nicht weiter geht als	Merkur	, nicht so weit als Jupiter.
32:	[dta:parthey_mgermqu1711_1828:492]	... weiter er geht, er desto mehr an	Wasser	verliert.
33:	[dta:nn_n0171u1_1828:328]	... auch Flammen aufſteigen geſehen; es ſcheint nicht	Waſſerſtoffgas	zu ſein, ſondern andern chemiſche Subſtanzen.
34:	[dta:nn_oktavgfeo79_1828:366]	In Verbindung mit	Sauerſtoff	, als Waſſergas, ſteigt wie bekanntlich jeden...
35:	[dta:nn_mgermqu2345_1827:165]	Der Geiser und Raiko in Island enthalten 3/10	Kieſelerde	und kohlensaures Natron, ja sogar eine vegetabilisch-animaliſche...
36:	[dta:nn_oktavgfeo79_1828:429]	.... ſo können ſich auch nicht Schnee und	Eis	darauf lagern, was eben die Nordweſtküſte der...
37:	[dta:nn_oktavgfeo79_1828:242]	... wo nach der Auflage der Einwohner ſich periodiſch	Feuer	zeigen ſoll.
38:	[dta:nn_n0171u1_1828:345]	Von dem	Baſalt	hat man das Hervordringen aus der Tiefe an...
39:	[dta:parthey_mgermqu1711_1828:589]	... wird mehr Sonnenſtralen abſorbiren, und daher die	Luft	erwärmen.
40:	[dta:nn_mgermqu2345_1827:261]	... durch ſehr kalten Winter, weil Schnee und	Eis	ſich darauf lagern können und ungeheuer viel Kälte...
41:	[dta:nn_mgermqu2345_1827:16]	Ueberall iſt ein ſolcher Halbſchatten von	Marmor	und körnigem Kalkſtein um den Granit, wo...
42:	[dta:parthey_mgermqu1711_1828:595]	... man einigermaassen warm gekleidet war, und kein	Wind	wehte; beim Winde aber wurde eine Kälte...
43:	[dta:nn_oktavgfeo79_1828:374]	Man könnte das	Salzwaſſer	als die Urſache davon anſehen, allein bei...
44:	[dta:parthey_mgermqu1711_1828:440]		Zechſtein	oder Alpenkalk, waſſerhaltiger Gyps; Steinſalz eingelagert...
45:	[dta:nn_mgermqu2124_1827:39]	– Muſchel und Jurakalk, Salzthon,	Steinſalz	, Sandſtein, und zuletzt die Kreide gehören...

Abb. 4: DDC-Suche im Deutschen Textarchiv nach Begriffen aus dem GermaNet-Synset „Grundstoff; Urstoff“

Wechselwirkung zwischen Projektdesign und Produzentenperspektive

Ebenfalls im Anschluss an die DHd-2015-Präsentation in Graz – und seitdem des Öfteren wiederholt – wurde die manche/n Geisteswissenschaftler/in offenbar beunruhigende Frage gestellt, worin denn noch die Aufgaben des Herausgebers einer wissenschaftlichen (Manuskript-)Edition bestünden, wenn wie im Projekt Hidden Kosmos grundständige Arbeiten wie das Transkribieren und Annotieren (fast) aller Textzeugen an einen Dienstleister ausgelagert werden, wenn bestehende Datenmodelle und Infrastrukturen einfach mit- oder nachgenutzt und wenn immer mehr Arbeitsschritte automatisiert werden können.

Auf der DHd 2016 möchte ich mit Bezug auf die oben skizzierten Arbeitsschritte einige aus meiner Sicht notwendige Änderungen in der ‚Job Description‘ des Editors bzw. allgemeiner: der/des Geisteswissenschaftlerin/s in einem DH-Projekt anregen. Diese liegen m.E. nach wie vor im traditionellen Bereich der hermeneutisch-interpretierenden Quellenkritik, aber wesentlich mehr noch in den Bereichen der Datenkuration, tieferen Annotation, der Analyse und ständigen Wiederausführung und Optimierung automatisierter Prozesse. Und selbstverständlich bleibt bei jedem computergestützten Prozess der Datenanalyse ein nicht-automatisierbarer ‚Rest‘, der interpretiert werden muss und der durch weitere, wiederum computergestützte Optimierung der Datenbasis oder durch eine Rekonfiguration des automatisierten Prozesses verringert oder zumindest verändert werden kann. Dabei wird dieser Rest wohl auch durch die Optimierung der Daten und die fortschreitende Verbesserung der Tools und Services nicht verschwinden, möglicherweise aber immer interessanter werden.

Notes

1. Alle URLs in diesem Text abgerufen am 15. Oktober 2015.
2. Humboldts eigenhändige Manuskripte liegen, soweit sie sich überhaupt erhalten haben, verstreut und z. T. nicht sicher identifizierbar in dessen Nachlass in der Staatsbibliothek zu Berlin und der Jagiellonen-Bibliothek in Krakau. Bis voraussichtlich Ende 2016 werden beide Nachlassteile komplett digitalisiert sein. In einem die Hidden Kosmos-Idee weiterführenden Anschlussprojekt sollen dann die ursprünglichen Vortragsmanuskripte mit Hilfe der Hörernachschriften identifiziert und in die laufende Edition integriert werden.
3. Für eine detailliertere Darstellung der Kooperation zwischen Hidden Kosmos, DTA und CLARIN-D s. Thomas 2014/15.
4. Siehe auch dazu Sahle (2013: 107), der zu den oben skizzierten Auswirkungen digitaler Editionsformen festhält: „Die Visualität und Materialität der Überlieferung kann besser sichtbar gemacht werden, die Aufforderung zur Konstruktion der *einen* autoritativen editorischen Fassung, die alle anderen Fassungen nahezu unsichtbar macht, wird schwächer.“ (Hervorhebung CT)

Bibliographie

Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) (2007-):*DTA*. Deutsches Textarchiv http://www.deutschestextarchiv.de/ [letzter Zugriff 15. Oktober 2015].

CLARIN-D http://clarin-d.net [letzter Zugriff 15. Oktober 2015].

Erdmann, Dominik / Thomas, Christian (2014): "»… zu den wunderlichsten Schlangen der Gelehrsamkeit zusammengegliedert«. Neue Materialien zu den ›Kosmos-Vorträgen‹ Alexander von Humboldts, nebst Vorüberlegungen zu deren digitaler Edition". In: *HiN – Humboldt im Netz*. Internationale Zeitschrift für Humboldt-Studien (Potsdam – Berlin) XV, 28: 34-45 http://hin-online.de/hin28/erdmann-thomas.htm [letzter Zugriff 15. Oktober 2015].

Hamp, Birgit / Feldweg, Helmut (1997): "GermaNet – a Lexical-Semantic Net for German." In: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid 9-15.

Henrich, Verena / Hinrichs, Erhard (2010): "GernEdiT – The GermaNet Editing Tool". In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta 2228-2235 http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf [letzter Zugriff 15. Oktober 2015].

Humboldt, Alexander von (1845): *Kosmos*. Entwurf einer physischen Weltbeschreibung, 1. Stuttgart u. a.:

Deutsches Textarchiv http://www.deutschestextarchiv.de/humboldt_kosmos01_1845 [letzter Zugriff 15. Oktober 2015].

Humboldt, Alexander von ([1803/04]): *Tagebücher der Amerikanischen Reise*: IX: Varia. Obs. Astron. de Mexico a Guanaxuato, Torullo, Tiluca, Veracruz, Cuba. Voy. De la Havana à Philadelphia. Geologie de Guanaxato, Volcans de Torullo et de Toluca. Voayge de Veracruz à la Havana et de la Havana à Philadelphia. Torulla. Berlin: Staatsbibliothek zu Berlin – Preußischer Kulturbesitz http://resolver.staatsbibliothek-berlin.de/SBB0001527C00000000 [letzter Zugriff 15. Oktober 2015].

Lehrstuhl für Kulturtechniken und Wissensgeschichte (2014-): *Hidden Kosmos*. Reconstructing Alexander von Humboldt's »Kosmos-Lectures« https://www.culture.hu-berlin.de/de/forschung/projekte/hidden-kosmos/ [letzter Zugriff 15. Oktober 2015].

N.N. (1827): *Die physikalische Geographie von Herrn Alexander v. Humboldt, vorgetragen im Semestre 1827/28*. Berlin: Deutsches Textarchiv: http://www.deutschestextarchiv.de/nn_oktavgfeo79_1828/7 [letzter Zugriff 15. Oktober 2015].

Sahle, Patrick (2013): *Digitale Editionsformen*. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. 2: Befunde, Theorie und Methodik. Norderstedt: Books on Demand.

Thomas, Christian (2014/15): "Hidden Kosmos – Humboldts ‚Kosmos-Vorträge' als Probe der Digital Humanities", in: DHd-Jahrestagung 2015 „Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation", 23.-27.2.2015, Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities an der Universität Graz https://www.culture.hu-berlin.de/de/forschung/projekte/hidden-kosmos/media/c-thomas-dhd-graz-paper-hidden-kosmos-20150126.pdf [letzter Zugriff 15. Oktober 2015].

Thomas, Christian (28.09.2015): "99 unselbständige Schriften Humboldts als Volltext im Deutschen Textarchiv verfügbar", in: *avhumboldt.de*. Alexander von Humboldt Informationen online http://www.avhumboldt.de/?p=10922 [letzter Zugriff 15. Oktober 2015].

Dramen als small worlds? Netzwerkdaten zur Geschichte und Typologie deutschsprachiger Dramen 1730-1930

Trilcke, Peer
trilcke@phil.uni-goettingen.de
Georg-August-Universität Göttingen, Deutschland

Fischer, Frank
frank.fischer@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Göbel, Mathias
goebel@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Kampkaspar, Dario
kampkaspar@hab.de
Herzog August Bibliothek Wolfenbüttel

Ansatz

Neben dem ›klassischen‹ strukturalistischen Paradigma, das sich wesentlich an Theoremen der Linguistik orientiert (u. a. Lotman 1972; Titzmann 1977), gibt es in der Literaturwissenschaft bereits seit Jahrzehnten Ansätze zu einer Strukturanalyse, die sich auf die empirische Soziologie – insbesondere auf die *Social Network Analysis* – bezieht und Struktur entsprechend nicht über basale semantische Relationen (etwa als Opposition oder Äquivalenz) definiert, sondern über soziale Interaktionen (Marcus 1973; Stiller et al. 2003; de Nooy 2006; Stiller / Hudson 2005; Elson et al. 2010; Agarwal et al. 2012). Im Kontext der Digital Humanities haben diese Ansätze zu einer literaturwissenschaftlichen Netzwerkanalyse (Trilcke 2013) in den letzten Jahren eine neue Dynamik gewonnen (Moretti 2011; Rydberg-Cox 2011; Park et al. 2013). Aus literaturwissenschaftlicher Sicht versprechen diese Analyseverfahren dabei auf umfangreichen Korpora basierende, von quantitativen Daten gestützte Erkenntnisse über die Literaturgeschichte wie auch über die generischen Eigenarten literarischer Texte. Im Projekt *dlina. Digital Literary Network Analysis* haben wir einen Workflow zur Extraktion, Analyse und Visualisierung von Netzwerkdaten aus dramatischen Texten mit rudimentärer TEI-Auszeichnung entwickelt (Fischer et al. 2015). Der hier projektierte Vortrag wird Ergebnisse der netzwerkanalytischen Auswertung dieser Daten präsentieren und vor dem Hintergrund etablierter fachwissenschaftlicher Fragestellungen diskutieren.

Datenerhebung und -analyse

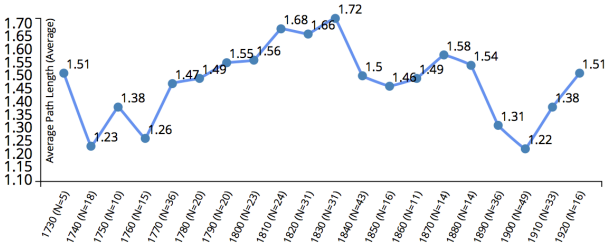
Unser derzeitiges Korpus umfasst 465 deutschsprachige Dramen (Zeitraum 1730 bis 1930), die aus dem Textgrid Repository extrahiert wurden. Die für die Netzwerkanalyse relevanten Strukturdaten dieser Dramen (Segmentierung, Figurenidentifikation)

wurden in einem regelbasierten Prozess händisch ediert, um OCR- und TEI-Tagging-Fehler zu beheben sowie solchen ›Eigenarten‹ der literarischen Texte zu begegnen, die die Analyseergebnisse verfälschen würden (u. a. unterschiedliche Bezeichnungen identischer Figuren; Bezeichnung von Figurengruppen mit unbestimmten Numeralien wie ›beide‹ oder ›alle‹; etc.). Die edierten Strukturdaten liegen in einem eigens entwickelten Datenformat, dem dlina-Format, in Form von XML-Dateien vor. Die Visualisierung der Netzwerke und die Berechnung netzwerkanalytischer Werte erfolgt – mittels Python- und D3-Skripten – automatisiert auf Basis der in den dlina-Dateien gespeicherten Strukturdaten. Neben Graphen und basalen Werten, die die Netzwerke global beschreiben (Network Size, Density, Average Degree, Average Path Length), werden dabei auch Zentralitätswerte für sämtliche Figuren eines Dramas erhoben (u. a. Degree, Average Distance, Closeness Centrality, Betweenness Centrality). Die Implemtierung weiterer Berechnungsrountinen (u. a. Clustering Coefficient, logarithmierte Degree Distribution-Tabellen) ist für den Winter 2015/16 vorgesehen. Sämtliche Daten und Visualisierungen werden frei verfügbar im Netz publiziert (https://github.com/dlina und https://dlina.github.io/linas/).

Literaturwissenschaftliche Auswertung 1: Dramengeschichte

Die diachrone Erstreckung unseres Dramenkorpus über ca. 200 Jahre deutscher Literaturgeschichte macht es möglich, größere Entwicklungen im Bereich der strukturellen Komposition von dramatischen Texten zu beobachten (erste Überlegungen dazu haben wir in einem Blogpost skizziert: https://dlina.github.io/200-Years-of-Literary-Network-Data/). Neben Werten, die sich auf die Gesamtnetzwerke der einzelnen Dramen beziehen (u. a. Network Size, Density, Average Degree; s. exemplarisch zur Average Path Length, Abbildung 1), werden dabei auch figurenbezogene Werte, v.a. Zentralitätsmaße, einbezogen, die Aufschluss etwa über die Streuung des Personals eines Dramas bzw. dessen Zusammensetzung aus ›zentralen‹ und weniger ›zentralen‹ Figuren gibt. Auf Grundlage dieser Werte sollen im Vortrag einige globale Thesen der Literaturgeschichte diskutiert werden. So werden wir *erstens* diskutieren, inwieweit sich anhand der netzwerkanalytischen Werte eine Ausdifferenzierung der strukturellen Komposition von dramatischen Texten am Ende des 18. Jahrhunderts beobachten lässt: Eine solche Ausdifferenzierung wäre angesichts des Nebeinanders von ›geschlossenen‹, in der Tradition der Französischen Klassik stehenden Dramen und ›offenen‹ Dramen, die sich u. a. an der Drramatik Shakespeares orientieren, zu erwarten. *Zweitens* werden wir einige geläufige literaturwissenschaftliche Periodisierungshypothesen testen (u. a. aus dem Strukturalismus und der Sozialgeschichte); gefragt werden soll hier, inwieweit die Entwicklung der netzwerkanalytischen Werte mit den von der Forschung vorgeschlagenen Periodisierungen korreliert.

Abb. 1: Average Path Length (Mean; nach Dekaden)

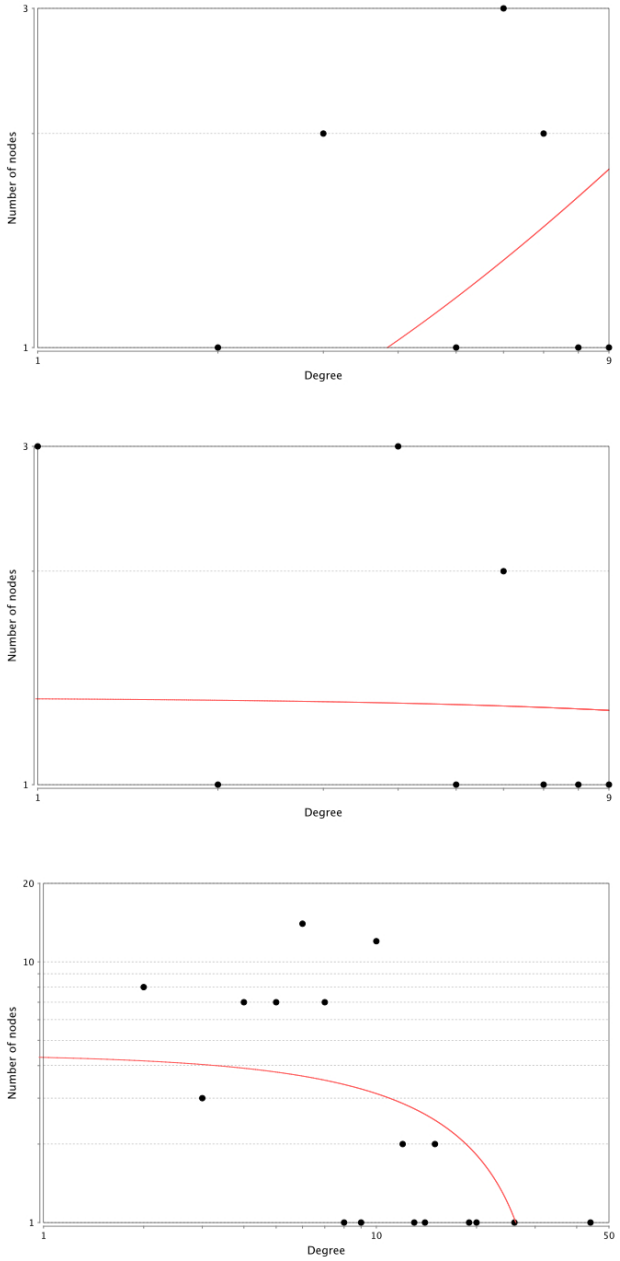


Literaturwissenschaftliche Auswertung 2: Dramentypen

Die von uns bisher erhobenen Werte zeigen, dass Dramen in dem untersuchten Zeitraum auf sehr unterschiedliche Weise strukturiert wurden. In der ›traditionellen‹ Literaturwissenschaft wurden für solche unterschiedlichen ›Bauformen‹ diverse Typologien entwickelt, in der Germanistik am bekanntesten ist Volker Klotz' Unterscheidung in eine ›offene‹ und eine ›geschlossen‹ Dramenform (Klotz 1960). Diesen typologischen Impuls wollen wir aufgreifen und einen Vorschlag unterbreiten, wie sich mittels netzwerkanalytischer Daten bestimmte Typen der strukturellen Komposition von Dramen unterscheiden (und dann wiederum historisch verorten) lassen. Unser Vorschlag greift dabei Überlegungen aus der Forschung zu sog. Small-world-Netzwerken auf. Diese Forschungen setzen bei der Beobachtung an, dass die Werte von empirisch erhobenen Netzwerken nicht selten signifikant von entsprechenden Random-Netzwerken (also z. B. nach dem Erdős-Rényi-Modell erstellten Graphen) abweichen. Abweichungen sind dabei insbesondere beim Clustering Coefficient, bei der Averge Path Length sowie bei der Degree Distribution zu beobachten (Albert / Barabási 2002). Für den hier projektierten Vortrag werden wir diese Werte – sowie die Werte für die entsprechenden Random-Netzwerke – für unser Gesamtkorpus erheben (sowie einen Workflow für die automatisierte Erhebung entwickeln) und diskutieren. Erste Testläufe deuten dabei darauf hin, dass sich auf diese Weise tatsächlich unterschiedliche Typen der strukturellen Komposition von Dramen beschreiben lassen könnten. So zeigen sich z. B. auffällige Unterschiede bei der Degree Distribution (s. exemplarisch die Tabellen für vier Dramen in Abbildung 2); und mit Blick auf den Clustering Coefficient zeigt sich, dass im Vergleich zu Random-Netzwerken signifikant höhere Werte, wie sie bei Small-world-Netzwerken zu erwarten sind, zwar in mehreren Fällen vorkommen, jedoch keineswegs

für alle Dramennetzwerke charakteristisch sind (siehe exemplarisch die Werte in Abbildung 3). Im Vortrag werden wir diese Werte für alle Dramen unseres Korpus präsentieren; wir werden diskutieren, inwieweit sich hier – aufbauend auf dem Small-world-Konzept – netzwerkanalytisch basierte Typen der strukturellen Komposition von Dramen unterscheiden lassen und wir werden literarhistorisch fundiert erörtern, welche Eigenschaften der Dramen für die unterschiedlichen Werte verantwortlich sind.

Abb. 2.1 bis 2.4: Node Degree Distribution für »Der sterbende Cato« (1731), »Emilia Galotti« (1772), »Götz von Berlichingen« (1773) und »Die Räuber« (1781)



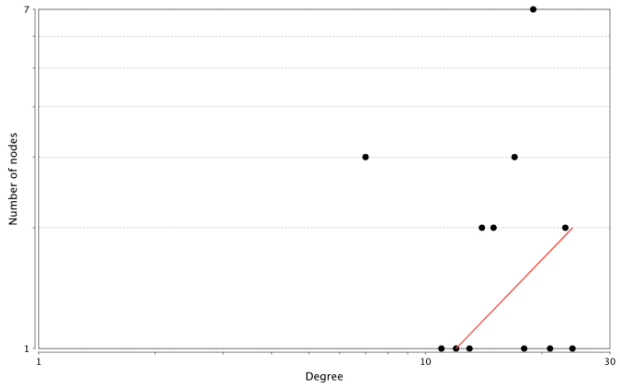


Abb. 3: Vergleich des Clustering Coefficent des Dramen-Netzwerks mit dem eines jeweils entsprechenden Random-Netzwerks

	Clustering Coefficient	Clustering Coefficient (Random: Erdős-Renyi)	Abw. Clustering Coefficient des Dramennetzwerks vom entsprechenden Random-Netzwerk
Faust II	0,941	0,126	746,83%
Hannibal	0,918	0,352	260,80%
Dantons Tod	0,909	0,419	216,95%
Die Journalisten	0,884	0,753	117,40%
Lucie Woodvil	0,883	0,742	119,00%
Die Räuber	0,867	0,693	125,11%
Götz von Berlichingen	0,852	0,116	734,48%
Der gestiefelte Kater	0,834	0,381	218,90%
Die Jungfrau von Orleans	0,769	0,156	492,95%
Der Hofmeister	0,764	0,261	292,72%
Der sterbende Cato	0,75	0,481	155,93%
Kabale und Liebe	0,745	0,496	150,20%
Canut	0,738	0,619	119,22%
Die Soldaten	0,727	0,296	245,61%
Emilia Galotti	0,517	0,367	140,87%

Bibliographie

Albert, Réka / Barabási, Albert-László (2002): "Statistical mechanics of complex networks", in: *Reviews of Modern Physics* 74: 47–97.

Agarwal, Apoorv / Corvalan, Augusto / Jensen, Jacob / Rambow, Owen (2012): "Social Network Analysis of *Alice in Wonderland*" in: *Proceedings of the Workshop on Computational Linguistics for Literature*. Montréal 88–96.

de Nooy, Wouter (2006): "Stories, Scripts, Roles, and Networks" in: *Structure and Dynamics* 1, 2 http://escholarship.org/uc/item/8508h946#page-1 [letzter Zugriff 12. Oktober 2015].

Elson, David K. / Dames, Nicholas / McKeown, Kathleen R. (2010): "Extracting Social Networks from Literary Fiction", in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala 138–147.

Fischer, Frank / Kampkaspar, Dario / Göbel, Mathias / Trilcke, Peer (2015): "Digital Network Analysis of Dramatic Texts", in: *DH 2015, Sydney, 2. Juli 2015* https://dlina.github.io/Our-Talk-at-DH2015/ [Skript] und [Slides] [letzter Zugriff 12. Oktober 2015].

Klotz, Volker (1960): *Geschlossene und offene Form im Drama*. München: Hanser.

Lotman, Jurij M. (1972): *Die Struktur literarischer Texte*. München: Wilhelm Fink.

Marcus, Solomon (1973): *Mathematische Poetik*. Frankfurt am Main: Editura Academiei.

Moretti, Franco (2011): "Network Theory, Plot Analysis" in: *Stanford Literary Lab Pamphlets* 2 http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [letzter Zugriff 12. Oktober 2015].

Park, Gyeong-Mi / Kim, Sung-Hwan / Cho, Hwan-Gue (2013): "Structural Analysis on Social Network Constructed from Characters in Literature Texts", in: *Journal of Computers* 8, 9: 2442-2447 http://ojs.academypublisher.com/index.php/jcp/article/view/jcp080924422447/7672 [letzter Zugriff 12. Oktober 2015].

Rydberg-Cox, Jeff (2011): "Social Networks and the Language of Greek Tragedy", in: *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1, 3 https://letterpress.uchicago.edu/index.php/jdhcs/article/view/86/91 [letzter Zugriff 12. Oktober 2015].

Stiller, James / Nettle, Daniel / Dunbar, Robin I. M. (2003): "The Small World of Shakespeare#s Plays", in: *Human Nature* 14: 397–408.

Stiller, James / Hudson, Matthew (2005): "Weak Links and Scene Cliques Within the Small World of Shakespeare", in: *Journal of Cultural and Evolutionary Psychology* 3: 57–73.

TextGrid: *TextGrid Repository* https://textgridrep.de [letzter Zugriff 10. Februar 2016].

Titzmann, Michael (1977): *Strukturale Textanalyse*. Theorie und Praxis der Interpretation. München: Wilhelm Fink.

Trilcke, Peer (2013): "Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft", in: Ajouri, Philip / Mellmann, Katja / Rauen, Christoph (eds.): *Empirie in der Literaturwissenschaft*. Münster: mentis 201–247.

Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte

Vertan, Cristina
cristina.vertan@uni-hamburg.de
Universität Hamburg, Deutschland

Ellwardt, Andreas
andreas.ellwardt@uni-hamburg.de
Universität Hamburg, Deutschland

Hummerl, Susanne
sanne.hummel@gmx.de
Universität Hamburg, Deutschland

Kurze Darstellung des Altäthiopischen (Ge#ez)

Das südsemitische Ge#ez ist die Sprache des Königreichs Aksum in der heutigen nordäthiopischen Provinz Tigray, von wo aus die im 4. Jahrhundert beginnende Christianisierung Äthiopiens ihren Anfang nahm. Die in der Folge entstehende reiche Literatur ist in großem Umfang geprägt von Übersetzungen aus dem Griechischen und später, ab dem 13. Jahrhundert, aus dem Arabischen, was durch grammatische Interferenzphänomene reflektiert wird. Während seine Verdrängung als gesprochene Sprache bereits im 9./10. Jahrhundert beginnt, bleibt es als Schriftsprache sehr viel länger erhalten und ist bis in die Gegenwart hinein Liturgiesprache des äthiopischen und eriträischen Klerus. Das Altäthiopische hat aus einer südsemitischen Schrift ein eigenes Silbenalphabet entwickelt, das bis heute in mehreren modernen Sprachen Äthiopiens und Eritreas Verwendung findet. Innerhalb der semitischen Sprachen fällt es durch die verwendete Rechtsläufigkeit auf; außerdem werden die Vokale vollständig geschrieben. Beides unterscheidet das Ge#ez von verwandten Sprachen wie Altsüdarabisch, Arabisch, Hebräisch und Syro-Aramäisch. Des weiteren sind Grapheme, die ursprünglich distinkten Phonemen zugeordnet waren, schon früh in identischer phonetischer Realisierung zusammengefallen, was sich konkret bereits in den ältesten überlieferten Handschriftzeugnissen (aber noch nicht in den aksumitischen Inschriften) niederschlägt, wo eine beliebige Austauschbarkeit der Laryngale und Sibilanten jeweils untereinander zu konstatieren ist. Mit den genannten eng verwandten semitischen Sprachen teilt das Altäthiopische die nichtkonkatenative Morphologie. Hierbei muss das einzelne Lexem als Kombination von zwei Elementen beschrieben werden, nämlich der Wurzel und dem Schema: Die konsonantische Wurzel gibt veränderliche Positionen zwischen ihren, zumeist drei, Wurzelkonsonanten vor, die durch die Vokale des Schemas aufgefüllt werden, häufig, jedoch nicht zwingend, ergänzt um (vokalische oder konsonantische) Affixe. Das äthiopische Silbenalphabet bringt dabei mit sich, dass Morphemgrenzen in der Schrift nicht darstellbar sind, sodass beispielsweise ein einzelner Vokal als Bestandteil einer Silbe eine eigenständige Wortart darstellen kann und tokenisiert werden muss; z. B. ist im

zweisilbigen Wort ### /be·tu/ das /u/ als pronominales Suffix zu bet- *u* (*sein* Haus) zu segmentieren.

Ein Tagest für die morphologische Annotation des Ge#ez

Zur Analyse morphologischer Merkmale des Altäthiopischen wurde erstmals ein feingliedriges Tagset von 30 Wortarten entwickelt. Wir unterscheiden vier Klassen von Wortarten: Nomina, Verben, Existentiale, Partikel. Diese Klassen untergliedern sich weiter in die folgenden Wortarten:

- Nomina: Nomen (Eigennamen und Substantive), Pronomen (mit 10 Wortarten) und Zahlen (Kardinal- und Ordinalzahlen)
- Verben
- Existentiale (affirmativ und negativ)
- Partikel (14 Wortarten, z. B. Konjunktionen, Präpositionen, Adverbien)

Den Wortarten wurden entsprechende grammatische Kategorien zugewiesen (Genus, Numerus, Kasus, Person usw.). Ein Sonderfall ist die Bestimmung des Genus für das Nomen. Hier ist das Altäthiopische häufig uneindeutig sowohl in der morphologischen Markierung, als auch in der syntaktischen Kongruenz. Bei eindeutiger Kennzeichnung des grammatischen Genus wird daher weiter dahingehend spezifiziert, wodurch das jeweilige Genus bestimmt ist: durch das morphologische Schema, durch die Syntax und/oder aufgrund des natürlichen Geschlechts (z. B. Mutter). Daher kann ein Nomen im selben Satz im Genus mehrfach bestimmt sein (z. B. syntaktisch maskulin und dem Schema nach feminin). Ist das grammatische Genus nicht eindeutig zu bestimmen, wird es als „unmarkiert“ annotiert.

Das Annotationstool

Die Komplexität des in Sektion 2 dargestellten Annotationstools wird zwar zum einen sehr vielfältige linguistische Anfragen und eine detaillierte Analyse der Sprache ermöglichen, andererseits jedoch verhindert ebendiese Komplexität eine automatische Annotation. Ein Vectorspace-Modell (das für maschinelle Lernverfahren benutzt werden muss) das alle morphologischen Merkmale abdecken würde, wäre zu groß. Vorstellbar ist lediglich eine flache automatische Annotation (z. B. der Wortarten); jedoch wird auch für eine solche zunächst eine relativ große Menge an Trainingsdaten benötigt. Daher ist die Entwicklung eines Werkzeugs für die manuelle Annotation ein obligatorischer Schritt.

Das eigens für das Altäthiopische entwickelte Annotationstool berücksichtigt die spezifischen Besonderheiten der Sprache, von denen einige oben in Sektion 1 kurz skizziert wurden. Aufgrund der dargestellten Eigenheiten sowohl des Silbenalphabets als auch der semitischen Morphologie kann der Text nicht unmittelbar in der Originalschrift annotiert werden. Eine Annotation kann daher ausschließlich in der Transliteration erfolgen (siehe obiges Beispiel *bet-u*).

Die Transliteration wiederum ist nur bedingt durch automatische Regeln beschreibbar. Phänomena wie Konsonantenverdoppelung oder Kontraktion von zwei Silben können nicht durch klare Regeln beschrieben werden. Einige solcher Phänomena ließen sich zwar mittels weiterer Ressourcen automatisch regeln (z. B. Konsonantenverdoppelung bei bestimmten Verbklassen), jedoch müssten derartige Informationen aus einem (bisher noch nicht vorhandenen) digitalen Lexikon extrahiert werden. Auch über die genannten Schwierigkeiten hinaus wäre die Entwicklung einer (semi-)automatischen Transliteration ein äußerst zeitaufwendiger Prozess. Daher haben wir uns für die folgenden Arbeitsschritte entschieden:

Die Texte werden mittels eines automatischen regelbasierten Prozesses transkribiert.
Die Transkription wird manuell korrigiert (entspricht der Transliteration des Textes)

Aus den oben genannten Gründen ist der in der Arbeit mit anderen Sprachen gängige Arbeitsablauf – zuerst Textkorrektur, dann Annotation – hier nicht möglich. Ein solcher ist jedoch bei bereits existierenden Tools Bedingung, wenn eine Mehrebenen-Annotation angestrebt wird. Das CorA-Tool (vg. Bollmann et al.) ermöglicht zwar Korrekturen synchron mit der Annotation, jedoch sind nicht mehr als zwei Annotationsebenen möglich; auch eine Mehrwortannotation ist nicht erlaubt. Für die Annotation muss ein XML-Schema des Tagsets vorliegen und es werden alle möglichen Kombinationen von morphologischen Merkmalen je Wortart generiert. Da sämtliche Kombinationsmöglichkeiten dem Benutzer in Form einer Dropdown-Liste präsentiert werden, ist das Tool in der Anwendung mit dem sehr umfangreichen Tagset für das Altäthiopische ungeeignet. Ein anderes Tool, das relativ häufig angewendet wird, ist WebAnno . Dieses Tool ermöglicht eine Annotation mit mehr als zwei Ebenen, jedoch sind Korrekturen im Text während der Annotation nicht möglich.

Im TraCES Projekt implementieren wir eine neuartige Architektur, die sowohl Änderungen im Text als auch eine Mehrebenen-Annotation ermöglicht.

Wir betrachten als Grundtext den Originaltext in der altäthiopischen Schrift. Die Transliteration bildet die erste und die morphologische Annotation die zweite Ebene, wobei die Transliteration und der Originaltext bei allen Arbeitsschritten synchronisiert bleiben. Im folgenden Abschnitt beschreiben wir das Datenmodell, das diese Architektur ermöglicht.

Die Basiseinheit in unserem System ist ein Wort, das eine einmalige ID zugewiesen erhält. Ein Wort hat folgende Komponenten:

Eine Liste der einzelnen Fidal [1] -Objekte, wo ein Fidal-Objekt aus einer ID und einem Label (dem Fidal-Buchstaben) besteht.
Eine Liste einzelner Silben-Objekte, wo ein Silben-Objekt aus einer ID und einer Liste von einzelnen Buchstaben-Objekten besteht
Ein Buchstaben-Objekt hat immer eine ID und ein Label (das graphische Symbol)

Die Zusammengehörigkeit aller Komponenten wird durch die ID-Zusammensetzung gesichert:
Eine Wort ID ist aus vier Komponenten zusammengesetzt

Projekt ID + Dokument ID + W + automatisch generierte Random ID

Ein Fidal-Buchstaben-Objekt wird dann durch

Wort ID + F + Random-Nummer

identifiziert, während ein Transliterationssilben-Objekt:

Wort ID + TF + Random-Nummer

als ID hat.
Ein Transliterationsbuchstabe wird durch:

Wort ID + Transliterationssilben ID + L + Random-Nummer

identifiziert.
Durch dieses System ist es zu jeder Zeit möglich, jeden einzelnen Buchstaben zu identifizieren und zu referenzieren. Da wir jeden Buchstaben als Objekt betrachten, trennen wir die graphische Realisierung von der linguistischen Annotation; so sind etwa die Annotation und die graphische Repräsentation eines Buchstabens in der Transliteration Labels für ein und dasselbe Objekt.
In der Abbildung 1 wird dieses Modell für die Wortgruppe ###### „ *und vor allem nämlich*“ (eigentlich # - ## - ## - #, aber graphisch quasi ein „Wort“) dargestellt:

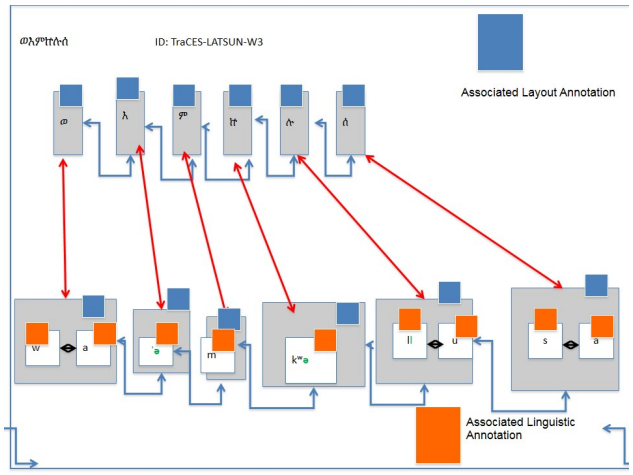


Abb. 1: Annotations-Modell für ######

Zusammenfassung

In diesem Artikel beschreiben wir ein neuartiges Modell für ein Annotationstool, das sowohl eine Annotation mit gleichzeitiger Korrektur, als auch eine Mehrebenen-Annotation ermöglicht. Wir begründen, warum die Entwicklung eines speziellen Modells für die Annotation des Altäthiopischen notwendig war. Möglicherweise könnte das Modell mit wenigen Änderungen auch für andere Sprachen benutzt werden. Eine Demonstration des ersten Prototyps wird auch möglich sein.

Acknowledgements

Das Projekt TraCES wird durch einen Grant der European Science Foundation unterstützt (Grant Agreement 338756).

Die in diesem Artikel beschriebenen Ergebnisse sind das Resultat der Arbeit des gesamten TraCES-Teams: Alessandro Bausi (Projektleiter), Wolfgang Dickhut, Daria Elagina, Andreas Ellwardt, Susanne Hummel, Vitagrazia Pissani, Eugenia Sokolinski, Cristina Vertan.

Notes

1. „Fidal“ ist der Terminus technicus für das äthiopische Silbenalphabet.

Bibliographie

Bollmann, Marcel / Petran, Florian / Dipper, Stefanie / Krasselt, Julia (2014): "CorA: A web-based annotation tool for historical and other non-standard language data", in: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Gothenburg, Sweden 86-90 https://aclweb.org/anthology/W/W14/W14-0612.pdf [letzter Zugriff 30. Dezember 2015].

Castilho, Richard Eckart de / Biemann, Chris / Gurevych, Iryna / Yimam, Seid Muhie (2014): "WebAnno: a flexible, web-based annotation tool for CLARIN", in: *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands http://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf [letzter Zugriff 10. Februar 2016].

Marquez, Lluis / Rodriguez, Horacio / Carmona, Josep / Montolio, Josep (1999): "Improving POS Tagging Using Machine-Learning Techniques", in: *Proceedings of the 1999 Joint SIGDAT conference on Empirical Methods in Natural Language Processing and very large corpora* 53-62 http://www.aclweb.org/anthology/W99-0608 [letzter Zugriff 30. Dezember 2015].

„Jesus ist keine App“ - Fachsprachliche Konzeptualisierungen des ›Computers‹ und Ansätze computergestützter Fachsprachenlinguistik am Beispiel der Domänen Medizin und Theologie

Vogel, Friedemann
friedemann.vogel@medienkultur.uni-freiburg.de
Albert-Ludwigs-Universität Freiburg, Deutschland

Sowohl die Medienkulturwissenschaft als auch die Sprach- und Diskurslinguistik hat sich in der Vergangenheit wiederholt verschiedenen Facetten des Computer- oder Technikdiskurses gewidmet – vgl. etwa Sybille Krämer (1998) zu Computer als Medium, Wichter (1991) und Busch / Wichter (2000) zur Rolle des Computers im Mediendiskurs, Schlobinski et al. (1998), Schlobinski (2006) u.v.a. zur Linguistik des Internets, Vogel (2012) zu Überwachungstechnik oder Müller / Vogel (2014) zu Risikodiskursen. Gemein haben diese Studien alle, dass sie dem ›Computer‹ oder anderen Aspekten der Mediatisierung nahezu ausschließlich im Kontext von gemeinsprachlichen Medientexten, also Mediendiskursen v.a. in Zeitung, Zeitschriften und verschiedenen Online-Formaten nachgehen. Mein eigener Beitrag stellt erste Ergebnisse eines Forschungsprojektes vor, dass sich demgegenüber dem Konzept des ›Digitalen‹ in verschiedenen Fachkulturen widmet. Dabei wird davon ausgegangen, dass sich das ›Digitale‹ nicht allein objektseitig in Gegenständen ‚der Technik‘ fassen lässt, sondern vielmehr als sich verändernde versprachlichte Denkschemata im Umgang mit unserer kulturellen Lebenswelt verstanden werden muss. Drei Beispiele sollen dies kurz illustrieren: (1) Der *Computer* kann im Strafprozess- und Polizeirecht zwar auch ein ›PC‹ (*Rechner* mit Textverarbeitung und Tabellenkalkulation) sein, darüber hinaus aber ›illegitime Form eines neuartigen Fernzugriffs auf die Intimsphäre‹ (klassisches Beispiel aus der Umgangssprache: *Staatstrojaner*) oder ›lediglich ‚digitale‘ Variante der legalen ‚analogen‘ Hausdurchsuchung‹ (*Online-Durchsuchung*). Die verschiedenen Ausdrücke und Konzepte sind mit komplexen Wissensrahmen (u.a.) aus dem Verfassungsrecht und also mit divergierenden Gesellschaftsentwürfen sowie Handlungsnormen verbunden. (2) Das ›Internet‹ wird in theologischen Predigten nicht nur als ‚Rechner-Netzwerk‘ verstanden, sondern mit einem theologischen Leitkonzept, nämlich dem der ›Allgegenwärtigkeit‹, verhandelt (z.B. ›Anmaßung des Menschen, Gott zu spielen‹). (3) Und in der (Korpus-)Linguistik traut man der *Introspektion* nicht mehr über den Weg, hofft aber anthropomorphisierend-paradox auf die *Unbestechlichkeit* der Muster-berechnenden Maschine.

Dem Untersuchungsprojekt zugrunde liegt ein umfassendes, diachrones, annotiertes Korpus aus Fachtexten (derzeit etwa 250.000 Fachaufsätze bzw. 0,53 Milliarden Wortformen) und aus einem Erscheinungszeitraum von 1950 bis 2015, das mit korpus- und computergestützten Methoden diskursanalytisch ausgewertet wird. Im Fokus stehen die vier Domänen Linguistik, Medizin, Theologie und Recht.

Der Vortrag konzentriert sich auf die Konzeptualisierung des ›Computers‹ in Medizin (auch sprachvergleichend) und Theologie. Im ersten Schritt werden zunächst Datengrundlage und eingesetzte Untersuchungsmethoden illustriert und dabei Ansätze korpuslinguistischer Methoden zur Auswertung fachsprachlicher Massendaten diskutiert. Dabei geht es insb. um die Frage, wie sich (Sub-)Korpora dahingehend strukturieren lassen, dass sie möglichst nahe das Zielkonzept der Untersuchung repräsentieren, ohne introspektiven Zirkelschlüssen oder fehlerhaften, rein automatischen Semantik-Annotationen zu erliegen. Im zweiten Schritt werden Untersuchungsergebnisse zu Gemeinsamkeiten und Unterschieden aller vier Fachdomänen und anschließend im vierten Schritt die Konzeptualisierung des ›Computers‹ in Medizin- und Theologie-Diskurs im Detail vorgestellt. Im fünften und abschließenden Schritt wird ein kurzes Resümee gezogen sowie Desiderata zukünftiger computergestützter Fachkommunikationsforschung akzentuiert.

Das hier vorgestellte Thema versteht sich als Beitrag zur gegenstandsbezogenen Forschung sowie zur Entwicklung und Erprobung computergestützter Methoden in den Digital Humanities.

Bibliographie

Busch, Albert / Wichter, Sigurd (eds.) (2000): *Computerdiskurs und Wortschatz. Corpusanalysen und Auswahlbibliographie* (= Germanistische Arbeiten zu Sprache und Kulturgeschichte, 40). Frankfurt am Main, Bern, New York, Paris: Lang.

Krämer, Sybille (ed.) (1998): *Medien, Computer, Realität. Wirklichkeitsvorstellungen und Neue Medien* (= Suhrkamp-Taschenbuch Wissenschaft 1379). Frankfurt am Main: Suhrkamp.

Müller, Marcus / Vogel, Friedemann (2014): „Risikotechnologien in europäischen Mediendiskursen: Der korpuslinguistische Zugriff am Beispiel ‚Biotechnologie'", in: *Technikfolgenabschätzung – Theorie und Praxis*: Risikodiskurse/Diskursrisiken – Sprachliche Formierungen von Technologierisiken und ihre Folgen (= Sonderheft): 40-48.

Runkehl, Jens / Schlobinski, Peter / Siever, Torsten (1998): *Sprache und Kommunikation im Internet. Überblick und Analysen*. Opladen: Westdt. Verl.

Schlobinski, Peter (ed.) (2006): *Von *hdl* bis *cul8r*. Sprache und Kommunikation in den neuen Medien* (= Thema Deutsch 7). Mannheim: Dudenverl.

Vogel, Friedemann (2012): *Linguistik rechtlicher Normgenese. Theorie der Rechtsnormdiskursivität am Beispiel der Online-Durchsuchung* (= Sprache und Wissen 9). Berlin, Boston, Peking, Basel, München: De Gruyter.

Wichter, Sigurd (1991): *Zur Computerwortschatz-Ausbreitung in die Gemeinsprache. Elemente der vertikalen Sprachgeschichte einer Sache* (= Germanistische Arbeiten zu Sprache und Kulturgeschichte 17). Frankfurt am Main, Bern, New York, Paris: Lang.

Annotation und Distant Reading: Probleme, Synergien, Perspektiven

Zirker, Angelika
angelika.zirker@uni-tuebingen.de
Eberhard Karls Universität Tübingen, Deutschland

Bauer, Matthias
m.bauer@uni-tuebingen.de
Eberhard Karls Universität Tübingen, Deutschland

In unserem Vortrag möchten wir Methoden des *close* und *distant reading* miteinander in Beziehung setzen und Probleme sowie Synergien im Rahmen der Digital Humanities diskutieren. Unser Ziel ist dabei, mögliche Perspektiven im Zusammenspiel der beiden Herangehensweisen aufzuzeigen. Ausgangspunkt ist ein Projekt, das bei der Tagung der Digital Humanities im deutschsprachigen Raum 2015 in Graz vorgestellt wurde und das sich mit der erklärenden Annotation von literarischen Texten im Kontext der Digital Humanities befasst.

Unsere Ausgangsfrage ist, inwiefern Methoden des *distant reading* bei Annotationen hilfreich sein können und wo ihre (derzeitigen) Grenzen liegen bzw. in welchen Fällen sie sogar hinderlich oder nicht zielführend sind. Franco Moretti (in seiner Essaysammlung zum *Distant Reading*) wie auch Fotis Jannidis haben aufzeigen können, inwieweit qualitative Methoden Aufschluss über bestimmte Entwicklungen und Trends in der Literaturgeschichte oder in der Geschichte von Gattungen geben können; ebenso ermöglichen quantitative Methoden z. B. die Aufschlüsselung von Charakterkonstellationen. Dabei werden jedoch auch Probleme deutlich, etwa wenn für den identischen Charakter unterschiedliche Namen und Referenzen gebraucht werden: an diesen Punkten versagen automatisierte Verfahren häufig. Eine weitere Schwierigkeit ergibt sich daraus, dass bei der Ermittlung von Worthäufigkeiten die Semantik unberücksichtigt bleibt: eine rein quantitative Analyse etwa des Wortes „bank" in einem englischen Text kann ggf. nicht zwischen den verschiedenen Bedeutungen des Wortes unterscheiden und übersieht somit Ambiguitäten ebenso wie unterschiedlichen Funktionen von Wörtern in der Syntax und Grammatik eines Satzes oder Textes. Dieses Problem ergibt sich beispielsweise bei x-ray, das Verlinkungen auf Wikipedia-Einträge anbietet, dabei aber häufig Ambiguitäten nicht erkennt bzw. lediglich einen Link auf die Disambiguierung von Wikipedia selbst liefert (der dann wiederum dem Leser, dem ja eigentlich geholfen werden soll, die Interpretation des Begriffs überlässt, die Ambiguität also erkennt, aber nicht auflöst). Ebenso ergibt sich ein Problem im Verhältnis von Quantität und Qualität: wenn ein Wort in einem Text nicht häufig genannt wird, heißt das nicht zwangsläufig, dass es nicht wichtig ist und für die Gesamtbedeutung des Textes relevant ist. Vor diesem Hintergrund stellt sich somit auch die Frage, wie Entscheidungen über Bedeutungen im Verhältnis zu distant reading-Methoden getroffen werden können: gibt es hierzu systematische Ansätze? Und (wie) können im Hinblick auf diese Probleme Annotationen von quantitativen Verfahren profitieren?

Im zweiten Teil des Vortrags werden Synergieeffekte von Methoden der Annotation und des *distant reading* vorgestellt. Tools wie etwa der *google Ngram Viewer* erlauben die sehr schnelle Durchsicht von großen Daten- und Textmengen, die manuell nicht zu leisten ist. Sie verschafft dem Annotierenden – einen Überblick, der die erklärende Annotation von Texten erleichtert und etwa auch Querverweise und interne Verlinkungen erleichtert. Der Leser profitiert vom Zusammenspiel der Methoden, denn die individuelle Annotation, ausgerichtet an dem Bedarf sowohl von individuellen Nutzern wie auch von

social communities und Lernern, kann die automatisierten und quantitativen Verfahren anreichern.

Aus diesen Synergieeffekten ergibt sich der Anschluss an Perspektiven zum Verhältnis von (erklärender) Annotation und quantitativen Methoden des *distant reading*. Es ist denkbar, dass Datenbanken es künftig ermöglichen, Vorgänge des erklärenden Annotierens zu automatisieren, etwa in Fällen von Ambiguität, deren Semantik erkannt und die entsprechend aufgelöst wird. In diesem Zusammenhang bedarf er der Ergänzung unserer literaturwissenschaftlichen Perspektive durch technische Expertise. Unser Vortrag stellt somit auch Fragen, die insbesondere im Zuge der Tagung diskutiert werden können. Dabei sollen auch Ideen diskutiert werden, welche Möglichkeiten es geben könnte, dass die manuelle Markierung dessen, was erklärend annotiert (also nicht im Sinne von markup) werden soll, entfällt und Erklärungskontexte definiert werden, innerhalb derer eine datenbankgestützte – und damit „automatische" – Annotation erfolgen kann.

Der Vortrag bewegt sich an der Schnittstelle von Automatisierung und individuellen hermeneutischen Akten und damit entlang des Problems, wie im Markup eines Textes Entscheidungen getroffen werden können, welche Aspekte in einem Text relevant sind und die dem individuellen Text gerecht werden können. Wir möchten verschiedene Fallstudien aus englischsprachigen literarischen Texten vorstellen, etwa anhand von automatisierten Annotationssystemen wie x-ray, die oben geschilderte Probleme exemplarisch aufzeigen, die aber zugleich auch Synergieeffekte deutlich machen. Letztlich sollten bei dem Verhältnis von qualitativen Methoden des *close reading* und quantitativen Herangehensweisen die Nutzerfreundlichkeit sowie die Individualisierung im Vordergrund stehen. Diese Individualisierung ist dabei zweifach: zum einen bezieht sie sich auf Informationen aus Datenbanken, zum anderen auf den Text. Texte, die annotiert werden sollen, müssen mit Datenbanken korrelieren, und die erklärenden Annotationen müssen offenlegen, was im Text gemeint ist sowie was der potentielle Leser erfahren und wissen möchte. Eine solche Anreicherung von Texten ist von quantitativen Methoden bislang nicht zu leisten; umgekehrt sind qualitative Methoden momentan dadurch eingeschränkt, dass sie sich manueller Verfahren bedienen müssen. Das Zusammenspiel von Quantität und Qualität, von *close* und *distant reading*, erklärender Annotation und computergestützter Textanalyse öffnet neue Perspektiven im Bereich der Digital Humanities und kann auch einen Beitrag zu fächerübergreifenden Paradigmen leisten.

Bibliographie

Battestin, Martin C. (1981): „A Rationale of Literary Annotation: The Example of Fielding's Novels", in: *Studies in Bibliography* 34: 1-22.

Bauer, Matthias / Angelika Zirker (2015): „Whipping Boys Explained: Literary Annotation and Digital Humanities", in: Siemens, Ray / Price, Kenneth M. (eds): *Literary Studies in the Digital Age: An Evolving Anthology* http://dlsanthology.commons.mla.org/under-review-matthias-bauer-and-angelika-zirker-whipping-boys-explained-literary-annotation-and-digital-humanities/ [letzter Zugriff 09. Januar 2016].

Cummings, James (2013): „The Text Encoding Initiative and the Study of Literature", in: Siemens, Ray / Schreibman, Susan (eds.): *A Companion to Digital Literary Studies*. Oxford: Blackwell 451-76.

Drucker, Johanna (2012): „Humanistic Theory and Digital Scholarship", in: Gold, Matthew K. (ed.): *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press 85-95.

Jannidis, Fotis (2010): „Methoden der computergestützten Textanalyse", in: Nünning, Vera (ed.): *Methoden der literatur- und kulturwissenschaftlichen Textanalyse*. Stuttgart: Metzler 109-132.

Jannidis, Fotis / Flanders, Julia (2015): „Knowledge Organization and Data Modeling in the Humanities. A whitepaper" http://www.wwp.northeastern.edu/outreach/conference/kodm2012/ [letzter Zugriff 09. Januar 2016].

Jannidis, Fotis / Krug, Markus / Toepfer, Martin / Puppe, Frank / Reger, Isabella / Weimer, Lukas (2015): „Automatische Erkennung von Figuren in deutschsprachigen Romanen". Abstract für die DHd 2015 in Graz.

McCarty, Willard (2012): „Collaborative Research in the Digital Humanities", in: Deegan, Marilyn / McCarty, Willard (eds.): *Collaborative Research in the Digital Humanities*. Farnham: Ashgate 1-10.

Meister, Jan-Christoph (2012): „Crowd Sourcing 'True Meaning': A Collaborative Approach to Textual Interpretation", in: Deegan, Marilyn / McCarty, Willard (eds.): *Collaborative Research in the Digital Humanities*. Farnham: Ashgate 105-122.

Moretti, Franco (2013): *Distant Reading*. New York: Verso.

Stroud, Matthew D. (2006): "The Closest Reading: Creating Annotated Online Editions", in: Bass, Laura R. / Greer, Margaret R. (eds.): *Approaches to Teaching Early Modern Spanish Drama*. New York: The MLA of America 214-129.

Emosaic Visualisierung von Emotionen in Texten durch Farbumwandlung zur Analyse und Exploration

von Lupin, Martin
martin.von.lupin@fh-potsdam.de

FH Potsdam, Deutschland

Geuder, Philipp
philipp.geuder@fh-potsdam.de
FH Potsdam, Deutschland

Leidinger, Marie-Claire
marie.leidinger@fh-potsdam.de
FH Potsdam, Deutschland

Schröder, Tobias
schroeder@fh-potsdam.de
FH Potsdam, Deutschland

Dörk, Marian
doerk@fh-potsdam.de
FH Potsdam, Deutschland

Abstract

Das computergestützte Extrahieren und Visualisieren von Emotionen in Texten ist eine etablierte Technik des “Distant Reading”. Die generelle Stimmung eines Textes kann schnell erfasst werden ohne den gesamten Text lesen zu müssen. Da Emotionen sehr komplex und die Eigenschaften zwischen verschiedenen Emotionen fließend sind, ist die visuelle Charakterisierung von Emotionen schwierig. Wir stellen *Emosaic* vor, ein Online-Tool welches Emotionen aus benutzerdefinierten Texten filtert und durch systematische und nachvollziehbare Farbumwandlung zur Exploration und Analyse innerhalb einer interaktiven Visualisierung bereitstellt. Die durch drei Dimensionen beschreibbaren Emotionen werden dabei in klar definierte Farbparameter übersetzt. Ein von uns entwickelter öffentlich zugänglicher Web- Prototyp (vgl. Geuder et al. 2014-) zeigt anhand interaktiver Visualisierungen erste Analyse- und Explorationsmöglichkeiten dieser Methode.

Einleitung

Wörtern wurden bereits in verschiedenen Studien Emotionen anhand von Emotionsdimensionen wie Valenz (V, engl. valence), Erregung (A, engl. arousal) und Dominanz (D, engl. dominance) zugewiesen (Osgood 1962). Diese Dimensionen sind eine Möglichkeit, Emotionen anhand numerischer Werte zu beschreiben und nach der Affect Control Theory Grundlage aller sozialen Interaktionen (Rogers et al. 2013). Mithilfe der in diesem Forschungskontext empirisch entstandenen Emotionswörterbücher können einem Text somit verschiedene Emotionen numerisch zugeordnet werden. Die dreidimensionale numerische Beschreibung von Emotionen ist hingegen dem intuitiven Verständnis nicht leicht zugänglich. Eine visuelle Entsprechung für sämtliche Emotionszustände würde helfen, die mehrdimensionale Emotionalität eines Textes schnell erfassbar und zudem explorierbar zu machen.

Verwandte Arbeiten

Die Analyse und Visualisierung von benutzergenerierten Inhalten ist ein spannendes und aktuelles Forschungsgebiet. Sowohl wegen des Entstehens großer Textmengen durch Trends wie Online-Weblogs, durch maschinell auslesbare Schnittstellen zu solchen Diensten und durch die Leistungssteigerung von Hardware und Software, hat das Interesse daran zugenommen.

We feel fine (Kamvar et al. 2011) ist ein interaktives Tool welches die Exploration von Emotionen in Weblogs ermöglicht mit dem Ziel den Menschen dabei behilflich zu sein sich selbst bzw. andere Personen besser zu verstehen. Die Ausgabe erfolgt anhand einer experimentellen Visualisierung. Es kann als Suchmaschine verstanden werden, welche das Web nach Emotionen durchsucht. Jede gefundene Emotion wird als Kreis dargestellt. Die Füllfarbe ist abhängig von der zu repräsentierenden Emotion. Glückliche Emotionen werden durch ein helles Gelb vertreten, während Ärger durch ein Rot dargestellt wird. Diese Farbkodierung hilft dem Benutzer bei der groben Unterscheidung von Emotionen, es können diese jedoch nicht weitergehend charakterisiert werden.

We feel (Milne et al. 2015) ist ein webbasiertes Tool welches Emotionen in sozialen Medien verfolgt und aufzeichnet. Das Ziel ist die Aufzeichnung der Weltstimmung bzw. der Stimmung eines Landes um die Verbreitung von physischen Problemen zu erforschen. Die Ausgabe erfolgt anhand interaktiver Graphen und Diagramme. Die Emotionen werden durch deren Namen in Textform dargestellt. Eine Farbkodierung innerhalb der Diagramme hilft bei der Orientierung wobei auch hier keine differenzierte Identifikation von Emotionen bzw. Nuancen zwischen ähnlichen Emotionen möglich ist.

Während beide Tools als Datengrundlage Emotionen aus Webeinträgen nutzen, ist es hingegen unser Ziel die benutzerdefinierte Eingabe eines Textes zu ermöglichen. Zusätzlich sollen sowohl differenzierte als auch nachvollziehbare farbliche Repräsentationen für Emotionen verwendet werden. Die Farbkodierung von Emotionen zur besseren Orientierung innerhalb visueller Darstellungen scheint eine etablierte Methode zu sein, jedoch lassen die bisher vorliegenden Kodierungssysteme keine fundierte Emotionsbeschreibung zu. *Emosaic* hingegen nutzt eine direkte Umwandlung der drei Emotionsebenen in klar definierte Farbparameter, sodass Worte nicht notwendig sind, um Emotionen in Texten unterscheiden und klassifizieren zu können.

Methodisches Vorgehen

Farbzuweisung

Da die Emotionalität eines Wortes nach der Affect Control Theory durch genau einen Punkt im dreidimensionalen Emotionsraum beschrieben werden kann, ist eine Farbzuweisung anhand eines dreiparametrigen Farbraums plausibel. Nach mehreren Iterationen in verschiedenen Farbräumen, wählten wir schließlich den HSV-Farbraum. Die drei Parameter des HSV-Farbraums weisen einen hohen eigenständigen Einfluss auf die resultierende Farbe auf (Farbwert, Farbsättigung und Hellwert). Da der Einfluss eines Parameters auf das Farbresultat klar sichtbar ist, eignet sich dieser Farbraum gut für die Farbumwandlung der drei Emotionsebenen. Von Probanden als plausibel wahrgenommene Ergebnisse erzielten wir, wenn Valenz den Farbwert bestimmt (H, engl. hue), Erregung die Farbsättigung (S, engl. saturation) und Dominanz den Hellwert (V, engl. value).

Während die Sättigung und die Helligkeit Minima und Maxima analog zu Dominanz und Erregung beschreiben, stellt der Farbwert einen kontinuierlichen Farbverlauf dar. Um eine Farbwertannäherung an den Rändern zu vermeiden, haben wir bei der Farbzuweisung einen Farbwertbereich bewusst ausgespart, sodass eine Grenze zwischen Minima und Maxima deutlich hervortritt. Blau entspricht dem Minimum, rot dem Maximum und grün einem mittleren neutralen Valenzwert.

Für die Beurteilung der Zuordnung orientierten wir uns an den sechs Basisemotionen (Liebe, Überraschung, Freude, Wut, Trauer und Angst), wobei wir die Zuweisung der Emotionsdimensionen auf die Farbdimensionen so wählten, dass Liebe einem Rot- / Pinkton entspricht, um der tradierten Farb-Emotions-Zuweisung in der westlichen Kultur zu entsprechen (Abbildung 1a). Eine informelle Studie zeigte, dass diese Form der Zuordnung als intuitiv bewertet wurde. Daneben zeigte sich in Übereinstimmung mit unserer Farbzuweisung, dass negative Gefühle eher dunkel sind und kühlen Farbtönen wie blau oder grün zugeordnet werden (Abbildung 1b), dagegen positive Gefühle eher hell sind und mit warmen Farbtönen wie gelb oder orange in Verbindung gebracht werden (Abbildung 1c).

Datengrundlage für die Charakterisierung der Emotionen

Wir verwenden für die Farbumwandlung das ANEW-Wörterbuch (Warriner et. al 2013) mit über 13.000 englischen Wörtern. Das Wörterbuch wurde ebenfalls vom Projekt “We feel” verwendet.

Möglichkeiten der Farbübersetzung

Durch die Farbübersetzung können emotionale Stimmungen in Texten visuell miteinander verglichen werden. Zusätzlich können Emotionen eines Textes in verschiedenen Ebenen analysiert werden. Sowohl die Gesamtstimmung eines Textes als auch einzelne Sätze können in den Fokus gerückt werden. Stimmungsänderungen im Text können so visuell dargestellt werden. Zudem ist das Filtern von Emotionen möglich. Die farbliche Kodierung unterstützt dabei die Regulierung der Filter.

Aufbau und Funktionsweise des Tools

Grundlegend für die Funktionsweise des Tools ist die Eingabe eines Textes. Der Nutzer kann aus vorgegebenen Texten verschiedenster Länge wählen oder einen eigenen Text innerhalb eines Textfeldes platzieren. Nach der serverseitigen Textanalyse sind verschiedene statische und interaktive Darstellungen verfügbar. Die Darstellung der Emotionsanalyse teilt sich in drei Bereiche auf: Makroansicht, Textansicht und Mikroansicht (Abbildung 2). Die drei Bereiche sind miteinander verlinkt. Grundlegend für die dynamische Änderung einer Ansicht ist die Auswahl von einzelnen Emotionen bzw. Wörtern oder einem Bereich innerhalb einer Emotionsdimension.

Die **Textansicht** ist der zentrale Bereich des Tools (Abbildung 3b). Oberhalb befindet sich ein Histogramm mit permanenter Positionierung und darunter der zu Beginn eingegebene Text, welcher durch Scrollen in voller Länge gelesen werden kann. Ist noch keine Auswahl getroffen, werden alle emotionsrelevanten Wörter innerhalb des Textes durch Hinterlegung mit der korrespondierenden Farbe hervorgehoben. Bei einer Auswahl tritt die Farbhinterlegung von Wörtern, welche außerhalb der Auswahl liegen, in den Hintergrund. Im Histogramm werden alle im Text vorkommenden Farben angeordnet. Die x-Achse kann mit einer der Emotionsdimensionen belegt werden. Durch diesen Ansicht wird die Verteilung innerhalb der Dimensionen sichtbar. Das Histogramm dient zum einen als emotionaler Fingerabdruck und zum anderen als Auswahlwerkzeug. Der Nutzer hat die Möglichkeit durch Klicken und Ziehen einen Bereich im Histogramm auszuwählen. Hierdurch verändert sich dynamisch die Auswahl an Emotionsworten. Durch dieses Brushing ändern sich entsprechende Elemente in der Makro- und Mikroansicht. Alternativ zur Mehrfachauswahl kann eine Einzelauswahl durch einen Mausklick auf das entsprechende Wort vorgenommen werden.

Die **Makroansicht** bietet einen ersten Überblick über allgemeine Emotionstendenzen und -entwicklungen im

Text (Abbildung 3a). Mit dem links platzierten Diagramm kann untersucht werden, inwiefern sich die einzelnen Werte (V, A, D) innerhalb des Textes verändern und wie sie in Beziehung zueinander stehen. Hat der Nutzer eine Auswahl getroffen bietet der Index daneben eine Übersicht darüber, wo sich die entsprechenden Textstellen befinden. Klicken auf den Index ermöglicht schnelles Springen zur entsprechenden Textpassage.

Die **Mikroansicht** ermöglicht das Explorieren des Textes im Detail (Abbildung 3c). Die Wörter innerhalb der getroffenen Auswahl werden hier aufgelistet. Zu der Emotionsfarbe und der Häufigkeit des Vorkommens erfährt der Nutzer hier auch die Zusammensetzung aus den vad-Werten. Hat der Nutzer nur ein Emotionswort gewählt, werden zusätzlich emotionsverwandte Worte aufgelistet, um weiteres Explorieren des Textes zu ermöglichen.

Zukünftige Arbeiten

Die Analyse und Exploration von Emotionen in Texten anhand von Farben, basierend auf deren Emotionsdimensionen ist ein innovativer Ansatz zur Textanalyse. Der Umgang durch Nutzer mit unserem Tool ist dabei Bestand weiterer Untersuchungen. Hierzu ist bereits eine langfristig angesetzte Evaluation gestartet worden, die neben automatisch generierten Parametern der Texte wie Länge und Emotionalität auch von Nutzern angegebenes Feedback wie zum Beispiel dem Verwendungszweck des Tool beinhaltet. Auf Basis der Evaluationsergebnisse planen wir Erkenntnisse zu der Wirkung von Farbumwandlung von Emotionen und dem generellen emotionsbezogenen Interessensfokus von Nutzern zu gewinnen. Limitiert durch unser verwendetes Wörterbuch ist die Analyse momentan ausschließlich auf englische Texte beschränkt. Wir planen weitere Sprachen zu integrieren, insofern ähnliche Wörterbücher für diese Sprachen zur Verfügung stehen. Zudem muss ein möglicher Mehrwert neuer bzw. alternativer Textanalysemethoden untersucht werden. Da Farben und deren emotionale Empfindung kulturabhängig sind, wäre die Untersuchung von Sprachen aus verschiedenen Kulturräumen ebenfalls besonders interessant. Des Weiteren stellt sich die Frage, wie mit der Kombination von Adjektiven und bedeutungsverändernden Partikeln („a little bit“, „very“, „not“, „only“) und der damit einhergehenden Veränderung des emotionalen Gehalts des Wortes umgegangen wird.

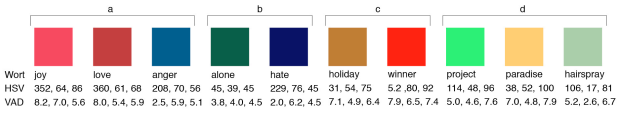


Abbildung 1: Farbumwandlung verschiedener Wörter inklusive der Farbwerte (HSV) und Emotionsdimensionen (VAD): Ausschnitt der Basisemotionen (a), negative Wörter (b), positive Wörter (c), auffällige Farben (d).

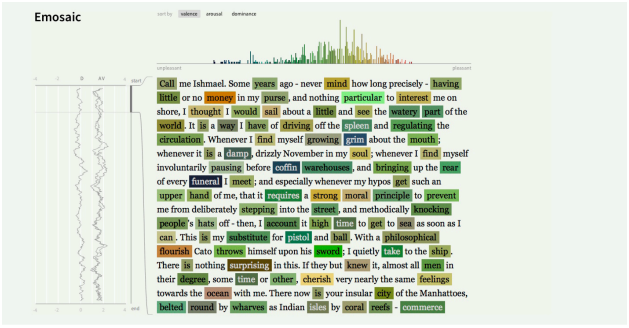


Abbildung 2: Ansicht nach Eingabe des zu analysierenden Textes.

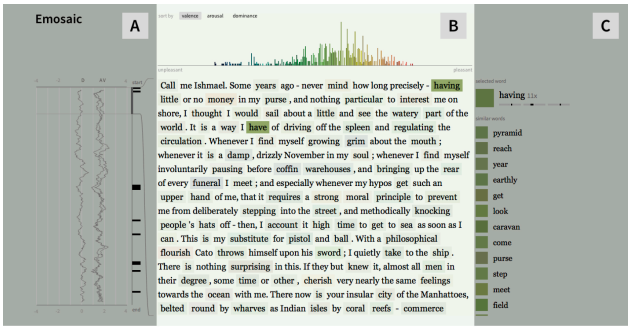


Abbildung 3: Darstellung des Tools nach Auswahl eines einzelnen Wortes. Die grauen Farbüberlagerungen dienen hier zur Erklärung der verschiedenen Bereiche: (A) Makroansicht mit Werteverlauf innerhalb des Textes von valence, arousal und dominance und Index zur Lokalisierung der Textpassagen, (B) Histogramm und eingegebener Text mit farblich hervorgehobenen emotionsrelevanten Wörtern, (C) Mikroansicht mit ausgewähltem Wort und verwandten Wörtern innerhalb des Textes.

Bibliographie

Geuder, Philipp / Leidinger, Marie-Claire / von Lupin, Martin (2014-): *emosaic* http://emosaic.de/ [letzter Zugriff 10. Februar 2016].

Kamvar, Sepandar D. / Harris, Jonathan (2011): "We feel fine and searching the emotional web", in: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. ACM 117–126.

Milne, David / Paris, Cecile / Christensen, Helen / Batterham, Philip / O'Dea, Bridianne (2015): "We Feel: Taking the emotional pulse of the world", in: *Proceedings of the 19th Triennial Congress of the International Ergonomics Association (IEA 2015)*.

Moretti, Franco (2013): *Distant Reading*. London / New York: Verso.

Osgood, Charles E. (1962): "Studies on the generality of affective meaning systems", in: *American Psychologist* 17, 1: 10-28.

Rogers, Kimberly B. / Schröder, Tobias / von Scheve, Christian (2014): “Dissecting the sociality of emotion: A multilevel approach”, in: *Emotion Review* 6: 24-33.

Warriner, Amy Beth / Kuperman, Victor / Brysbaert, Marc (2013): "Norms of valence, arousal, and dominance for 13,915 English lemmas", in: *Behavior Research Methods* 45: 1191-1207.

Poster

Das Tool LAKomp und seine Anwendung auf Texte nichtstandardisierter Sprachstufen

Aehnlich, Barbara
Barbara.Aehnlich@uni-jena.de
Friedrich-Schiller-Universität Jena, Deutschland

Kösser, Sylwia
sylwia.koesser@germanistik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg

Die Verarbeitung historischer Sprachdaten des Deutschen birgt zahlreiche Probleme: Sie weisen einen hohen Grad an Variation auf, insbesondere auf den Ebenen Phonologie und Graphematik, aber auch in den Bereichen der Morphologie, Syntax und Lexik. Die bisher entwickelten Tools, z. B. im Bereich der automatischen Wortarten-Annotation, sind auf Daten des Gegenwartsdeutschen trainiert und können deshalb nur bedingt oder gar nicht auf Daten historischer Sprachstufen angewandt werden.

Für die Lemmatisierung und Annotierung mit Part-of-Speech-Tags existieren bereits linguistische Werkzeuge, die nach einer Trainingsphase auf bereits annotierten Texten weitere Texte automatisch annotieren können. Angewendet auf frühneuhochdeutsche Texte liefern diese Werkzeuge aber hohe Fehlerraten, denn eine Voraussetzung für ihr Funktionieren ist hier schwer erfüllbar: das Erkennen von Wortformen. Hier stellt die stark variierende Graphie ein Hindernis dar.

Im Projekt SaDA (Semiautomatische Differenzanalyse von komplexen Textvarianten) (Bremer et al. 2012-2015) werden deshalb elektronische Werkzeuge entwickelt, die der Aufbereitung eines historischen Korpus dienen sollen und zur Anwendung in verschiedenen philologischen Bereichen gedacht sind. Zur Erstellung eines strukturierten Korpus ist die Anreicherung der Überlieferungszeugen mit verschiedenen Informationen Voraussetzung. Zu diesem Zweck wurde das Werkzeug LAKomp [1] entwickelt, mit dessen Hilfe alle im Zuge der Bearbeitung dem Text hinzugefügten Informationen gespeichert und für die spätere Nutzung aufbereitet werden.

LAKomp wird unter anderem an der "Wundarznei" des Heinrich von Pfalzpaint (weiter)entwickelt. Nach der Transkription der Überlieferungszeugen nach den Konventionen und Kodierungen der Mittelhochdeutschen Grammatik, des Referenzkorpus Mittelhochdeutsch und des Referenzkorpus Frühneuhochdeutsch werden die Texte lemmatisiert und annotiert.

Die morphologische Annotation reichert das Textmaterial zunächst mit der Angabe der Wortart an, wobei Verben und Nomina weiter spezifiziert, also mit Angaben zu den verbalen und nominalen Kategorien versehen werden. Syntaktische Informationen werden teilweise durch die Unterscheidung attributiver, prädikativer oder adverbialer Verwendung bei Adjektiven und Partizipien geliefert.

Durch Lemmatisierung und Annotation werden die Wortformen der einzelnen Handschriften einem tertium comparationis gegenübergestellt. Durch diese Abstraktion, die Zuweisung einer der Einzelgraphie übergeordneten Wörterbuchform (bei parallelem Erhalt der konkreten Handschriften-Graphie), wird ein sehr konkreter maschineller Vergleich möglich.

Mit der vorgenommenen Kodierung des Quellenmaterials ist ein semi-automatischer Textzeugenvergleich möglich. Zunächst durch die Segmentierung, aber vor allem durch die Lemmatisierung und noch stärker durch die grammatische Auszeichnung können die einzelnen Handschriften konkret aufeinander abgebildet werden, sodass Abweichungen und damit Filiationsverhältnisse deutlich sichtbar werden. Für die Darstellung der Unterschiede und Gemeinsamkeiten der Textzeugen werden diese in einem sogenannten Partiturtext vertikal dargestellt, miteinander verglichen und die Unterschiede zusätzlich farbig markiert. Der Partiturtext wird von LAKomp unter Zuhilfenahme der vorher beigegebenen Informationen automatisch erzeugt.

Neben der einfachen Suchfunktion kann das zuvor im textspezifischen Wörterbuch abgelegte und mit Informationen angereicherte Wortmaterial auch mit der Analysefunktion gezielt durchsucht werden. So bietet sich dem Nutzer beispielsweise die Möglichkeit, alle Graphieformen eines Lemmas abzurufen und ihre statistische Verteilung in den Handschriften und Drucken abzufragen. Neben der prozentualen Verteilung werden ebenso die Belegzahlen und die einzelnen Graphieformen ausgegeben.

Im Rahmen eines an der MLU Halle geplanten Projekts zu medizinischen Sachtexten des Mittelalters soll LAKomp weiterentwickelt werden, um die Untersuchung der medizinischen Inhalte (Texte und Objekte) hinsichtlich verschiedener Fragestellungen (Verschlagwortung, Datenbank, Verknüpfung von Informationen) und eine optimierte nutzerbezogene Darstellung der Ergebnisse gewährleisten (Analysefunktion, Satzprogramm zum Edieren der Texte, kartographische Darstellung) zu können. Die Überlieferung der Zeit von 1350 - 1650 ist vor allem durch Kompilationen medizinischer Texte geprägt, was eine Einordnung einzelner Texte in Überlieferungswege und -zusammenhänge bedeutend erschwert. Grundvoraussetzung für die Entwicklung und Verifizierung von Werkzeugen ist ein geeignetes Korpus. Text- und Objektbasis dieser Pilotstudie ist die "Wundarznei" des Heinrich von Pfalzpaint aus dem Jahre 1460. Anhand dieses Textes sollen die Möglichkeiten

zur Beantwortung verschiedenster Fragen exemplarisch erprobt und Werkzeuge zur Umsetzung und Darstellung entwickelt werden.

Ein weiteres Projekt, das sich auf das Tool LAKomp stützt, befasst sich mit Rechtstexten aus der Rezeptionszeit des römischen Rechts (Aehnlich 2016). Es beruht auf einem Korpus zweier frühneuhochdeutscher Rechtsbücher des 15. und 16. Jahrhunderts. Der Klagspiegel ist das mit Abstand älteste populärwissenschaftliche Rechtsbuch der Rezeptionszeit und bildet mit dem Laienspiegel zusammen die wichtigste Grundlage an rechtswissenschaftlichen populären Texten des 15. und 16. Jahrhunderts. Davon ausgehend ist ein Projektantrag zu einem Korpus von Strafrechtstexten der frühen Neuzeit in Arbeit, welches ebenfalls mithilfe von LAKomp strukturiert und aufbereitet werden soll. Durch semantische und linguistische Annotationen soll eine umfassende Forschungsgrundlage geschaffen werden, die für die Schließung rechts- und sprachhistorischer Forschungslücken einen zentralen Beitrag leistet.

Das Poster stellt das Werkzeug LAKomp mit seinen Einsatzmöglichkeiten und -gebieten vor. Am Beispiel des Pfalzpaint und des Laienspiegels wird gezeigt, dass das Tool einfach und intuitiv bedienbar ist.

Notes

1. **L**emmatisierung, **A**nnotation, **Komp**aration.

Bibliographie

Aehnlich, Barbara (2016): *Sprachwissenschaftliche Untersuchungen zum Klagspiegel Conrad Heydens (1436) und zum Laienspiegel Ulrich Tenglers (1509)*. Universität Jena http://www.sprachwissenschaft.uni-jena.de/Lehrbereiche/Geschichte+der+deutschen+Sprache/Dr_+Barbara+Aehnlich/Projekt-p-1881.html [letzter Zugriff 28. Januar 2016].

Bremer, Thomas / Molitor, Paul / Ritter, Jörg / Solms, Hans-Joachim (eds.) (2012-2015): *SaDA*. Semi-automatische Differenzanalyse von komplexen Textvarianten. Martin-Luther-Universität Halle http://www.informatik.uni-halle.de/ti/forschung/ehumanities/sada/ [letzter Zugriff 08. Januar 2016].

Visualisierung von Ortsnamen im Deutschen Textarchiv

Barbaresi, Adrien
adrien.barbaresi@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich; Berlin-Branbenburgische Akademie der Wissenschaften, Deutschland

Textarchiv und Umfang der Studie

Das DFG-geförderte Projekt „DeutschesTextarchiv“ (DTA) der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) stellt deutschsprachige Drucke als Bilddigitalisate und TEI-XML-annotierte Volltexte aus mehr als 300 Jahren, vom Beginn des 17. bis zum frühen 20. Jahrhundert, über das Internet zur freien Nutzung bereit. Das DTA mit seinen Erweiterungskorpora umfasst derzeit knapp 2800 Dokumente mit mehr als 630 000 digitalisierten Seiten und ca. 1,1 Mrd. Zeichen (Stand: 7.10.2015). Neben dem Anspruch, vielseitig nutzbare und qualitativ hochwertige Primärquellen frei verfügbar zumachen, liegt der Fokus des DTA-Projekts auf der korpus- bzw. computerlinguistischen Analyse der elektronischen Volltexte. Alle Quellen stehen in verschiedenen Formaten zum Herunterladen und auch zum „Harvesten“ über eine API bereit (BBAW 2007-2015).

Im Rahmen des DTA wurden in den letzten beiden Jahren verschiedene automatische und semiautomatische Ansätze zur Erkennung von Personen- und Ortsnamen evaluiert. Immer wieder bedeuteten dabei die Heterogenität des Korpus und die große sprachliche Varianz innerhalb des Textkorpus eine Herausforderung für die Tools: Je früher die Entstehungszeit eines Textes liegt, desto größer werden sprachliche und sachliche Differenz bei der Benennung von Eigennamen. Der Fokus wird im Folgenden auf Ortsnamen in historischen Texten des DTA liegen.

Das Ziel der Studie besteht darin, die Verteilung der im DTA erwähnten Ortsnamen darzustellen, um ein synthetisches Bild der Sammlung zusammenzustellen und gleichzeitig Rückschlüsse auf den Inhalt zu ermöglichen. Sie erfolgt im Rahmen einer Kooperation zwischen der Berlin-Brandenburgischen Akademie der Wissenschaften (Zentrum Sprache) und der Österreichischen Akademie der Wissenschaften (ICLTT, Institut für Corpuslinguistik und Texttechnologie), beide Zentren verfügen über digitalisierte historische Textkorpora.

Erkennung von Ortsnamen

Spezialisierte Werkzeuge aus dem Gebiet der Computerlinguistik werden im Rahmen dieser Studie eingesetzt. Erstens wird für die Tokenisierung (Segmentierung in Wortformen) die Software WASTE (Jurish / Würzner 2013) benutzt, die speziell für Texte verschiedener Epochen im Rahmen des DTA entwickelt worden ist. So lassen sich Sprachqualitäten besser annähern, die von den heutigen Standards abweichen.

Die deutsche Version des „Wikiwörterbuchs“ Wiktionary der Wikimedia Stiftung, das von Internetnutzern gepflegt wird, wird verwendet, um lexikalische Informationen über Wörtern zu sammeln. Ziel dieses Vorgehens ist es unter anderem, solche Token zu erkennen, die mit Sicherheit keine Eigennamen sind. Ein weiteres mögliches Problem besteht bei Eigennamen, die keine Ortsnamen sind, jedoch aus verschiedenen Gründen als solche ausgezeichnet worden sind, u. a. Namen von bekannten Autoren so wie fiktive Namen und Vornamen. Listen werden also benutzt, um bereits bekannte Eigennamen auszugrenzen.

Die Erkennung von Ortsnamen beruht oft auf Verfahren aus der künstlichen Intelligenz sowie named-entity recognition (Leidner / Lieberman 2011). Wissensbasierte Methoden zeigen jedoch auch vielversprechend Ergebnisse, so wie zum Beispiel anhand von Datenbanken aus Wikipedia (Hu et al. 2014).

Unsere Erkennung der Ortsnamen erfolgt über Datenbanken, die die Vorteile von geisteswissenschaftliche Sorgfalt und Opportunismus aus Big Data Herangehensweisen kombiniert. Über ein gleitendes Fenster wird nach Treffern (einschließlich Mehrwortausdrücken) gesucht. Aus der passenden Datenbank werden Koordinaten und gegebenenfalls weitere geographisch relevante Informationen extrahiert, diese Daten werden wiederum in einer weiteren für das gesamte Verfahren angelegten Datenbank zusammengefasst. Falls mehrere Möglichkeiten bestehen, ist ein Disambiguierungsverfahren nötig, das Informationen wie Distanz, Kontext und aktuelle Bevölkerungszahlen benutzt.

Die Erkennung erfolgt über die Durchsuchung von Listen unterschiedlichen Ranges: als Erstes wird nach aktuellen sowie ehemaligen Ländern und vergleichbaren Hoheitsgebieten gesucht (z. B. Österreich-Ungarn), dann wird die Suchanfrage um Regionen oder regionale Landschaften erweitert (z. B. Schwaben), bei einem negativen Ergebnis wird anschließend nach Städten und schließlich nach geographischen Merkmalen wie Flüssen oder Bergen gesucht. Die dafür nötigen Informationen wurden zum Teil händisch (Staaten und Regionen) und zum Teil automatisch gesammelt und händisch zusammengefasst oder überprüft (Städte und Geographie). Da mancherorts die Staatsgrenzen bis ins 20. Jahrhundert instabil geblieben sind und da gewisse Staaten sich durchaus als multinational verstehen lassen, wurden insbesondere für Mitteleuropa Listen einschließlich der aktuellen oder ehemaligen deutschen Namen erstellt, u. a. anhand von bereits im Web auffindbaren Listen wie zum Beispiel Kategorien oder Listen von Wikipedia.

Jedem eindeutigen Ortsnamen wurden dann Koordinaten hinzugefügt, entweder durch automatische Abfrage von Wikipedia und Wikidata oder händisch unter Heranziehungen historischer Beschreibungen oder Atlanten. Bei politischen Entitäten wurde bisher Europa im 19. und 20. Jahrhundert in Betracht gezogen. Die Listen werden regelmäßig erweitert, sie umfassen derzeit 78 Hoheitsgebiete, 858 Regionen, 9.846 Städte und 13.962 geographische Merkmale.

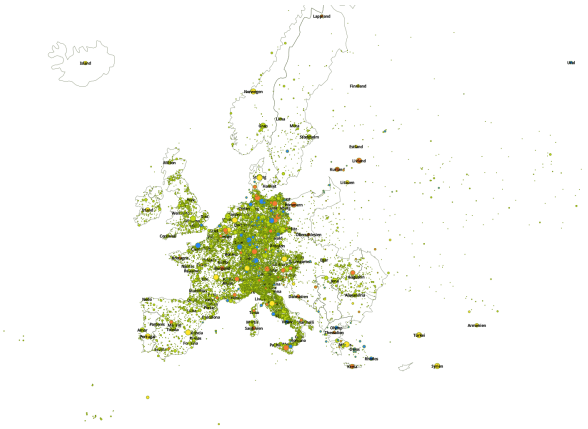
Wenn kein Treffer in den Listen gefunden wird, werden größere, automatisch erstellte Ortsregister in Betracht gezogen. Geographische Informationen über Orte stammen dann aus den Geonames-Datenbanken, die zum Beispiel von dem Openstreetmap Projekt benutzt werden, und dessen Creative Commons Attribution Lizenz eine Wiederverwendung der Daten ermöglicht. Alle Datenbanken für aktuelle europäische Länder sind gesammelt und verarbeitet worden: gewisse Ortstypen (nämlich Region und bewohnter Ort) sind ausgewählt worden, und existierende Varianten in diversen europäischen Alphabeten sind extrahiert worden, um mögliche Änderungen im Laufe der Geschichte zu reflektieren.

Projektion auf einer Karte

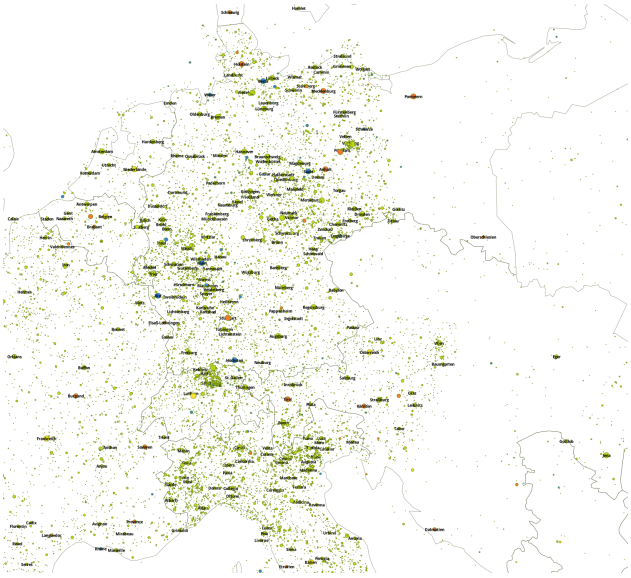
Schließlich werden die Ergebnisse auf eine Karte Europas projiziert, die die tatsächliche politische Lage dieser Zeit spiegelt. Dafür werden die Grenzen von 1914 gezeigt. Der quantitative Schwerpunkt des Korpus liegt nämlich auf dem 19. Jahrhundert liegt und der Stand vor dem ersten Weltkrieg gibt ein vernünftiges Bild von Europa während des 'langen 19. Jahrhunderts'. Die Qualität der Daten so wie des graphischen Resultats wurde in mehreren Durchläufen geprüft, dabei wurden jeweils verbliebene Fehler eliminiert: die Karte bzw. die Projektion der Daten wird so sukzessive verbessert und feiner justiert.

Zur Projektion wird die Kartographieumgebung TileMill benutzt, die eine Anpassung anhand der Stylesheet-Sprache CartoCSS ermöglicht. So können wichtige Punkte im Graphen hervorgehoben werden. Die Wahl verschiedener Farben erleichtert den Überblick über das visualisierte Ergebnis und dessen Interpretation, da im Feld der Visualisierungsstudien bekannt ist, dass das menschliche Auge instinktiv unterscheiden und klassifizieren kann (Bertin 1967).

Karte



Karte von Europa in den Grenzen von 1914 (Hoheitsgebiete in Gelb, Regionen in Orange, Städte sowie aus Geonames extrahierte Ortsnamen in Grün, geographische Merkmale in Blau).



Zoom: Karte von Mitteleuropa in den Grenzen von 1914 (Hoheitsgebiete in Gelb, Regionen in Orange, Städte sowie aus Geonames extrahierte Ortsnamen in Grün, geographische Merkmale in Blau).

Diskussion

Wir hegen die Hoffnung, dass solche Visualisierungsstudien den Weg nach einer größeren Sichtbarkeit von digitalem Kulturerbe und von literarischer Forschung im digitalen Zeitalter ebnen. Genauer betrachtet glauben wir, dass detailreiche Annäherungsweisen gefragt werden, die sowohl auf technischer Kompetenz als auch auf historisches und literarisches Wissen aufbauen. In diesem Sinne planen wir, mehr Metadaten einzubeziehen sowie vielseitige Visualisierungen zu erzeugen.

Es sollte immer berücksichtigt werden, dass die linguistischen Korpora, die als Basis für die Karte benutzt werden, immer schon ein Konstrukt sind, woraus folgt, dass die auf diesen Daten basierenden Projektionen ebenso Konstrukte sind: Auch wenn sie unmittelbar interpretierbar scheinen, spielen Qualität der Daten, Spezialisierungsgrad der Verarbeitungskette und Qualitätsprüfung eine maßgebende Rolle. Deswegen sind wir der Meinung, dass eine gewisse Dekonstruktion des Prozesses nötig ist, im Sinne einer Öffnung der black box, die dem Betrachter das originelle Moment der Entzückung vielleicht wegnimmt, aus wissenschaftlicher Sicht jedoch wünschenswert ist. So möchten wir keine Fehler kaschieren, eventuelle Verzerrungen nicht verschweigen, und für die Reproduzierbarkeit des gesamten Prozesses sorgen, einerseits durch detailreiche Dokumentierung des Prozesses, andererseits durch die Herausgabe möglichst aller dabei verwendeten Tools und Komponenten als Open Source Software.

So können unmögliche, in diesem Kontext falsche Verbindungen vermieden werden: es gibt zum Beispiel einen Ort namens "Hermann" in Norwegen, was Fakt und Datum zugleich ist. Es ist jedoch nötig, sich nicht allein auf diese Datengrundlage zu verlassen, wenn man nach Orten sucht, sonst wird das Endprodukt – d. h. die Karte – verfälscht.

Interessanterweise bietet die Karte für einen möglichen Hermeneuten genau diese Fälle in Perspektive, die Existenz eines möglicherweise falschen Knotens bleibt nicht unbemerkt und wirft Fragen auf. Auf dieser Weise ist uns zum Beispiel ein systematischer Fehler mit den Vornamen aufgefallen. Durch diese hermeneutische Schleife wird die Analyse nach und nach verschärft.

Die Säuberung der Daten ist in dieser Hinsicht entscheidend, die Anzahl und Vielfalt von Filtern, die eingesetzt werden, erheben unsere Arbeit von einer massiven Datensammlung und -Analyse auf das Niveau einer Studie in Digital Humanities, die Rücksicht auf Besonderheiten einer Sprache und einer Epoche nimmt.

Bibliographie

BBAW (2007-2015): *Deutsches Textarchiv*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften http://www.deutschestextarchiv.de/ [letzter Zugriff 30. Januar 2016].

Bertin, Jacques(1967): *Sémiologie graphique*. Paris: Mouton / Gauthier-Villars.

Hu, Yingjie / Janowicz, Krzysztof / Prasad, Sathya (2014): „Improving Wikipedia-Based Place Name Disambiguation in Short Texts Using Structured Data from Dbpedia“, in: *Proceedings of the 8th ACM Workshop on Geographic Information Retrieval, Dallas, Texas* 8–16.

Jurish, Bryan / Würzner, Kay-Michael (2013): „Word and Sentence Tokenization with Hidden Markov Models“, in: *Journal for Language Technology and Computational Linguistics* 28, 2: 61–83.

Leidner, Jochen L. / Lieberman, Michael D. (2011): „Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language“, in: *SIGSPATIAL Special* 3, 2: 5–11.

Ähnlichkeitssuche in den Digital Humanities: Semi-automatische Identifikation von Kostümmustern

Barzen, Johanna
johanna.barzen@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Falkenthal, Michael
Falkenthal@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Hentschel, Frank
Frank.Hentschel@uni-koeln.de
Universität zu Köln, Deutschland

Leymann, Frank
Leymann@iaas.uni-stuttgart.de
Universität Stuttgart, Deutschland

Strehl, Tino
Tino.Strehl@student.reutlingen-university.de
Hochschule Reutlingen, Deutschland

Ausgangslage

Kostüme in Filmen sind ein wichtiges Gestaltungselement der diegetischen Welt. Mit MUSE [1] (MUster Suchen und Erkennen) verfolgen wir das Ziel, Konventionen zu identifizieren und darstellbar zu machen, die sich entwickelt haben, um Kostüme als kommunikatives, bedeutungstragendes Element zu nutzen. Um diese Konventionen zu identifizieren, verwenden wir das Konzept des Musters nach Christopher Alexander et al. (1977). In dieser Tradition kann ein Kostümmuster als abstrakte und bewährte Lösung eines wiederkehrenden Designproblems, wie beispielsweise der adäquate textile Ausdruck eines bestimmten Charakters, verstanden werden.

Um die Identifikation und das Verfassen von Mustern zu unterstützen, haben wir ein Lösungs- und ein Musterrepository konzipiert und implementiert. Während das Lösungsrepository ein detailliertes Erfassen der Kostüme aus Filmen ermöglicht (konkrete Lösungen für Designprobleme), können im Musterrepository abstrakte Designlösungen (Kostümmuster) abgelegt werden (Fehling et al. 2014). Wie aber identifiziert man diese Kostümmuster aus der Menge der multidimensional beschriebenen Kostümdaten?

Einen ersten Ansatz haben wir mittels der Analyse aufbauend auf OLAP Cubes vorgestellt (Barzen 2015). Dieser Ansatz erlaubt multidimensionale Abfragen auf den Kostümdatenbestand, beschränkt sich allerdings auf die Analyse der Kostümdaten durch konkrete Abfragen. Bei konkreten Abfragen nicht vermutete Zusammenhänge im Datenbestand können dabei nicht identifiziert werden. Um solche Zusammenhänge der Daten sichtbar zu machen, gewinnen besonders in Industrie und Naturwissenschaften Techniken aus dem Bereich des Data Minings an Gewicht. Diese erlauben mögliche „Auffälligkeiten“ oder Cluster in Datensätzen zu finden. Was wir in diesem Poster vorstellen möchten, ist eine Werkzeugumgebung, die verschiedene Algorithmen und entsprechende Visualisierungen der Analyseergebnisse zur Identifikation von „Kostümmusterkandidaten“ unterstützt. Dem vorgegebenen Umfang geschuldet, beschränken wir uns in diesem Abstrakt auf das Vorstellen einer der angewandten Methoden: Wie kann man die Ähnlichkeit der Daten selektiv auswerten um durch die Visualisierung ähnlicher Ausprägungen von Kostümen aus dem Lösungsrepository Hinweise auf Kostümmuster zu erhalten?

Methodischer Ansatz (exemplarisch)

Um ähnlich wirkende Artefakte (hier die konkreten Kostüme und deren Basiselemente wie Hosen, Pullover, etc.) zu identifizieren und zu visualisieren, machen wir uns die taxonomische Strukturierung (Barzen 2013) des Datenbestandes als Hintergrundwissen zunutze. Um eine detaillierte und strukturierte Erfassung der Kostüme zu gewährleisten, werden sie durch die Eingabe der kostümrelevanten Parameter (Attributbeschreibungen wie Farbe, Material, Zustand etc.), deren Wertebereich durch zugrundeliegende Taxonomien vorstrukturiert ist, beschrieben und im Lösungsrepository gespeichert. In der Literatur gibt es bewährte Verfahren, um aus einer Taxonomie, die Ähnlichkeiten von Objekten berechnen zu können. Insbesondere in der Biologie (Lord 2003), aber auch in der Linguistik (Jiang 1997) haben sich beispielsweise Verfahren zur Ähnlichkeitsbestimmung von Genotypen oder Sprachbausteinen bewährt. Dieser Ansatz soll auf die Kostümdaten übertragen werden.

Um die Ähnlichkeit von Artefakten zu bestimmen, wird die Struktur der Taxonomie als Hintergrundwissen einer Distanz-Funktion als Graph bereitgestellt. Aufbauend auf der Distanzmetrik, die Wu und Palmer für die Bestimmung konzeptueller Entfernung zwischen

Begriffen (Palmer 1994) entwickelt haben, soll die Ähnlichkeit von Artefakten über die jeweiligen Distanzen ihrer Attributsausprägungen bestimmt werden. Eine Anwendung dieser Metrik auf die Attributsausprägungen „Farbe“ soll in Abbildung 1 demonstriert werden. Hier wird den Farbklassen „Hellblau“ und „Gelbtöne“ über Bestimmung des gemeinsamen Elternknotens (C3) und der Kantenanzahlen von jeder Klasse (C1 und C2) zu dem Elternknoten (N1 und N2), sowie von Elternknoten zu Wurzelknoten (N3) durch die Anwendung der Distanzmetrik ein Ähnlichkeitswert von 0,4 zugeordnet (wobei 1 mit Identität und 0 mit völliger Verschiedenheit korrespondiert).

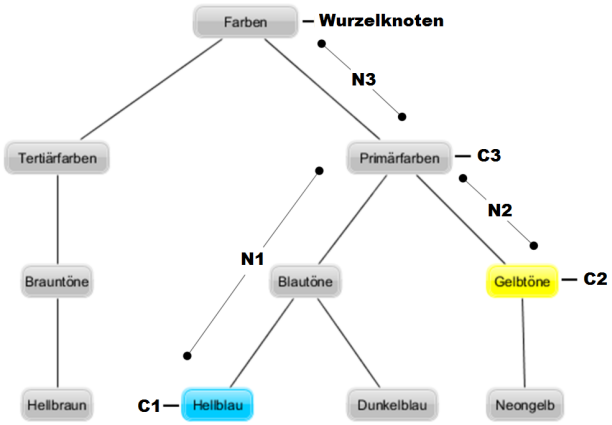


Abb. 1: Distanzbestimmung der Attribute

Visualisierung: Hinweise auf Musterkandidaten

Die Ergebnisse der Ähnlichkeitsanalyse können dann als Graph visualisiert werden. Abbildung 2 zeigt eine Beispielauswertung. Der Übersichtlichkeit halber haben wir die Anfrage auf Basiselemente, welche mit „negativ belegten“ Charaktereigenschaften assoziiert und von „weiblichen“ Rollen getragen werden, sowie auf die Kostümeigenschaften „Design“, „Farbe“ und „Zustand“ in der Ähnlichkeitsanalyse beschränkt. Die größte Ähnlichkeit bei den abgefragten Kostümen liegt bei „Unifarben“, „Gold/Silber“ und „Sauber“.

Diese so identifizierten Häufungen bzw. Cluster ähnlicher Attributsausprägungen können als Hinweise auf mögliche Kostümmuster gewertet werden.
Wie die Ergebnisse bewertet werden und ob ein gehäuftes Auftreten ähnlicher Ausprägungen als Kostümmuster bewertet werden kann, bedarf einer weiterführenden Interpretation der Ergebnisse durch einen Domänenexperten.

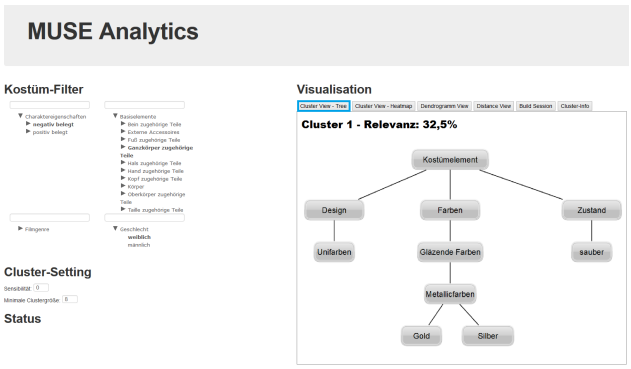


Abb. 2: Visualisierung der gemeinsamen Merkmale

Um die Analyse und Visualisierung einfach zugänglich zu machen, ist das Tool über ein Web Frontend erreichbar und erlaubt über Filtermöglichkeiten und unterschiedliche Visualisierung ein differenziertes Auswerten der Daten. Einen kleinen Ausblick auf die unterschiedlichen Ansätze und Diagrammarten, die das Tool unterstützt, soll durch die folgenden Screenshots (Abbildungen 3 und 4) gegeben werden.

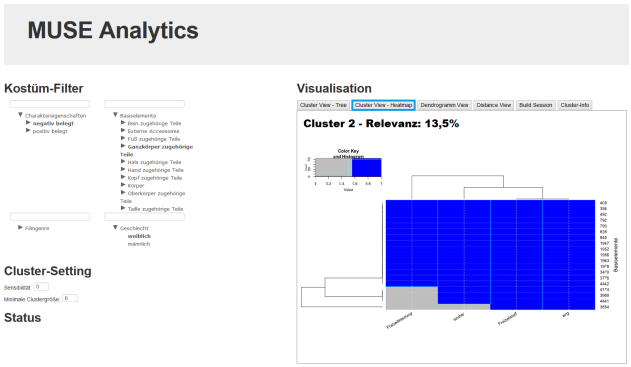


Abb. 3: Web Frontend: Heatmap

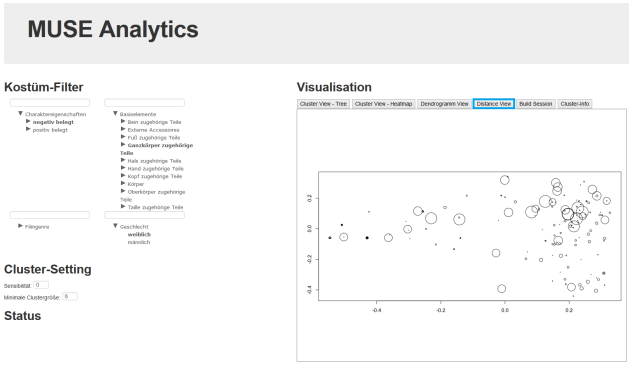


Abb. 4: Web Frontend: Distanzdiagramm

Über das Kostüm hinaus kann dieser Ansatz auch für andere Domänen der Digital Humanities fruchtbar gemacht werden. So zum Beispiel ist der Einsatz bei der Identifikation musikalischer Muster angedacht. Hier wurde bereits mit der Erstellung musikalischer Taxonomien als Grundlage begonnen.

Notes

1. Zur Projektbeschreibung s. auch http://www.iaas.uni-stuttgart.de/forschung/projects/MUSE/ .

Bibliographie

Alexander, Christopher / Ishikawa, Sara / Silverstein, Murray / Jacobson, Max / Fiksdahl-King, Igrid / Angel, Shlomo (1977): *A Pattern Language*. Towns, Buildings, Constructions. Oxford: Oxford University Press.

Barzen, Johanna (2013): *Taxonomien kostümrelevanter Parameter.* Annäherung an eine Ontologisierung der Domäne des Filmkostüms. Technischer Bericht Nr. 2013/04, Universität Stuttgart.

Barzen, Johanna / Falkenthal, Michael / Hentschel, Frank / Leymann, Frank (2015): „Musterforschung in den Geisteswissenschaften: Werkzeugumgebung zur Musterextraktion aus Filmkostümen“, in: *Book of Abstracts zur Tagung der Digital Humanities im deutschsprachigen Raum 2015*, Graz 59-64 http://gams.uni-graz.at/o:dhd2015.abstracts-poster [letzter Zugriff 21. Januar 2016].

Fehling, Christoph / Barzen, Johanna / Falkenthal, Michael / Leymann, Frank (2014): „PatternPedia – Collaborative Pattern Identification and Authoring“, in: *Proceedings of PURPLSOC (Pursuit of Pattern Languages for Societal Change)*. The Workshop 2014. Krems 252-284.

Palmer, Martha / Wu, Zhibiao (1994): „Verb Semantics and Lexical Selection“, in: *ACL '94 Proceedings of the 32nd annual meeting of the Association for Computational Linguistics*, Stroudsburg, USA 133-138.

Schmitt, Ingo (2005): *Ähnlichkeitssuche in Multimedia-Datenbanken*. Retrieval, Suchalgorithmen und Anfragebehandlung. München: Oldenbourg Wisschaftsverlag.

Jiang, Jay J. / Conrath, David W. (1997): „Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”, in: *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan.

Lord, Phillip W. / Stevens, Robert D. / Brass, Andrew / Goble, Carole A. (2003): “Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation”, in: *Bioinformatics* 19, 10: 1275-1283.

Das Dortmunder Chat-Korpus in CLARIN-D: Modellierung und Mehrwerte

Beißwenger, Michael
michael.beisswenger@tu-dortmund.de
TU Dortmund, Deutschland

Axel, Herold
herold@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Harald, Lüngen
luengen@ids-mannheim.de
Institut für deutsche Sprache, Mannheim, Deutschland

Angelika, Storrer
astorrer@mail.uni-mannheim.de
Universität Mannheim, Deutschland

Einleitung und Projekthintergrund

Die Kommunikation im Internet bzw. mit sozialen Medien hat in den vergangenen zwei Jahrzehnten in den geisteswissenschaftlichen Disziplinen eine zunehmende Aufmerksamkeit erfahren. Zahlreiche sprach-, sozial- und medienwissenschaftliche Analysen haben die sprachlichen und interaktionalen Besonderheiten bei der Kommunikation in Chats, Foren, Weblogs und sozialen Netzwerken, per SMS und WhatsApp als einen neuen Gegenstand geisteswissenschaftlicher Forschung erschlossen. Durch ihre digitale Verfügbarkeit sind Sprachdaten aus solchen Genres – im Gegensatz etwa zu Aufzeichnungen von Gesprächen – einfach zu gewinnen und für Forschungszwecke speicherbar. Trotzdem gibt es bislang wenige Korpora zur Sprachverwendung in sozialen Medien, die für Analysezwecke im Bereich der Digital Humanities aufbereitet sind und die der Scientific Community zur Nutzung zur Verfügung stehen. Das hat zum einen mit unklaren rechtlichen Rahmenbedingungen in Bezug auf die Nutzung und Bereitstellung digitaler Kommunikationsdaten für Forschungszwecke zu tun, zum anderen mit dem Fehlen geeigneter Standards für die Strukturbeschreibung und linguistische Annotation von Social-Media-Genres sowie der Notwendigkeit, automatische Annotationswerkszeuge für Daten dieses Typs anzupassen.

In unserem Beitrag präsentieren wir Ergebnisse aus dem Projekt „ChatCorpus2CLARIN“, das als Kurationsprojekt der fachspezifischen Arbeitsgruppe F-AG 1 „Deutsche Philologie“[1] . von Mai 2015 bis Februar 2016 vom BMBF gefördert wird. Ziel des Projekts ist es, das *Dortmunder Chat-Korpus*, ein existierendes Korpus zur Sprachverwendung und Sprachvariation in der deutschsprachigen Chat-Kommunikation, in die Korpus-Infrastrukturen der CLARIN-D-Zentren an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) und am Institut für Deutsche Sprache (IDS) Mannheim zu integrieren. Dabei geht es insbesondere um die Herstellung einer Interoperabilität der Zielressource mit Korpora zur gesprochenen und geschriebenen Sprache (DWDS-Korpora, DeReKo, FOLK), die an der BBAW und am IDS bereits vorhanden sind. Die Bereitstellung des Chat-Korpus in CLARIN-D soll einen systematischen, korpusgestützten Vergleich der Sprachverwendung in Chats mit der Sprachverwendung in mündlichen Gesprächen und in redigierten Texten erlauben und der empirischen, sprachdatengestützten Forschung zur Sprache und Interaktion in sozialen Medien somit neue Möglichkeiten eröffnen.

Um Interoperabilität mit existierenden CLARIN-D-Ressourcen herzustellen und es Forscher_innen zu ermöglichen, die unterschiedlichen Ressourcen im Forschungsprozess vernetzt zu nutzen, wird das Chat-Korpus bei der Integration unter Rückgriff auf Standards im Bereich der Digital Humanities remodelliert und um zusätzliche linguistische Annotationen erweitert. Der Beitrag beschreibt die Modellierung der Ressource und ihre Integration in CLARIN-D und zeigt, welche Mehrwerte sich für Nutzer des Korpus durch die Integration und die zusätzlichen Annotationen ergeben.

Die Ausgangsressource

Das *Dortmunder Chat-Korpus* (Beißwenger 2013) ist eine Sammlung von Chat-Mitschnitten aus vier verschiedenen Handlungsbereichen (Freizeit, Bildung, Beratung, Medien), die ca. 140.000 Chatter-Beiträge und 1,06 Mio. Token umfasst und die 2002–2008 am Lehrstuhl für Linguistik der deutschen Sprache und Sprachdidaktik der TU Dortmund aufgebaut wurde. Die Daten sind in einem XML-Format repräsentiert, das zentrale Strukturelemente von protokollieren Chatverläufen (sog. ‚Logfiles‘) abbildet, unterschiedliche Typen von Chat-Beiträgen unterscheidet und ausgewählte Stilelemente internetbasierter Kommunikation erfasst. Teile des Korpus werden seit 2005 über die Website http://www.chatkorpus.tu-dortmund.de zusammen mit einem einfachen, Java-basierten Abfragewerkzeug zur Verfügung gestellt. Das Korpus wird in diversen linguistischen und computerlinguistischen Projekten sowie im Bildungskontext (Schule und Hochschule) als Ressource in Forschung und Lehre genutzt.

Interoperabilität durch Anschluss an Standards im Bereich der Digital Humanities

Strukturmodellierung und Repräsentation in TEI

Für die Repräsentation der im Korpus dokumentierten Chat-Verläufe greifen wir auf die Formate der *Text Encoding Initiative* (*TEI)* zurück. In den TEI-Guidelines (TEI-P5) gibt es bislang keine Modelle für die Darstellung von Social-Media-Genres, dafür umfangreiche Module für die Strukturrepräsentation von Textgenres und von transkribierten Gesprächen. Die in den Guidelines vorgesehene Möglichkeit der *customization* macht das Encoding-Framework aber flexibel genug, um es an die Erfordernisse auch von (neuen) Genres anzupassen.

Seit 2013 beschäftigt sich in der TEI eine Special Interest Group (SIG) „Computer-mediated communication“ [2] mit der Entwicklung eines Standards für die Modellierung von Social-Media-Genres (Beißwenger et al. 2012; Chanier et al. 2014; Margareta / Lüngen 2014). Das Projekt greift den aktuellen Stand der in der SIG diskutierten Schemaentwürfe auf, testet diese an den Daten des Chat-Korpus sowie an Ausschnitten ausgewählter weiterer Social-Media-Genres (Wikipedia-Diskussionsseiten, WhatsApp-Dialoge, News-Diskussionen, Tweets) und entwickelt sie weiter. Das dabei entstehende TEI-Schema wird in Form eines ODD[3] dokumentiert und bildet die Grundlage für die TEI-Modellierung des kompletten Korpus. Zugleich wird das ODD, dessen Fertigstellung für Herbst 2015 vorgesehen ist, in die weitere Arbeit der SIG eingespielt.

Linguistische Basisannotation mit „STTS 2.0“

Um die Recherchemöglichkeiten im Korpus zu verbessern, wird der Ausgangsressource eine zusätzliche Annotationsebene hinzugefügt, deren Kern Part-of-speech-Informationen (PoS) bilden. Das im Projekt verwendete PoS-Tagset („STTS 2.0“, Beißwenger et al. 2015) verwendet die Kategorien des *Stuttgart-Tübingen Tagset* (STTS, Schiller et al. 1999) und erweitert diese einerseits um Tags für typische Einheiten bei der schriftlichen Sprachverwendung in Social-Media-Genres (u. a. Emoticons, Hashtags, Adressierungen) sowie um Einheiten für die Darstellung von Phänomenen, die typisch sind für Kontexte informeller, dialogischer Kommunikation (u. a. Abtönungs- und Intensitätspartikeln, Diskursmarker). Die Erweiterungen sind abgestimmt auf Erweiterungen, die

am IDS für die PoS-Annotation des FOLK-Korpus zur gesprochenen Sprache zum Einsatz kommen.

Um die Annotationen nach STTS 2.0 zu erzeugen, wurde das komplette Chat-Korpus 2015 mit einem POS-Tagger annotiert, für den im BMBF-Projekt "Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen" (IDS 2014-2016) neue Taggermodelle speziell für den Umgang mit Social-Media-Genres entwickelt wurden (Horbach et al. 2014). Um das Ergebnis der automatischen Annotation manuell nachzukorrigieren und zusätzlich einzelnen Tokens normalisierte Formen zuzuordnen, wurde das Werkzeug *OrthoNormal* (Schmidt 2012) auf die Bearbeitung von Chat-Daten angepasst.

Zielressource und Mehrwerte

Die Integration in die Infrastrukturen der beteiligten CLARIN-D-Zentren umfasst die Archivierung in den Repositorien an der BBAW und am IDS, die Aufnahme der Metadaten in das Virtual Language Observatory (VLO), die Einbindung der Daten in die korpusübergreifende Suchmaschine *CLARIN Federated Content Search* sowie die Bereitstellung über Webservices.

Die rechtlichen Bedingungen der Bereitstellung werden über ein Rechtsgutachten geklärt. Je nach Ergebnis kommen für die Ressource unterschiedliche Lizenzmodelle in Frage: Als Idealfall wird eine CLARIN-Endnutzer-Lizenz vom Typ PUB („publicly available", Oksanen et al. 2010) angestrebt, gegebenenfalls aber auch der Lizenztyp ACA-NC (akademische, nicht-kommerzielle Nutzung zum vollständigen Kopieren / Download freigegebener Ressourcen) oder, falls erforderlich, eine Beschränkung auf eine Nutzung über eine Korpusrecherchesoftware durch bei CLARIN registrierte Nutzer (Lizenztyp QAO-NC, gemäß Vorschlag in Kupietz / Lüngen 2014).

Nach der Integration wird die Zielressource für Nutzer im Bereich der Digital Humanities gegenüber der Ausgangsressource die folgenden Mehrwerte aufweisen:

- **Erweiterung der Möglichkeiten des Zugriffs und der Durchsuchbarkeit** der Ressource.
- **Interoperabilität auf der Ebene der Dokumentstruktur (TEI):** Durch die Remodellierung in einem TEI-Format wird die Ressource interoperabel mit anderen in TEI repräsentierten Sprachressourcen und Annotations- bzw. Analysewerkzeugen.
- **Linguistische Annotation:** Die Anreicherung um zusätzliche linguistische Basisannotationen wird die Möglichkeiten zur Nutzung der Ressource für die korpusgestützte Sprachanalyse erweitern und anspruchsvollere linguistische Suchanfragen ermöglichen.
- **Interoperabilität auf der Ebene der linguistischen Annotation (STTS):** Durch die Kompatibilität der Part-of-speech-Annotationen mit STTS wird die Ressource interoperabel mit anderen nach STTS annotieren Sprachressourcen.
- **Vernetzung mit Korpusressourcen anderen Typs:** Durch die Integration in CLARIN-D und die genannten Interoperabilitätsmerkmale werden die Möglichkeiten zu einem korpusgestützten Vergleich sprachlicher Besonderheiten im Chat-Korpus mit Korpora gesprochener Sprache und Korpora redigierter Schriftlichkeit verbessert.
- **Verbesserte Auffindbarkeit der Ressource** durch die Bereitstellung standardisierter Metadaten und die Aufnahme in das VLO.

Die Ergebnisse aus dem Projekt können zum gegenwärtigen Zeitpunkt z. T. nur perspektivisch formuliert werden. Zum Termin der Konferenz werden die Projektarbeiten abgeschlossen sein und die Ergebnisse vorliegen.

Notes

1. Für weitere Informationen siehe http://www.clarin-d.de/de/wissenschaftsbereiche/germanistik
2. Sie hierzu die Webseite der TEI unter http://www.tei-c.org/Activities/SIG/CMC/ .
3. Siehe http://www.tei-c.org/Guidelines/Customization/odds.xml .

Bibliographie

Beißwenger, Michael (2013): "Das Dortmunder Chat-Korpus", in: *Zeitschrift für germanistische Linguistik* 41, 1: 161-164. Erweiterte Fassung online: http://tinyurl.com/chatkorpus [letzter Zugriff 18. September 2015].

Beißwenger, Michael / Ermakova, Maria / Geyken, Alexander / Lemnitzer, Lothar / Storrer, Angelika (2012): "A TEI Schema for the Representation of Computer-mediated Communication", in: *Journal of the Text Encoding Initiative (jTEI)* 3. http://jtei.revues.org/476 [letzter Zugriff 18. September 2015].

Beißwenger, Michael / Bartz, Thomas / Storrer, Angelika / Westpfahl, Swantje (2015): *Tagset und Richtlinie für das PoS-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation*. https://sites.google.com/site/empirist2015/home/annotation-guidelines [letzter Zugriff 18. September 2015].

Chanier, Thierry / Poudat, Celine / Sagot, Benoit / Antoniadis, Georges / Wigham, Ciara / Hriba, Linda / Longhi, Julien / Seddah, Djamé (2014): "The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres", in: *Journal of Language Technology and Computational Linguistics* 2: 1-30. http://

www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf [letzter Zugriff 18. September 2015].

Horbach, Andrea / Steffen, Diana / Thater, Stefan / Pinkal, Manfred (2014): "Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication", in: *Proceedings of KONVENS 2014* 171-177.

IDS = Institut für Deutsche Sprache (2014-2016): *Projekt Schreibgebrauch*. Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen http://www.schreibgebrauch.de/index.html .

Kupietz, Marc / Lüngen, Harald (2014): "Recent developments in DeReKo", in: Calzolari, Nicoletta / Choukri, Khalid / Declerck, Thierry / Loftsson, Hrafn / Maegaard, Bente / Mariani, Joseph / Odijk, Jan / Piperidis, Stelios (eds): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Margaretha, Eliza / Lüngen, Harald (2014): "Building Linguistic Corpora from Wikipedia Articles and Discussions", in: *Journal of Language Technology and Computational Linguistics* 2: 59-82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf [letzter Zugriff 18. September 2015].

Oksanen, Ville / Lindén, Krister / Westerlund, Hanna (2010): "Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN", in: *Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*, Malta.

Schmidt, Thomas (2012): "EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language", in: *Proceedings of LREC2012* http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf [letzter Zugriff 18. September 2015].

Schiller, Anne / Teufel, Simone / Stöckert, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Universität Stuttgart: Institut für maschinelle Sprachverarbeitung.

TEI Consortium (eds.) (2007): *TEI P5: Guidelines for Electronic Text Encoding and Interchange* http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 18. September 2015].

Die computergestützte Erschließung und Visualisierung historischer Itinerare

Blank, Daniel
daniel.blank@uni-bamberg.de
Universität Bamberg, Deutschland

Henrich, Andreas
andreas.henrich@uni-bamberg.de
Universität Bamberg, Deutschland

Einleitung und Ziele

Die Itinerarforschung beschäftigt sich mit der Erschließung historischer Straßennetze. Sie differenziert zwei Itinerararten. Als Itinerare werden einerseits historische Reisewege hochstehender Personen und Herrscher bezeichnet. Diese Reisewege wurden und werden meist anhand historischer Dokumente und Urkunden rekonstruiert, die Auskunft darüber geben, zu welcher Zeit sich gewisse Personen an bestimmten Orten aufgehalten haben. Andererseits bezeichnen Itinerare auch Reisewegverläufe, einzeln oder in Form von Sammlungen, die unmittelbar als solche zusammengetragen wurden (Szabó 2009: 85).

Die Erforschung historischer Itinerare ist ein wichtiger Arbeitsschwerpunkt in verschiedenen Wissenschaftsdisziplinen. Dies ist in Teilen dadurch bedingt, dass historische Personen, die die Itinerare entweder direkt erstellt haben oder auf Basis deren Vita die Itinerare durch Dritte erstellt wurden, häufig in verschiedenen Rollen unterwegs waren. So tritt etwa Hieronymus Münzer auf seiner Spanien- und Frankreichreise gleichzeitig als Arzt, Historiker, Kaufmann, Pilger und Geograph in Erscheinung (Hurtienne 2009: 268). Nicht zuletzt deshalb ist die „Altwegeforschung“ ein stark interdisziplinäres Forschungsfeld (Veling 2014). Charakteristisch für die Itinerarforschung ist eine manuell geprägte und zeitaufwändige Arbeitsweise. Ein wesentlicher Aspekt bei der Erschließung ist etwa die Identifizierung der in den Itineraren genannten Orte (Hurtienne 2009: 269).

Der Ansatz, der in dieser Arbeit beschrieben wird, versucht Werkzeuge zu entwerfen, die Forscher_innen in der Itinerar- und Altwegeforschung in verschiedenen Wissenschaftsbereichen unterstützen. Ziel ist es, die zeitaufwändige, manuelle Erschließung der Itinerare effizienter zu gestalten und später auch den Vergleich verschiedener Itinerare im großen Stile zu fördern. Außerdem soll es ermöglicht werden, leichter Fehler und Inkonsistenzen in den Itinerarquellen zu identifizieren. Darüber hinaus soll die Erweiterung von Ortsverzeichnissen, insbesondere um historische Informationen, erleichtert werden. Ortsverzeichnisse, sog. *Gazetteers*, sind häufig unvollständig und lückenhaft, insbesondere wenn es um historische Informationen geht. Ferner beschränken sich historische Gazetteers häufig auf bestimmte geografische Gebiete und/oder Zeitperioden. Für die Anreicherung der Gazetteers stellen Itinerare eine wesentliche Datenbasis dar, aus der

sich computerunterstützt mit Hilfe des hier skizzierten Ansatzes wichtige Informationen ableiten lassen.

Während Blank und Henrich (2015) bereits die grundlegende Idee und eine Abgrenzung gegenüber verwandten, technischen Arbeiten im Geografischen Information Retrieval adressieren, beleuchtet die vorliegende Arbeit insbesondere die Anwendbarkeit und Einsatzmöglichkeiten des Ansatzes.

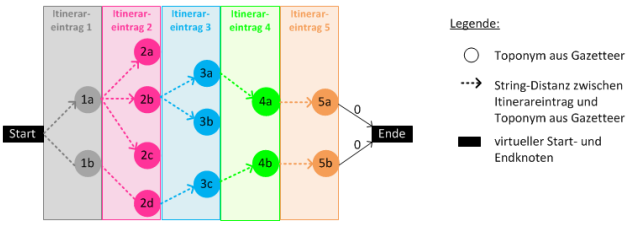
Vorgehen und Systembeschreibung

Die computergestützte Erschließung historischer Itinerare lässt sich in vier Teilschritte zerlegen. Nachdem in Schritt (1) eine optische Zeichenerkennung durchgeführt wird und die Dokumente, die in der Regel als Scans vorliegen, eingelesen werden, muss in Schritt (2) die Struktur der Itinerare erfasst werden. Diese liegen gelegentlich in Tabellenform vor und enthalten beispielsweise wichtige Distanzangaben, die für eine Auflösung von Mehrdeutigkeiten in den Ortsnamen in Schritt (3) ein wesentliches Kriterium sind. Abschließend geht es in Schritt (4) darum, die exakten Wegeverläufe im Gelände zu rekonstruieren.

Diese Arbeit fokussiert auf Schritt (3). Grundlage des Verfahrens sind Itinerare, die neben potentiell mehrdeutigen bzw. in historischer Schreibweise enthaltenen Toponymen auch geografische Distanzen zwischen den einzelnen Wegpunkten erfassen. Distanzen können explizit (s. Spalte *Eingabe* in der folgenden Tabelle; Zahlen entsprechen Meilen) oder implizit zum Beispiel durch eine Angabe von Tagesetappen vorhanden sein.

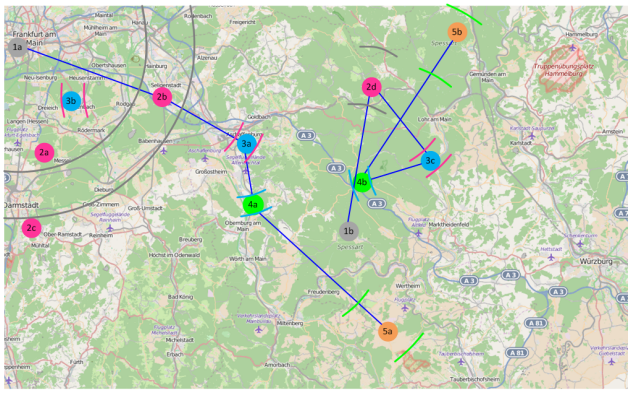
Eingabe	Ergebnis	Korrekte Lösung	Distanz
Franckhfurt, -	Frankel	Frankfurt am Main	66,2 km
Aschenburg, 5	Scheringen	Aschaffenburg	57,3 km
Mildtenburg, 4	Kaltenbrunn	Miltenberg	9,1 km
Bischoffhaim, 4	Tauberbischofsheim	*Tauberbischofsheim*	0 km
Würtzburg, 4	Würzburg	*Würzburg*	0 km
Detelbach, 4	Dettelbach	*Dettelbach*	0 km
Haßfurt, 5	Westheim bei Haßfurt	Haßfurt	6,5 km
Bamberg, 4	Bischberg	Bamberg	5,1 km
Liechtenfelß, 4	Steinfeld	Lichtenfels	19,9 km
Khulmbach, 4	Kulmbach	*Kulmbach*	0 km

Das Verfahren generiert einen Entscheidungsgraph mit einem virtuellen Start- und Zielknoten. Ein solcher Graph ist exemplarisch in der folgenden Abbildung für ein fiktives Itinerar mit fünf Wegpunkten skizziert.



Für jedes Toponym des Itinerars wird zunächst die minimale Distanz zu den Toponymen des deutschen Teils des Geonames-Gazetteers (Wick 2005-2016) berechnet. Dabei können Distanzen verwendet werden, die rein syntaktisch Zeichenketten vergleichen; auch phonetische oder semantische Distanzen sind denkbar (alle im Folgenden vereinfachend als String-Distanzen bezeichnet). Auf den so ermittelten, minimalen String-Distanzwert pro Toponym des Itinerars wird ein Delta addiert, um pro Itinerareintrag einen Schwellwert zu erhalten, mit dessen Hilfe eine Kandidatenmenge aus der Menge der Toponyme des Gazetteers identifiziert werden kann. Die String-Distanz aller Toponyme der Kandidatenmenge muss kleiner oder gleich dem Schwellwert sein. Alle Toponyme der Kandidatenmenge werden anschließend nach Anwendung diverser Filter als Knoten in den Entscheidungsgraph aufgenommen. Als Kantengewicht wird die String-Distanz zwischen dem jeweiligen Toponym des Itinerars und dem des Gazetteers erfasst. Dieser Verarbeitungsschritt wird für alle Toponyme des Itinerars wiederholt, sodass ein Graph wie in obiger Abbildung entstehen kann. Abschließend können die gemäß String-Distanz kürzesten Wege vom virtuellen Start- zum Endknoten und damit die Toponyme des Gazetteers mit den in Summe geringsten String-Distanzen zu den Itinerareinträgen ermittelt werden.

Als Filter sind verschiedene Kriterien denkbar. Ein erstes Kriterium ist der bereits genannte Schwellwert der String-Distanz, mit dessen Hilfe Toponyme des Gazetteers gefiltert werden. Ein zweites Kriterium ist die geografische Distanz. Hierzu wird die Entfernungsangabe im Itinerar in einen Kilometerkorridor transformiert. Die Nutzer_innen der Anwendung können die Breite des Korridors vorgeben bzw. diese je nach Kontext festgelegen. Der Korridor ist in der folgenden, fiktiven Abbildung durch Sphärenschalen angedeutet.



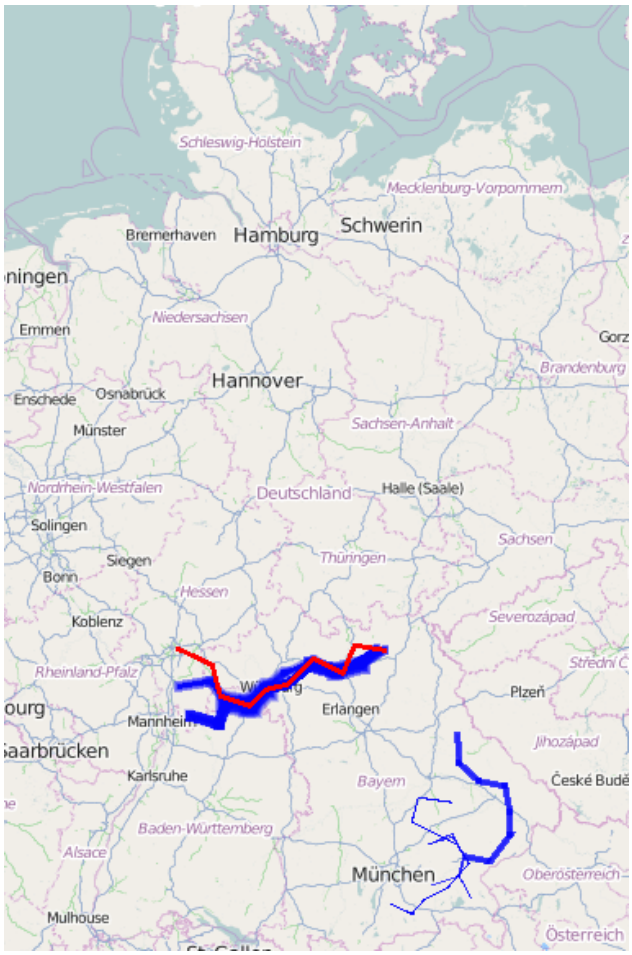
Die Orte 2a und 2c scheiden als nachfolgende Etappenziele von Ort 1a aus, da sie nicht innerhalb der grauen Sphärenschale um Ort 1a liegen. Nur Toponym 2b qualifiziert sich für die weitere Analyse, da es innerhalb liegt. Ein drittes Filter-Kriterium ist die Gerichtetheit der Wegverbindungen. Indem an jedem Zwischenstopp ein Peilungswinkel ermittelt wird, können Orte und Wegverläufe ausgeschlossen werden, die der Gerichtetheit eines Itinerars widersprechen. Somit würde unter Umständen von Ort 2b ausgehend das Toponym 3b verworfen und nur 3a weiter betrachtet, wenn nur hier die geforderte Gerichtetheit gegeben ist. Weitere Kriterien ermöglichen es, ganze Pfade herauszufiltern. Wenn beispielsweise die Vertauschung zweier aufeinanderfolgender oder beliebiger Wegpunkte dazu führt, dass sich die Gesamtdistanz vom Start- zum Zielpunkt des Itinerars reduziert, kann dies ein Indiz für eine inkorrekte Lösung sein, die der Algorithmus verwerfen kann. Ein solches Beispiel ist in der obigen Abbildung durch den Wegverlauf, der bei Toponym 1b beginnt, angedeutet.

Systemverwendung

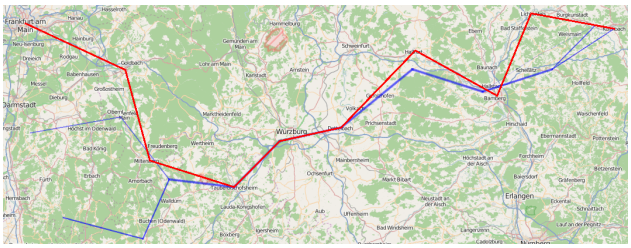
Die Evaluation basiert zunächst auf dem Itinerar Jörg Gails aus dem Jahr 1563 (Krüger 1974). Dies ist „der erste selbständig gedruckte Reiseführer des deutschen Schrifttums“ (Krüger 1974: 1). Erste Analysen basieren dabei auf einem Wegverlauf mit zehn Zwischenstopps aus Route 1 des Itinerars (s. obige Tabelle). Dabei findet die Jaro String-Distanz Anwendung (Winkler 1990). Geografische Distanzfilter werden definiert, indem von einem typischen Umwegfaktor von 1,2 ausgegangen wird. Der Radius der Sphären beträgt 6000 bzw. 8000 Meter, sodass eine mittelalterliche deutsche Meile zwischen 7200 und 9600 Metern modelliert wird (inspiriert durch den Wikipedia-Eintrag 'Meile' , 25.11.2015). Der Winkel für die Gerichtetheit beträgt 90 Grad. Den String-Distanzschwellwert setzten wir auf 0,24. Dieser ist grob per Hand abgestimmt, alle anderen Parameter wurden initial festgelegt und verblieben unverändert.

Die oben dargestellte Tabelle zeigt die durch den Algorithmus ermittelte beste Route. Kursiv dargestellte Toponyme (4/10 = 40%) werden korrekt identifiziert. Im Gazetteer gefundene Toponyme werden aufgetrennt, weil z. B. Frankfurt, der Ausgangspunkt der Route, als *Frankfurt am Main* im Gazetteer enthalten ist. Partielle Übereinstimmungen erhalten keinen Malus. *Frankfurt am Main* soll gleichberechtigt zu *Frankfurt* behandelt werden, da etwa auch *Frankfurt* allein als Toponym im Gazetteer enthalten ist, in Form eines kleinen Dorfs weit von der Route entfernt. Eine Konsequenz dieser Entscheidung ist, dass eine fünfte Übereinstimmung durch *Westheim bei Haßfurt* verhindert wird, das den Vorzug vor *Haßfurt* erhält. Es kann außerdem festgestellt werden, dass einige der nicht korrekt identifizierten Toponyme geografisch sehr nah an den tatsächlichen Orten liegen. Dies ist anhand der Entfernungsangaben in der letzten Spalte der obigen Tabelle ersichtlich. Es zeigt sich auch, dass es schwierig ist, den korrekten Ausgangspunkt der Route zu finden. Nur zwei der zehn Ortsnennungen des Itinerars sind als eindeutiger Eintrag im Gazetteer vorhanden (*Haßfurt* und *Bamberg*). Die besondere Behandlung solch eindeutiger Orte wird in zukünftigen Arbeiten in das Verfahren integriert, um die Qualität weiter zu verbessern.

Die folgende Abbildung zeigt eine Visualisierung der 30 besten Routen, die der Algorithmus findet. Dabei führen Mehrfachnennungen zu dickeren blauen Linien. Rot dargestellt ist die korrekte Lösung. Wegpunkte sind stets durch gerade Linien verbunden. Es lassen sich grob zwei geografische Regionen identifizieren, in denen der Algorithmus die korrekte Lösung vermutet.



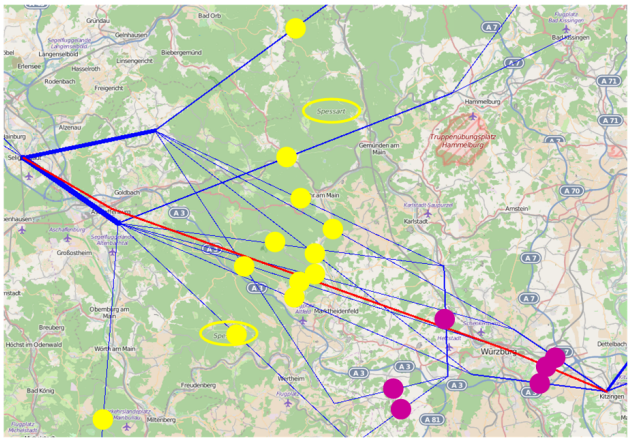
Eine Überlagerung der fünf besten Vorschläge zeigt die nachfolgende Abbildung. Der Fokus liegt nun auf nur einer geografischen Region. Hierbei ist zu erkennen, dass die Vorschläge (blau) nahe an der tatsächlichen Route (rot) liegen.



Exemplarisch soll nun gezeigt werden, wie die Techniken auch im Bereich Visual Analytics und im Speziellen beim Auffinden von Auffälligkeiten und Fehlern verwendet werden können. Hierzu wird ein Ausschnitt einer Route des Brügger Itinerars mit vier Wegpunkten und einer textuellen Anmerkung *Per nemora* analysiert (vgl. Hamy 1908: 170):

Eingabe	korrektes Ergebnis
Zelghenstat, -	Seligenstadt
Asscaffengherne, III	Aschaffenburg
Witsbuerch, IX	Würzburg
Per nemora	*durch den Wald*
Litsinghen, III	Kitzingen

Die Anmerkung *Per nemora* ist zwischen Würzburg und Kitzingen angesiedelt. Die hier dargestellte Karte, unter Verwendung der Levenshtein-Distanz (Levenshtein 1966) mit einem Schwellwert von 0,2, gibt Hinweise, dass in direkter Umgebung Würzburgs kein größeres bewaldetes Gebiet zu finden ist (magenta farbige Kreise: Kreismittelpunkt ist jeweils der Mittelpunkt einer Strecke zwischen möglichen Interpretationen von Witsbuerch und Litsinghen).



Eine Positionierung der Anmerkung eine Etappe früher zwischen Asscaffengherne und Witsbuerch (gelbe Kreise) deckt auf, dass mit der Bezeichnung *Per nemora* wohl die Wegstrecke durch den Spessart gemeint ist. Die Tatsache, dass diese Anmerkung im Itinerar verzeichnet ist, in dem Anmerkungen eher selten zu finden sind, deutet auf eine Besonderheit hin. Müller (1907: 175) schreibt über den entsprechenden Wegabschnitt, dass der Spessart im 15. Jahrhundert vom Verkehr gemieden wurde, obwohl die direkte Verbindung zwischen Nürnberg und Frankfurt durch den Spessart verlief.

Bibliographie

Blank, Daniel / Henrich, Andreas (2015): "Geocoding place names from historic route descriptions", in: *Proceedings of the 9th ACM Workshop on Geographic Information Retrieval, Paris, France*.

Hamy, Ernest-Théodore (1908): *Le Livre de la description des pays de Gilles le Bouvier, dit Berry*. Paris: Ernest Leroux.

Hurtienne, René (2009): “Ein Gelehrter und sein Text: Zur Gesamtedition des Reiseberichts von Dr. Hieronymus Münzer, 1494/95 (Clm 431)“, in: Neuhaus, Helmut (ed.): *Erlanger Editionen*. Grundlagenforschung durch Quelleneditionen (= Erlanger Studien zur Geschichte 8). Erlangen / Jena: Palm & Enke 255-272.

Krüger, Herbert (1974): *Das älteste deutsche Routenhandbuch.* Jörg Gails Raißbüchlin. Graz: Akademische Druck- und Verlagsanstalt.

Levenshtein, Vladimir I. (1966): "Binary codes capable of correcting deletions, insertions, and reversals”, in: *Soviet Physics Doklady* 10, 8: 707-710.

Müller, Johannes (1907): “Geleitswesen und Güterverkehr zwischen Nürnberg und Frankfurt im 15. Jahrhundert", in: *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte*5: 173-196, 361-409.

Szabó, Thomas (2009): “Die Itinerarforschung als Methode zur Erschließung des mittelalterlichen Straßennetzes“, in: Szabó, Thomas (ed.): *Die Welt der europäischen Straßen.*Von der Antike bis in die Frühe Neuzeit. Köln / Weimar / Wien: Böhlau Verlag 85-96.

Veling, Alexander (2014): "Altwegeforschung: Forschungsstand und Methoden", in: *aventinus - Geschichtswissenschaften im Internet*44: http://www.aventinus-online.de/no_cache/persistent/artikel/9847/ [letzter Zugriff 13. Oktober 15].

Winkler, William E. (1990): “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage”, in: *Proceedings of the Section on Survey Research Methods, American Statistical Association*: 354-359.

Wick, Marc (ed.) (2005-2016): *GeoNames*. Unxos GmbH, Switzerland http://www.geonames.org/ [letzter Zugriff 30. Januar 2016].

Wikipedia (25.11.2015) „Meile“, in: *Wikipedia*. Die freie Enzyklopädie https://de.wikipedia.org/wiki/Meile [letzter Zugriff 30. Januar 2016].

Dramenwerkbank - Automatische Sprachverarbeitung zur Analyse von Figurenrede

Blessing, Andre
andre.blessing@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Bockwinkel, Peggy
peggy.bockwinkel@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Willand, Marcus
Marcus.Willand@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

Einleitung

In diesem Beitrag stellen wir erste Einsichten aus einer quantitativen Analyse von Dramen vor, sowie unsere Konzeption für eine darauf aufbauende interaktive Werkbank, die einen Anstoß für eine Diskussion zur Tool-Unterstützung quantitativer Dramenanalyse geben soll. Die Werkbank unterstützt interessierte Forscherinnen und Forscher beim Einlesen von Dramen aus TEI-basierten Quellen und befindet sich noch in Entwicklung [1] . Neben den in Dramen schon explizit kodierten strukturellen Informationen (Wer spricht was?) stellt die Werkbank insbesondere Möglichkeiten zur Verfügung mit Werkzeugen zur maschinellen Sprachverarbeitung auch den Inhalt der Figurenrede zu analysieren. Inspektions- und Aggregationswerkzeuge und -sichten erlauben auch die Analyse größerer Korpora.

Um die Anwendungsgebiete der Werkbank aufzuzeigen, skizzieren wir # anhand einer *Pilotstudie* zur Analyse des Verhältnisses von dramatischer Figur zur dramatischen Handlung # den Problemhorizont quantitativer Literaturwissenschaft. Dabei interessieren uns insbesondere diese Fragen: Gibt es einen Zusammenhang zwischen angenommenen prototypischen Rollen (Protagonist, Intrigant, König usw.) und Länge bzw. Häufigkeit der Redebeiträge oder der Referenz auf die Figur? Wird über bestimmte Figuren(-rollen) auf bestimmte Arten gesprochen (abwertend / aufwertend, ...)? Gibt es Figuren(-rollen)konstellationen, die häufig kookurrieren, und zwar in Bezug auf ihren eigenen Rede- und Bühnenbeitrag als auch im Bezug auf die Referenzen auf die Figuren?

Dramenanalyse: Basics

Dramentexte unterscheiden sich insbesondere durch zwei zusammenhängende Eigenschaften von Prosatexten: a) Dramatische Texte sind im Gegensatz zu vielen anderen Textsorten auf allen Ebenen (Akt- bis Redefolge) ausgesprochen gut strukturiert und ermöglichen somit eine verhältnismäßig unaufwändige Datenerhebung. Die Kehrseite der guten Strukturiertheit ist dass dramatische Texte damit nicht dem Prototyp eines Textes entsprechen, wie er von vielen Werkzeugen zur Sprachverarbeitung angenommen wird. Die maschinelle Sprachverarbeitung auf dramatischen Texten ist damit nicht durch existierende Werkzeuge „out of the box“ zu leisten. b) Die dramatischen Figuren sprechen *unvermittelt*. Unterscheidungen zwischen Erzähler- und Figurenrede und -denken spielen in Dramen keine Rolle. Während Ansätze der Stilometrie, das Figurensignal vom Erzähler- und jenes wiederum vom Gattungssignal zu trennen (Jannidis 2014), noch in den Kinderschuhen stecken, muss sich die (teil-)automatische quantitative Dramenanalyse diesen interpretativen Problemen nicht stellen. Sie hat vor allem *technisch- methodische* Probleme zu lösen: a) Erfassung und Einlesen der Daten und b) (teil-)automatische Textanalyse in Dramen. Zu letzterem gehört auch der adäquate Einsatz von interpretierbaren Maßen und transparenten Verfahren sowie visuellen Repräsentationen von Ergebnissen.

Erfassung und Einlesen der Daten: TEI-Integration

Eine automatisierte Erfassung der Oberflächenstruktur inklusiver aller relevanten Metadaten dramatischer Texte ist die Grundvoraussatzung einer quantitativen Textanalyse im oben genannten Sinne. TEI / XML ist als Standard etabliert, um Texte und Korpora möglichst genau entsprechend der/einer gedruckten Edition digital zu kodieren (cf. TextGrid; DTA). Insbesondere erlaubt TEI auch die Kodierung von Seitenzahlen, Formatierungen, Zeilenumbrüchen, Kopf- und Fußzeilen und vieles mehr, was über den reinen Textinhalt hinausgeht.

Wie Trilcke et al. (2015) auch schon festgestellt haben, ist die Extraktion der inhaltlichen Textstruktur aus den TEI-Daten keineswegs trivial. Für Netzwerkanalyse ist die eindeutige Identifizierung von Figuren besonders relevant, für eine (maschinelle, computergestützte) Analyse des Inhaltes und der Häufigkeit der Figurenrede kommen o.g. Formatierungsmarkierungen noch als Herausforderung hinzu. In unserer Werkbank bieten wir einen Plausibilitätscheck an, der es erlaubt, Fehler im Importprozess (die sowohl durch Fehlannahmen im Importmodul als auch durch Fehlkodierungen in den Quelldaten verursacht werden können) direkt zu erkennen und zu beheben. Einmal identifizierte und behobene Fehler fließen in die Quelldaten zurück.

(Automatische) Textanalyse in Dramen

In den bereits existierenden Arbeiten zur Stilometrie auf Dramen werden komplette Dramen verglichen (z. B. durch Vorverarbeitung mit DIGIVOY). Ein differenzierter Vergleich, bei dem einzelne Figuren oder Gruppen von Figuren betrachtet werden, ist so noch nicht möglich gewesen.

Andere Projekte gehen genau den gegenteiligen Weg und verwerfen alle Dialoginhalte und beziehen ihre Netzwerkanalyse nur auf die Interaktion der jeweils in der Szene aktiven Figuren (cf. Trilcke et al.). Uns ist kein verfügbares System bekannt, das diese Lücke schließt und eine inhaltliche Analyse erlaubt, die sowohl die Interaktion der aktiven Figuren als auch deren Redeinhalt einbezieht.

In unserer Werkbank erfolgt die Textanalyse mit computerlinguistischen Werkzeugen, welche durch die CLARIN-D Infrastruktur (Mahlow et al. 2014) bereitgestellt werden. Der Aufbau von Dramen erfordert eine spezielle Herangehensweise bei der Textanalyse, da die in der Computerlinguistik oft getroffene Annahme, dass Texte aus vollständigen und grammatikalisch wohlgeformten Sätzen bestehen, in Dramen nicht zutrifft (wie auch in Texten aus sozialen Medien oder in gesprochener Sprache). Daneben weisen Dramen die oben genannte spezifische Struktur auf, die eine adäquate Vorverarbeitung bedingt. Um eine Verarbeitung mit einer nicht modifizierten CL-Verarbeitungskette zu ermöglichen, wird das Drama vorher in passende Textsegmente zerlegt. Segmente, die zu einem Dialog gehören müssen nach der Verarbeitung wieder der jeweiligen Figur zugeordnet werden. Im Kontext der Figurenanalyse sind insbesondere Eigennamenerkennung und Koreferenzresolution von Interesse. Wenn man den stilometrischen Blick weitet und auch syntaktische Konstruktionen (verwendet eine Figur mehr oder weniger komplexe Satzstruktur?) untersuchen möchte, sind auch andere linguistische Verarbeitungsschritte möglich.

Die Ergebnisse dieser Verarbeitung werden nicht fehlerfrei sein, deswegen bietet die Werkbank Möglichkeiten, die Ergebnisse zu korrigieren. Insbesondere die Zusammenführung von unterschiedlich genannten oder geschriebenen (z. B. „Emilia“ vs. „Emilie“ oder „die Soldaten“ vs. „erster Soldat“) Figuren ist nicht trivial und teilweise nur durch zusätzliches Weltwissen realisierbar. Damit dieser Schritt vereinfacht wird kommt hier ein halb-automatischer Figurenabgleich zum Einsatz. Das überarbeitete und manuell geprüfte Drama kann in einem TEI-konformen Format exportiert werden, damit die so kuratierte Ressource wieder der Community zur Verfügung gestellt werden kann. Linguistische Annotationen, die in TEI nicht direkt repräsentiert werden können, werden in einem geeigneten stand-off-Format exportiert.

Pilotstudie

In einer Pilotstudie haben wir anhand eines einzelnen Dramas exploriert, wie der Zusammenhang von (der zentralen) Dramenfigur zur dramatischen Handlung automatisiert sichtbar gemacht werden kann. Die (zentrale) Stellung im Figurennetzwerk wird dabei nicht (wie in der aktuellen Forschung gängig; vgl. Moretti 2011) lediglich durch häufige Präsenz oder Interaktion auf der Bühne repräsentiert, sondern durch differenziertere Analysen der Figurenaktivität. *Wie häufig* eine Figur spricht, *wie viel* sie spricht und *wie häufig über sie* gesprochen wird, sind dabei die Kerndaten der quantitativen Analyse, auf der weiter vorzustellende Analysen beruhen. Eine manuelle Datenerfassung übersteigt jedoch selbst bei einzelnen Dramen schnell den vom Menschen leistbaren Zeiteinsatz (wie die in Abbildung 1 manuell erstellte Erfassung der Redeteile in *Emilia Galotti* zeigt):

Figuren (Reihenfolge wie im Register)	Token von "x" im Text	Figuren-nennung	Redehäufig-keit der Figur	(Aktivitäts)-Quotient / Figur
Emilia Galotti				
emilia	126	62	64	0,45
tochter	72	72	0	
emilien	8	8	0	
emiliens	2	2	0	
emilie	1	1	0	
		145		
Odoardo Galotti				
odoardo	113	5	108	2,08
vater	47	47	0	
		52		
Claudia Galotti				
claudia	84	11	73	1,01
mutter	61	61	0	
		72		
Hettore Gonzaga, Prinz von Guastalla				
prinz	246	89	157	1,29
prinzen	33	33	0	
		122		
Marinelli, Kammerherr des Prinzen				
marinelli	301	80	221	2,76
			0	

Abb. 1: „ **Token von x**“ = Gesamthäufigkeit der Nennung jeder einzelnen Namensvariante. **Figurennennung** = Nennung der Namensvarianten in der Rede anderer Figuren. **Redehäufigkeit** = Wie oft spricht eine Figur. **Gesamtzahl der Wörter…** = Redelänge in Wörtern. **(Aktivitäts)Quotient** = Summe der Redehäufigkeit geteilt durch die Summe der Figurennennung: X > 1 = Aktiv (Redet häufiger als über sie geredet wird); X < 1 = Passiv.

NLP-Unterstützte Analysemöglichkeiten in Dramen

Die Kombination von in Dramen vorhandenen strukturellen Informationen und durch automatische Verarbeitung ermittelte inhaltlich-semantische Information erlaubt neue, feinkörnige Analysen von Dramen. Die im Folgenden genannten sollen durch die Werkbank unterstützt werden, entweder durch Integration existierender oder durch Entwicklung neuer Tools.

Oberflächenanalyse der Figuren

Möglich ist eine automatische Auswertung der Figurenreden nach inhaltlichen Kriterien. Ohne Vorwissen bereitstellen zu müssen, lassen sich wichtige Begriffe, durch deren Verwendung sich eine Figur von anderen unterscheidet, mit Verfahren wie TF*IDF ermitteln und z. B. als Tabelle oder als Wortwolke darstellen. Komplexere Verfahren wie topic modeling (Blei et al. 2003) oder Wortfeldanalysen können natürlich auch auf den Redeinhalt einer Person (ggf. auf Akte / Szenen o. ä. eingeschränkt) angewendet werden, erfordern aber zumindest die Einstellung von Parametern (z. B. Anzahl der topics im topic modelling) oder das Spezifizieren von Wortfeldern. Automatische Methoden zur Erweiterung von Wortfeldern (angelehnt an z. B. Query Expansion, vgl. Manning et al. 2008) können diesen Prozess unterstützen und sollen im Rahmen der Werkbank erprobt und integriert werden. Abbildung 2 zeigt eine visuelle Auswertung dieser Analyse.

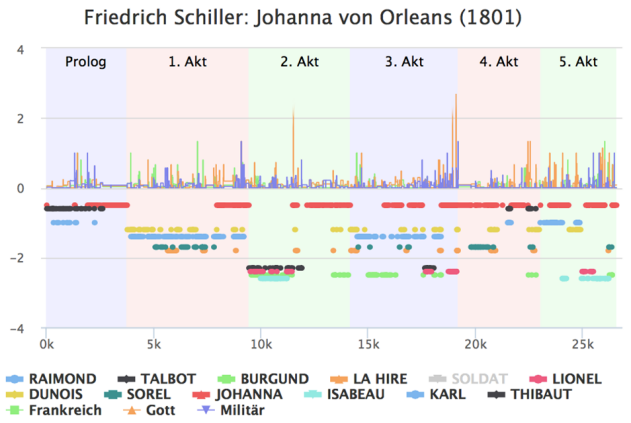


Abb. 2: Strukturelle und inhaltliche Analyse von Schillers Johanna von Orleans. Unten: Figurenaktivität. Oben: Prominenz ausgewählter semantischer Räume in der Figurenrede (Frankreich, Gott, Militär).

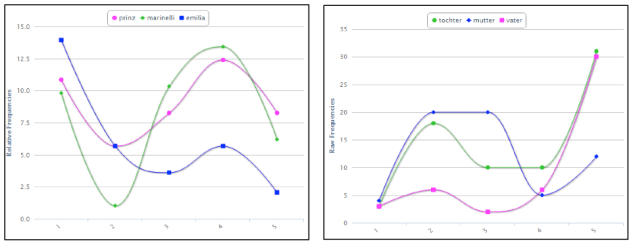


Abb. 3: Anhand der Häufigkeit der Figurennennung („Emilia“ vs. „Tochter“) kann der (bisher kaum erforschte) Diskursverlauf im Sinne einer Unterscheidung in private und öffentliche Konversation sehr gut nachvollzogen werden.

Stilometrische Analysen von Figurenreden

Stilometrische Analysen werden durch eine Schnittstelle ermöglicht, durch die man Figurenrede als Datenstrukturen in R abrufen und dann nach diversen Kriterien untersuchen kann, etwa mit Hilfe von stylo (Eder et al. 2013). Es ließe sich z. B. untersuchen, ob Könige bei Schiller anders sprechen als bei Lessing, oder ob Bürgerfiguren in einem bestimmten Dramenkorpus anders sprechen als Adelsfiguren:

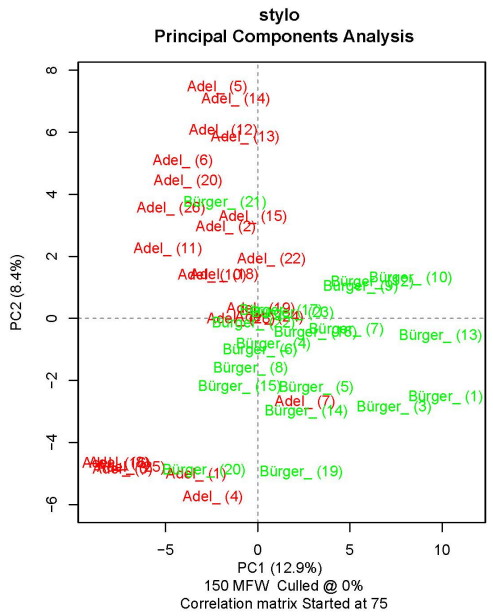


Abb. 4: Figurenreden, extrahiert aus 34 Dramen; nach Standeszugehörigkeit benannt.

Sentiment-Analyse

Durch Methoden aus der Sentiment-Analyse (die zur automatisierten Analyse von Produktreviews eingesetzt wird) ließe sich z. B. analysieren, wie und ob bestimmte Figuren über andere sprechen. Neben positiv / negativ wären auch feinere, dramenspezifische Unterscheidungen denkbar (Feigling, Hahnrei, ...).

Kombination mit Netzwerkanalyse

Die Kombination dieser Techniken mit Netzwerkanalyseverfahren würde es erlauben, im Netzwerk auch Entitäten darzustellen über die geredet wird, ohne dass sie direkt im Drama vorkommen (z. B. Gott), Kanten zwischen Knoten können dann (z. B. durch Farben) auch inhaltliche, relationale Informationen kodieren (X spricht viel / positiv über Y).

Eine Netzwerkdarstellung, in der die Position der Figuren nicht mehr zufällig (oder durch Layout-Algorithmen gesteuert) ist ist ebenfalls denkbar (Abbildung 4). Dabei werden prototypischen Figurenrollen feste Positionen in einem Raster zugewiesen, so dass große Mengen an Netzwerken schnell und direkt verglichen werden können.

Notes

1. http://www.ims.uni-stuttgart.de/short/dramen

Bibliographie

Blei, David / Ng, Andrew Y. / Jordan, Michael I. (2003): „Latent Dirichlet Allocation“, in: *Journal of Machine Learning Research* 3: 993–1022.

Eder, Maciej / Kestemont, Mike / Rybicki, Jan (2013): „Stylometry with R: a suite of tools“, in: *Digital Humanities 2013 Conference Abstracts* 487-89.

Jannidis, Fotis (2014): „Der Autor ganz nah. Autorstil in Stilistik und Stilometrie“, in: Schaffrick, Matthias / Marcus Willand (eds.): *Theorien und Praktiken der Autorschaft*. Berlin: De Gruyter 169-195.

Mahlow, Cerstin / Eckart, Kerstin / Stegmann, Jens / Blessing, Andre / Thiele, Gregor / Gärtner, Markus / Kuhn, Jonas (2014): „Resources, Tools, and Applications at the CLARIN Center Stuttgart“, in: *Akten der 12. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014)* 11-21.

Moretti, Franco (2011): *Network Theory, Plot Analysis*. LiteraryLab Pamphlet 2: http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [letzter Zugriff 20. August 2014].

Manning, Christopher D / Raghavan, Prabhakar / Schütze, Hinrich (2008): *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Trilcke, Peer / Fischer, Frank / Kampkaspar, Dario (2015): „Digital Network Analysis of Dramatic“, in: *Digital Humanities 2015 Conference Abstracts*: http://dh2015.org/abstracts/xml/FISCHER_Frank_Digital_Network_Analysis_of_Dramati/FISCHER_Frank_Digital_Network_Analysis_of_Dramatic_Text.xml [letzter Zugriff 16. Februar 2016].

Digitale Forschungsaktivitäten multilingual: TaDiRAH für die deutschsprachige DH-Community

Borek, Luise
borek@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Schöch, Christof
christof.schoech@uni-wuerzburg.de
Universität Würzburg

Thoden, Klaus
kthoden@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte

Bei der *Taxonomy of Digital Research Activities in the Humanities*, kurz TaDiRAH, handelt es sich um eine anwendungsorientierte Taxonomie, die unter Einbeziehung der Community dazu dient, Ressourcen aus dem Kontext der Digitalen Geisteswissenschaften nach bestimmten Kategorien des Forschungsprozesses zu klassifizieren. Sie trägt dazu bei, diese Ressourcen zu strukturieren, referenzierbar und damit gleichzeitig auffindbar und sichtbar zu machen. Zugleich ist die Taxonomie auch eine Modalität des Nachdenkens darüber, was die digitalen Geisteswissenschaften sind. Als ein Instrument, das in verschiedenen disziplinären Bereichen anwendbar sein soll, wurde die Struktur der Taxonomie in einem *bottom-up*-Verfahren so generisch wie möglich angelegt, ohne dabei jedoch seine Funktionalität zur Unzulänglichkeit einzuschränken. Ermöglicht wird dieser Ansatz z. B. zusätzliche Klassifizierungsparameter, die frei und somit spezifisch erweiterbar sind.

TaDiRAH ist die Initiative einer transatlantischen Kooperation zwischen DiRT und DARIAH-DE . Konzipiert wurde die Taxonomie anhand der Use-Cases dieser beiden Partner: dem Taggen von Tools innerhalb der DiRT Registry sowie der kuratorischen Verschlagwortung bibliographischer Daten in DARIAHs *Doing DH Bibliography* . Darüber hinaus ist TaDiRAH z. B. für das Erfassen von DH Studienangeboten im europäischen DH-Course Registry von DARIAH-EU implementiert. Für die ebenfalls europaweite DiMPO-Initiative stellt die Taxonomie einen Baustein dar, der in das komplexe Gebilde integriert werden kann. Weitere Anwendungsszenarien umfassen u. a. die Verwendung für Surveys (z. B. im Rahmen der Umfrage *practices4humanities*)[1] oder das Klassifizieren von Konferenzbeiträgen, wie in diesem Jahr erstmals bei der DHd-Konferenz erfolgt.

Bei der Konzeption profitierte die Taxonomie von verschiedenen Vorarbeiten in diesem Bereich, die allesamt englischsprachigen Ursprungs sind. Namentlich sei hier insbesondere auf das am King's College London entwickelte arts-humanities.net verwiesen, aus dem schließlich DHCommons hervorging. Auch die Kommunikation im Rahmen von TaDiRAH einschließlich der kollaborativen Phase, während der die Community im Vorfeld des ersten Releases aktiv mit eingebunden wurde, um die spätere Verwendbarkeit zu gewährleisten, nutzte das Englische als *lingua franca*. Entsprechend wird es wenig verwundern, dass auch TaDiRAH zunächst auf Englisch konzeptioniert wurde. Auf datensprachlicher Seite zeigt sich die unter CC-BY-Lizenz stehende Taxonomie ohnehin kompatibel: Neben der Projektwebsite auf GitHub mit vollständiger Dokumentation, Download als SKOS Core sowie einem Issue Tracker bietet eine Instanz des *TemaTres Vocabulary Server* zusätzlich einen SPARQL-Endpoint , der die Anwendung als Linked Open Data erlaubt. TemaTras unterstützt zudem das Implementieren multilingualer Taxonomien. Auf Initiative des argentinischen Projekts "Methodologies on Digital Tools for Research in the Humanities (MHeDI)" konnte zunächst eine spanische Version von TaDiRAH entwickelt werden.[2] Auch eine Ausweitung auf das Serbische ist bereits in Planung.

Inzwischen gibt es eine Vielzahl interessanter *state-of-the-art*-Projekte der deutschsprachigen DH-Community. Sie ist in einem eigenen Verband organisiert, dessen Veranstaltungen sich wachsender Teilnehmerzahlen erfreuen. An verschiedenen Universitäten und DH-Zentren wurden mit großem Erfolg spezifische DH-Angebote eingerichtet. Mit der *Zeitschrift für digitale Geisteswissenschaften* (ZfdG) existiert inzwischen zudem ein eigenes Publikationsorgan für deutschsprachige DH-Beiträge. Betrachtet man die stabile Entwicklung von Version 0.5 von TaDiRAH und ihrer Implementierung in weitere Sprachen und seine Integration in verschiedene EU-Initiativen, so scheint der Zeitpunkt für eine deutsche Version günstig, wenn nicht überfällig. Wir freuen uns daher sehr, der Community eine deutsche Version von TaDiRAH vorstellen zu können. Die Übersetzung profitiert von den Vorarbeiten, die im Rahmen der *practises4humanities*-Umfrage geleistet wurden. Mit dem vorliegenden Poster möchten wir einerseits einen Anstoß zur weiteren Verknüpfung und Erschließung deutschsprachiger DH-Aktivitäten geben, und andererseits die Selbstreflexivität der Disziplin im Abgleich und Austausch mit der internationalen Community fördern. Die Multidisziplinarität ist den Digitalen Geisteswissenschaften ebenso inhärent wie ihre Interdisziplinarität – diesem Sachverhalt möchte TaDiRAH nun auch für die deutschsprachige (Sub-)Community gerecht werden.

Notes

1. Bei "practises4humanities. Wissenschaftliche Forschungspraxis in den Geisteswissenschaften" handelt es sich um eine Online-Umfrage des HCC.lab Berlin in Kooperation mit dem Einstein-Zirkel Digital Humanities Berlin und dem Interdisziplinären Forschungsverbund Digital Humanities in Berlin (if|DH|b). Vgl. den Vortrag von Claudia Müller-Birn im Rahmen dieser Tagung (DHd 2016).
2. Die spanische Übersetzung wurde von Gimena del Rio angefertigt.

Bibliographie

Borek, Luise / Dombrowski, Quinn / Perkins, Jody / Schöch, Christof (2014): "Scholarly primitives revisited: towards a practical taxonomy of digital humanities research activities and objects", in: *Digital Humanities Conference 2014*, Lausanne, Switzerland http://dh2014.org/paper-session-details/ [letzter Zugriff 13. Februar 2016].

centerNet (o. J.): *DH commons*. A collaboration hub http://dhcommons.org/ [letzter Zugriff 13. Februar 2016].

DARIAH-DE (o. J.): *Digital Humanities Course Registry* https://dh-registry.de.dariah.eu/ [letzter Zugriff 13. Februar 2016].

DARIAH-DE (o. J.): *Digital Research Infrastructure for the Arts and Humanities* https://de.dariah.eu/ [letzter Zugriff 13. Februar 2016].

DARIAH-DE (o. J.): *Doing Digital Humanities.* Bibliographie. https://de.dariah.eu/bibliographie [letzter Zugriff 13. Februar 2016].

DiRT (o. J.): *DiRT*. Digital Research Tools http://dirtdirectory.org [letzter Zugriff 13. Februar 2016].

Hughes, Lorna / Constantopoulos, Panos / Dallas, Costis (Im Druck): "Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines", in Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A new companion to digital humanities*. Oxford: Wiley-Blackwell.

NeDiMAH (2012): "Network for Digital Methods in the Arts and Humanities (NeDiMAH)" http://www.esf.org/fileadmin/Public_documents/Publications/nedimah.pdf [letzter Zugriff 13. Februar 2016].

Schöch, Christof (2012): *Doing Digital Humanities*. A DARIAH-DE Bibliography. Göttingen: DARIAH-DE https://www.zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography [letzter Zugriff 13. Februar 2016].

TaDiRAH: *Taxonomy of Digital Research Activities in the Humanities* https://github.com/dhtaxonomy/TaDiRAH [13. Februar 2016].

Visualisierung mittelalterlicher Handschriften im Projekt eCodicology

Busch, Hannah
buschh@uni-trier.de
Universität Trier, Deutschland

Chandna, Swati
swati.chandna@kit.edu
Karlsruher Institut für Technologie, Deutschland

Tonne, Danah
danah.tonne@kit.edu
Karlsruher Institut für Technologie, Deutschland

Celia, Krause
krause@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Philipp, Vanscheidt
hegel@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Schmid, Oliver
oschmid@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Dank Bibliotheken und Archiven konnten große Bestände mittelalterlicher Handschriften über die Jahrhunderte erhalten werden. Zur Beantwortung zentraler Fragestellungen der buchhistorischen Forschung werden die Handschriften analysiert und insbesondere die verschiedenen Bestandteile des Layouts – beispielsweise Seitengröße, Schrift- und Bildraum – vermessen. Die meist manuelle Analyse ist jedoch sehr zeit- und arbeitsintensiv, so dass nur eine geringe Anzahl mittelalterlicher Handschriften auf diese Art und Weise untersucht werden kann. Mit Hilfe digitaler Methoden und Werkzeuge können vorhandene Digitalisate der Handschriftenseiten automatisch oder halbautomatisch ausgewertet werden. Im Gegensatz zum manuellen Ansatz kann in diesem Fall mit einer signifikanten Verbesserung im Hinblick auf Schnelligkeit, Genauigkeit sowie Reproduzierbarkeit gerechnet werden. Bisher mangelt es jedoch an grafischen Oberflächen, die die entstehenden hochdimensionalen Metadaten großer, elektronisch erfasster Bestände dynamisch visualisieren können. Als Mehrwert wird es Geisteswissenschaftlern ermöglicht, Zusammenhänge zwischen Handschriften einfach zu erkennen und neue Erkenntnisse aus den Daten zu gewinnen. Aus diesem Grund entwickelt das vom BMBF geförderte Verbundprojekt „eCodicology" (http://www.ecodicology.org) einen Software Workflow zur automatischen Annotation und Visualisierung von makro- und mikrostrukturellen Layoutmerkmalen mittelalterlicher Handschriften.

In diesem Kontext präsentieren wir das Visualisierungsframework CodiVis, das die Erforschung von Korrelationen im abstrakten Merkmalsraum digitalisierter mittelalterlicher Handschriften vereinfacht und unterstützt. Die Datengrundlage von CodiVis sind mittelalterliche Handschriften, die zwischen dem achten und neunzehnten Jahrhundert in der Bibliothek der Trierer Benediktinerabtei St. Matthias aufbewahrt wurden. Der Bestand wurde im Rahmen des Projektes „Virtuelles Skriptorium St. Matthias" (http://www.stmatthias.uni-trier.de) digitalisiert und mit bibliografischen Metadaten, wie Datierung, Beschreibstoff, Format und inhaltlichen Informationen, in TEI P5 konformen XML-Dateien angereichert. Nach Einspeisung der Digitalisate in das Datenrepositorium CodiStore können mit Hilfe von SWATI (Software Workflow for the Automatic Tagging of Images) verschiedene Layoutmerkmale der

Handschriftenseiten bestimmt sowie die bibliografischen Metadaten extrahiert werden.

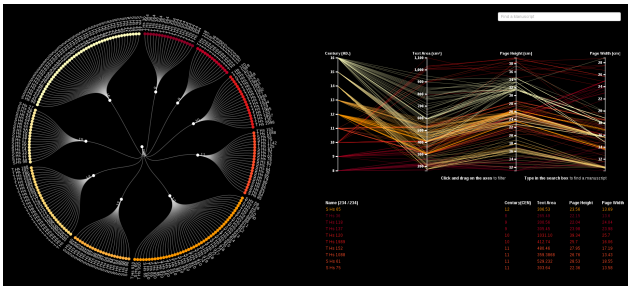


Abb. 1: CodiVis Prototyp, der beide Metadatenarten zur Visualisierung nutzt.

Für einen schnellen Überblick über den gesamten Datenbestand und eine gleichzeitige Darstellung der zugehörigen Handschriftendetails wird eine Kombination zweier Visualisierungsformen angeboten. Auf der linken Seite ist der Bestand mit Hilfe eines radialen Baumdiagramms illustriert, geordnet nach dem Jahrhundert der Entstehung. Auf der rechten Seite werden die extrahierten Merkmale Schriftraum, Seitenhöhe und Seitenbreite mittels paralleler Koordinaten dargestellt. Die verschiedenen Linien repräsentieren dabei die spezifischen Ausprägungen der Layoutmerkmale einzelner Handschriften. Zur Untersuchung der Korrelationen werden Markierungen im Radialbaum automatisch in die Ansicht der parallelen Koordinaten übernommen.

Bisherige Evaluationen des Visualisierungsframeworks zeigen, dass der überwiegende Teil der Nutzer durch die interaktive Zugangsweise erfolgreich Zusammenhänge zwischen ähnlichen Handschriften, fehlerhafte Informationen und Ausreißer erkennen konnte. Darüber hinaus eröffnet CodiVis neue Fragestellungen im Hinblick auf die Visualisierung von Unsicherheiten in den bibliografischen Daten sowie in den automatischen Messungen, die in einem nächsten Schritt zusätzlich zu Visualisierungsmöglichkeiten einzelner Seiten integriert werden. Insgesamt können durch CodiVis neue Möglichkeiten der intuitiven Erkundung historischer Daten aufgezeigt werden.

Die Präsentation gibt einen Einblick in die Entwicklung und Benutzung von CodiVis im Rahmen des Projektes eCodicology mit einem Ausblick auf die mögliche Weiternutzung mit anderen Beständen. Zum Ende der Projektlaufzeit ist die Veröffentlichung als Teil des DARIAH Portals geplant.

Bibliographie

DARIAH: *DARIAH-EU*. Digital Research Infrastructure for the Arts and Humanities http://www.dariah.eu/ [letzter Zugriff 11. Februar 2016].

Technische Universität Darmstadt / KIT Karlsruhe / Universität Trier (2014): *eCodicology*. Algorithmen zum automatischen Tagging mittelalterlicher Handschriften http://www.ecodicology.org [letzter Zugriff 16. Februar 2016].

Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften (2010-2014): Virtuelles Skriptorium St. Matthias. Der mittelalterliche Bibliotheksbestand der Trierer Abtei St. Matthias digital im Netz. Universität Trier: http://www.stmatthias.uni-trier.de/index.php [letzter Zugriff 16. Februar 2016].

Monasterium. Benutzerintegration in einem DH-Großprojekt

Bürgermeister, Martina
martina.buergermeister@uni-graz.at
ZIM, Universität Graz, Österreich

Makowski, Stephan
stephan.makowski@uni-koeln.de
CCeH, Universität Köln, Deutschland

Strecker, Bernhard
bernhard.strecker@uni-koeln.de
CCeH, Universität Köln, Deutschland

Jeller, Daniel
daniel.jeller@icar-us.eu
ICARUS, Wien, Österreich

Schneider, Gerlinde
gerlinde.schneider@uni-graz.at
ZIM, Universität Graz, Österreich

Bigalke, Jan
jan.bigalke@uni-koeln.de
CCeH, Universität Köln, Deutschland

Benutzerintegration in Monasterium

Monasterium ist das größte digitale Urkundenarchiv Europas. Es wurde im Jahr 2001 initiiert und beinhaltet heute etwa eine Million Urkundendatensätze aus dem Mittelalter und der Frühen Neuzeit.

Das Angebot von Monasterium hat sich mittlerweile entscheidend über die reine Präsentation von Urkunden hinaus entwickelt. Monasterium ist ein Portal mit vielen

Datenlieferanten und einer seit Projektstart offenen Datenaufnahmepolitik. Dieser Umstand begünstigte das kontinuierliche Wachstum des Datenbestandes, machte ihn aber gleichzeitig amorph hinsichtlich Erfassungs- und Erschließungstiefe. Aus diesem Grund ist für Monasterium die Integration der Benutzerinnen und Benutzer in der Datenerzeugung besonders wichtig. Die Integration verläuft (1) über kollaborativ-partizipative Nutzerveranstaltungen, sogenannte „MOMathons“, (2) über benutzerfreundliche Editionswerkzeuge wie WYSIWYM-Editoren, (3) über Visualisierung von vernetztem Wissen durch Georeferenzierung und (4) über verbesserten Retrieval durch „Drill Down“-Einsatz. Das Poster gibt einen Einblick in den aktuellen Stand der Entwicklungen.

Im Allgemeinen setzt die Bewältigung des Softwarepakets von Monasterium, genannt MOM-CA (Monasterium Collaborative Archive) eine stabile, effiziente und vor allem skalierbare Softwarearchitektur voraus. Grundlage des Datenbankmanagementsystems ist die Java-basierte XML-Datenbank „eXist“ mit darauf aufsetzendem XRX-Framework. Die Entwicklung von Monasterium erfolgte bis 2014 hauptsächlich am Kölner HKI. Heute übernehmen diese Aufgabe Digitale Geisteswissenschaftler und -innen vom CCeH in Köln, ZIM in Graz, ICARUS in Wien. Der Sourcecode von MOM-CA wird open source auf Github verwaltet und zur Verfügung gestellt.

Benutzerintegration durch „MOMathons“

Monasterium veranstaltet regelmäßig sogenannte „MOMathons“. Dabei haben Teilnehmerinnen und Teilnehmer die Möglichkeit auf Basis des Crowdsourcing-Prinzips bei der Tiefenerschließung von Urkunden mitzuwirken. Dieses kollaborativ eingebrachte Wissen vergrößert das Datenmaterial, das über Monasterium der Öffentlichkeit zugänglich gemacht wird. Die bei diesen Veranstaltungen entstehende Benutzererfahrung ist auch für die Fortentwicklung der Plattform wichtig. Durch „MOMathons“ kann eine direkte Rückmeldung unmittelbar in die Programmiertätigkeit einfließen. So eingebunden, hat jede Nutzerin und jeder Nutzer auf eigene Art und Weise die Möglichkeit zu dem Erfolg des Projektes beizutragen.

Benutzerintegration durch WYSIWYM-Editor

Das Ziel einer virtuellen Editionsumgebung ist seit jeher ein wichtiger Entwicklungsaspekt von Monasterium. Um Wissen in der Webdatenbank strukturiert sammeln und veröffentlichen zu können, mussten die Barrieren des Zugangs für die Benutzerinnen und Benutzer minimiert werden. Aus diesem Grund wurde ein XML-generierender WYSIWYM-Editor entwickelt. Aktuell ist die dritte Editorengeneration umgesetzt. Es handelt sich dabei um eine Javascript-basierte Editionsumgebung, die die Beschreibung der Urkunden intuitiv ermöglicht: Der Editor bietet eine Mischung aus strukturierten Eingabemasken und Freitextfeldern an, die den Zugriff auf kontrollierte Vokabularien sowie Text-Bild-Verknüpfungswerkzeuge bereitstellt. Durch den benutzerfreundlichen Aufbau des Editors können sich so Anwenderinnen und Anwender auf die Erfassung der eigentlichen Inhaltsdaten konzentrieren.

Benutzerintegration durch Georeferencing-Tool

Um den Nutzerinnen und Nutzern weiterhin intuitive Orientierung innerhalb des mitunter sehr schnell anwachsenden Bestandes zu bieten, wurde ein Georeferencing-Tool implementiert.

Die Kartenfunktionalität dient nicht nur einer intuitiven Darstellung von Ausstellungsorten, sondern lässt zudem eine direkte Bearbeitung der mit den Urkunden verknüpften Geoinformationen zu. Das Tool ist in enger Verbindung mit der neuen Suchfunktion in Monasterium zu sehen. So haben die Benutzerin und der Benutzer die Möglichkeit, auf einen Blick für sie räumlich interessante Urkunden wahrnehmen und deren Relevanz einschätzen zu können.

Benutzerintegration durch „Drill Down“

Der in dieser Größe einzigartige archivübergreifende Datenbestand lässt bei Monasterium herkömmliche Such- und Ordnungsfunktionalitäten an ihre Grenzen stoßen. Besonderes Augenmerk wurde und wird deshalb auf eine Erweiterung der Suchfunktion gelegt. Diese wird dahingehend verbessert, dass Treffermengen auf Basis verschiedener Urkunden-Attribute visualisiert und gefiltert werden können. Strukturen und Verbindungen innerhalb des Datenbestandes können besser erfasst und aufgedeckt werden. User werden insofern integriert, als das sie die Datenbank auf gezielte Forschungsfragen durchsuchen können und durch das „Drill Down“ ausschließlich relevante Treffer geliefert bekommen.

Zusammenfassung

Um eine ganzheitliche Integration der Benutzerinnen und Benutzer zur Datenerzeugung anbieten zu können, werden bei der Softwareweiterentwicklung

von Monasterium unterschiedlichste Zugänge berücksichtigt: Für direktes Benutzerfeedback sorgen „MOMathons“. Auf einer inhaltlich textbasierten Ebene wird die Datenerfassung durch WYSIWYM-Editoren erleichtert. Auf einer visuellen Ebene gestaltet die Kartenfunktionalität das Material anschaulicher und auf semantischer Ebene wird die Qualität der Informationsrückgewinnung durch die Granulierung der Suchparameter und entsprechender Suchtreffervisualisierung erhöht. Alle Entwicklungsebenen zusammen genommen lassen Monasterium als eine vielversprechende integrative und interaktive Forschungs-, Kommunikations- und Präsentationsumgebung erscheinen.

Bibliographie

ICARUS = International Centre for Archival Research (2001-2016): *Monasterium.net*. Virtuelles Urkundenarchiv http://icar-us.eu/cooperation/online-portals/monasterium-net .

Booksprint-Projekt: Lehrbuch „Forschungsdaten-management“

Büttner, Stephan
st.buettner@fh-potsdam.de
FH Potsdam, Deutschland

Heger, Martin
martin.heger@fh-potsdam.de
FH Potsdam, Deutschland

Heinrich, Marcus
marcus.heinrich@fh-potsdam.de
FH Potsdam, Deutschland

Keller, Carolin
carolin.keller@fh-potsdam.de
FH Potsdam, Deutschland

Lehmann, Anna
anna.lehmann@fh-potsdam.de
FH Potsdam, Deutschland

Meyer, Michaela
Michaela.meyer@fh-potsdam.de
FH Potsdam, Deutschland

Forschungsdaten sind sowohl in den Natur- als auch in den Geisteswissenschaften das Herzstück computergestützter Forschungsvorhaben (DARIAH-DE 2015). Die Unterschiede zwischen den Geisteswissenschaften und den STM-Fächern sind dabei sehr groß, sowohl was die Definition von Forschungsdaten innerhalb der Communities betrifft, als auch das Verständnis der entsprechenden Forschungsdateninfrastrukturen.

Die Informationswissenschaften fungieren bzw. verstehen sich in den letzten Jahren zunehmend als Mittler zwischen der Informatik einerseits und den Natur- und Geisteswissenschaften andererseits (Balck et al. 2015).

Studierende im Masterstudiengang „Informationswissenschaften“ der FH Potsdam stellen in diesem Poster ein aktuell laufendes kollaboratives „Booksprint-Projekt“ vor, das in einem Bachelor-Projekt im Sommersemester 2015 vorbereitet wurde. Dabei verfasst eine bestimmte Anzahl von Autoren in wenigen Tagen ein Buch in gemeinsamer, eng vernetzter Arbeit. Methodisch bedeutet dies, dass bereits nach wenigen Tagen ein handfestes Ergebnis vorhanden ist und das Problem mit z. B. nicht eingehaltenen Deadlines umgangen wird. Die Studierenden nehmen dafür verschiedene Rollen ein, wie die des Moderators und Begutachters und schaffen somit einen produktiven Rahmen, der die Autoren bereits während der Schreibphase mit konstruktivem Feedback versorgt. Die Projektgruppe orientiert sich dabei u. a. an dem Projekt „Handbuch Digital Humanities“ (Neuschäfer 2015). Unterstützt wird das Projekt vom Open Science Lab der TIB Hannover (Open Science Lab der TIB 2015). Innerhalb von drei Tagen haben 12 Autoren aus den Geistes- und Naturwissenschaften Anfang Januar 2016 gemeinsam ein „Lehrbuch Forschungsdatenmanagement“ geschrieben. Die Zielgruppe des Lehrbuchs sind einerseits Forscherinnen und Forscher, die zuvor noch nicht mit Forschungsdatenmanagement in Berührung kamen, unabhängig von der Disziplin sowie Praktiker (z. B. Informatiker, Datenbibliothekare etc.), die sich mit dem Thema beschäftigen, indem sie z. B. an ihrer Einrichtung Forschungsdatenmanagement betreiben oder / und anwenden.

Das Poster beschreibt anschaulich den Verlauf des Projekts und vermittelt strukturiert aufbereitet verallgemeinerungswürdige Erfahrungen bei kollaborativen Booksprint-Projekten.

Als besonders wichtig stellen sich bereits jetzt heraus:

• Erarbeitung dezidierter Vorgaben zu Zielgruppe, Zitierweise, Didaktik etc.

• Betreuung des Projekts durch „Booksprint Facilitators“,

• Ein intensives Marketing des Projekts, insbesondere innerhalb der verschiedenen Communities sowie in den sozialen Medien.

Bibliographie

Balck, Sandra / Büttner, Stephan / Ducks, Denise / Lehfeld, Ann-Sophie / Schneider, Eva / Vietze, Evelyn (2015): "Mit den Informationswissenschaften von Daten zu Erkenntnissen (Poster)", in: *Tagung der Digital Humanities im deutschsprachigen Raum (DHd 2015)*, Graz.

DARIAH-DE (2015): "Forschungsdaten in DARIAH-DE" https://de.dariah.eu/forschungsdaten [letzter Zugriff 14. Oktober 2015].

Neuschäfer, Markus (2015): *Handbuch Digital Humanities*. Public Beta http://dhd-blog.org/?p=5566 [letzter Zugriff 14.Oktober 2015].

Open Science Lab der TIB (2015): http://blogs.tib.eu/wp/opensciencelab/ [letzter Zugriff 14. Oktober 2015].

"Bleeding Edge" -- Datenmodellierung, Softwareentwicklung und die Freuden und Leiden forschungsgetriebener Entwicklung am Beispiel der Datenbank der Islamic Scientific Manuscript Initiative (ISMI)

Casties, Robert
casties@mpiwg-berlin.mpg.de
MPI für Wissenschaftsgeschichte, Deutschland

Wissenschaftliche Manuskripte des Islamischen Raums

Das ISMI Projekt (Islamic Scientific Manuscripts Initiative) ist ein gemeinsames Projekt des Max-Planck-Instituts für Wissenschaftsgeschichte und des Institute of Islamic Studies der McGill University in Montreal mit dem Ziel Informationen über alle islamischen Manuskripte in den "exakten Wissenschaften" (Astronomie, Mathematik, Optik, Mathematische Geographie, Musik, Mechanik und verwandte Disziplinen) in Arabisch, Persisch, Türkisch und anderen Sprachen aus der Zeit zwischen dem 8. und dem 19. Jahrhundert (CE) zu sammeln.

Im Projekt werden vor allem bibliographische und kodikologische Informationen zu den Manuskripten gesammelt, aber auch Informationen zu Nutzung und Besitz der Manuskripte und damit verbundenen Orten und Personen, die sich aus Kommentaren, Verkaufsbemerkungen und anderen Anmerkungen auf dem Manuskript ergeben können.

Arabische Manuskripte, aber auch andere alte Handschriften stellen besondere Anforderungen an die Datenmodellierung einer bibliographischen Datenbank. Informationen über Autor, Titel und Erstellungsdatum, die Standardfelder eines modernen bibliographischen Eintrags sind oft genug unbekannt. Dazu kommen viele unterschiedliche Schreibweisen des Namens einer Person und viele Fälle gleicher Namen, die zu unterschiedlichen Personen gehören. Es gibt viele Kopien des gleichen Textes von unterschiedlichen Kopisten und eine elaborierte Kultur von Kommentaren und Kommentaren höherer Ordnung.

Zusätzlich zu den bibliographischen Informationen werden auch Digitalisate präsentiert wenn sie vorhanden sind und die Lizenzbedingungen eine öffentliche Präsentation zulassen.

Von Tabellen zu Daten-Graphen

Die ursprüngliche ISMI Datenbank, erstellt von Prof. Jamil Ragep bestand aus einem Satz von relationalen Tabellen in MS-Access. In dieser Struktur gab es bereits die für die Arbeit mit Manuskripten wichtige Unterscheidung von abstraktem Text und der konkreten Manifestation in Form eines Manuskripts durch getrennte Tabellen für Texte und Manuskripte. Diese Unterscheidung erleichtert den Umgang mit vielen verschiedenen Kopien des gleichen Texts und erlaubt es den Blick auf die Beziehungen der Texte untereinander zu richten und beispielsweise durch zusätzliche Spalten Kommentarbeziehungen bis zu dritten Ordnung abzubilden.

Es schloss sich 2006 eine neue konzeptionelle Phase an, in der in Zusammenarbeit mit der IT-Gruppe des MPIWG ein neues Datenmodell entworfen wurde. Das neue Datenmodell orientiert sich an konzeptionellen Objekten, die es teilweise bereits im alten Datenbankschema gab: Texten, Manuskripten und Personen mit ihren jeweiligen Attributen und erweitert es durch weitere Objekttypen und individuelle Relations-Objekte mit frei definierbaren Typen und Attributen.

Das neue Modell sollte in der Software als möglichst flexible Attribut-Graph-Datenbank umgesetzt werden, so dass es jederzeit möglich sein sollte Objekten zusätzliche Attribute zu geben oder zusätzliche Relationen einzuführen, sobald dies während des Prozesses der Dateneingabe sinnvoll erscheint.

Im neuen Modell wurde beispielsweise die feste Relation des Author-ID Feldes der Text-Tabelle zur Autor-Tabelle durch ein Relationsobjekt des Typs "was_created_by" zwischen einem Text-Objekt und einem Personen-Objekt ersetzt.

Neue Möglichkeiten durch ein neues Datenmodell

Die mit dem neuen Datenmodell einhergehende Umstellung des Umgangs mit Daten von der vergleichsweise vertrauten Welt der Tabellen und tabellarischen Daten in eine Welt individueller Objekte und Relationen erweist sich bis heute einerseits als Chance, andererseits aber auch als Herausforderung in technischer und konzeptioneller Hinsicht.

Die Chancen zeigten sich in den Diskussionen über das Datenmodell und während der Dateneingabe. So entstanden neue Ideen wie das Konzept auch falsche Zuschreibungen (Misattribution) abzubilden: Wenn in der Literatur für lange Zeit eine Zuschreibung für die Autorschaft eines bestimmten Textes existiert, die sich im Lauf der aktuellen Forschung als falsch herausstellt, dann ist es wichtig nicht nur die korrigierte Information aufzunehmen und anzuzeigen, sondern auch die als inkorrekt markierte alte Zuschreibung, um Forscher darauf hinzuweisen, dass die alte Zuschreibung bekannt ist und dass sie durch neue Informationen überholt ist.

Herausforderungen durch neue Technologie

Der Preis für das neue Datenmodell war neben der konzeptionellen Arbeit des Umdenkens und der Umstellung der bestehenden Daten zunächst vor allem die Abwesenheit von existierender Software zur effektiven Umsetzung der geplanten Datenstrukturen. Zum Zeitpunkt der Umstellung existierte keine verbreitete Graphendatenbanksoftware, so dass in mehreren Anläufen eine spezifische Datenbank „OpenMind“ und ein webbasiertes Frontend entwickelt wurde. Der Aufwand für die Wartung der Software und die Implementierung neuer Anforderungen steigt jedoch ständig.

Eine weitere Herausforderung, die sich nach der Eingabe grosser Datenmengen (derzeit 4000 Texte in 14000 Manuskripten in 7500 Codices und 2200 Personen) stellt, ist die Suche und Analyse der Daten. Vorhandene Werkzeuge, wie tabellenorientierte Abfragen sind auf den Graphen von Daten-Objekten nicht ohne weiteres anwendbar und es müssen Abfrage-Oberflächen entwickelt werden, die es ermöglichen das Potential der vernetzten Daten auch für die Forschenden nutzbar zu machen.

Neben der eher klassischen Webpräsentation und Browsing-Umgebung werden derzeit verschiedene Graphen-Visualisierungs- und Abfragetools getestet. In einem geplanten Workshop soll eine grösserer Kreis von Fachwissenschaftlern Zugang zur Datenbank und den experimentellen Werkzeugen erhalten um eine weitere Diskussion in Gang zu setzen und Erfahrungen und mögliche Fragestellungen für die weitere Entwicklung zu sammeln.

Bibliographie

Ragep, Jamil F. / Ragep, Sally P. (2008): „The Islamic Scientific Manuscript Initiative (ISMI). Towards a Sociology of the Exact Sciences in Islam“, in: Calvo, Emilia / Comes, Mercè / Puig, Roser / Rius, Monica (eds.): *A Shared Legacy: Islamic Science East and West*. Homage to Professor J. M. Millàs Vallicrosa. Barcelona: University of Barcelona 15–21 https://www.rasi.mcgill.ca/ISMI_SharedLegacy.pdf [letzter Zugriff 17. Februar 2016].

Menschen und Monumente im Fokus. Semantische Modellierung im Baedeker Corpus

Czeitschner, Ulrike
ulrike.czeitschner@oeaw.ac.at
Österreichische Akademie der Wissenschaften, ACDH

Forschungskontext

Im Rahmen des Projekts „travel!digital“ [1] wird erstmals der Versuch unternommen, die vielfältige Terminologie in historischen Reiseführern systematisch zu erschließen. Der Fokus auf *Menschen* und *Monumente* stellt zwei dominante semantische Felder in den Mittelpunkt, die sowohl Schlüsselelemente der Textsorte Reiseführer als auch wesentliche Komponenten kultureller Narrative darstellen. Damit sind Reiseführer keineswegs bloße historische Quellen, sondern vielmehr als bedeutende diskurshistorische Artefakte [2] zu betrachten. Zudem eignet sich das in ihnen enthaltene enzyklopädische Wissen in besonderem Maße für die Datenmodellierung. Vor diesem Hintergrund erstaunt das Fehlen systematischer diachroner Forschung auf diesem Gebiet mindestens genauso wie der eklatante Mangel an digitalen Ressourcen, die für historische Untersuchungen geeignet wären. [3]

Mit dem *Baedeker Corpus*, einer digitalen Sammlung deutschsprachiger Reiseführer aus dem Zeitraum 1875–1914, zielt das Projekt nicht nur darauf ab, wertvolles kulturelles Erbe mit den Methoden der *Digital Humanities* nachhaltig und langfristig sicherzustellen, sondern besonders darauf, semantische Technologien verstärkt zur Erforschung des deutschsprachigen Repertoires

kultureller Diskurse an der Wende vom 19. zum 20. Jahrhundert einzusetzen. Als „komplexe Intertexte“ (vgl. Wierlacher 1997; Koshar 2000), die dominante Diskurse (re)produzieren bzw. (re)konstruieren, stellen Reiseführer „kodifizierte und autorisierte Versionen lokaler Kultur und Geschichte“ dar (vgl. Jaworski / Pritchard 2005). Die in den Reiseführern transportierten zeitgenössischen Lesarten zu Tourismus und kulturellem Erbe sowie Orientalismus und kolonialem Diskurs sind kultur- und diskurshistorisch von besonderem Interesse. Darüber hinaus lassen sich strukturell, linguistisch und insbesondere semantisch erschlossene digitale Reiseführer im Rahmen vergleichender literaturwissenschaftlicher Forschung, der historischen Lexikographie und Linguistik, der historischen Geographie und Kulturanthropologie nutzen.

Das vorliegende Korpus vereint *alle* Erstauflagen zu außereuropäischen Reisezielen aus dem Hause Baedeker, die vor der Zäsur des Ersten Weltkriegs erschienen sind. Es umfasst mehr als 1,5 Mio. *running words* und deckt eine Vielzahl an Regionen ab. [4] Mit Blick auf differenzierte Analysemöglichkeiten stehen neben der linguistischen Basis-Annotation der Volltexte (Lemmatisierung, *Part-of-Speech-Tagging*) der Aufbau kontrollierter Vokabulare und deren Anbindung an LOD-Ressourcen im Mittelpunkt. Die semantischen Repräsentationen werden mithilfe des RDF-basierten *Simple Knowledge Organization System SKOS* und dessen Erweiterung *SKOS-XL* realisiert, mit dem sich auch ambige lexikalische Einheiten und nicht-hierarchische Relationen sinnvoll organisieren lassen. Zur Erstellung der *SKOS*-Repräsentation des *Baedeker Corpus* wird der *OpenSKOS* Editor [5] eingesetzt; eine Entwicklung des *Meertens Institute*, die bereits für *CLAVAS (CLARIN OpenSKOS Vocabulary Service)* verwendet wird, das u. a. die *ISO-639-3 language codes* im *SKOS* Format enthält. Auch die *DARIAH-EU* Arbeitsgruppe *Thesaurus Maintenance* verwendet *OpenSKOS* für eine erste Version der *SKOS*-Repräsentation des in Entwicklung befindlichen *Backbone Thesaurus*. Der *OpenSKOS* Editor sichert daher die Kompatibilität mit aktuellen Standards und Entwicklungen.

bdk:ConceptScheme(s)

Neben Personennamen beinhalten Reiseführer auf Seiten der *Menschen* eine Vielzahl an Gruppenbezeichnungen, die generische Referenzen und als solche, Subjekte charakterisierender Eigenschaftszuschreibungen darstellen (vgl. Schmidt-Brücken 2015). Diese für diskurshistorische Analysen relevanten Belege des Sprachgebrauchs werden jeweils von einem *bdk:Descriptor*, einer Unterklasse von *skos:Concept*, repräsentiert. Ihnen zugeordnet sind einzelne Terme als *skosxl:prefLabel* und *skosxl:altLabel*. Mithilfe der *Properties hasTranslation / isTranslationOf* und *hasVariant / isVariantOf* finden die in den Reiseführern enthaltenen Übersetzungen und Varianten auf Ebene der Terme Berücksichtigung. Auf diese Weise wird der Wortschatz vollständig erfasst und in einem *skos:ConceptScheme* zusammengefasst. Der Strukturierung dieses bdk:ConceptSchemeGroups dienen sechs mit *skos:topConceptOf* zugeordnete Kategorien: 1) allgemeine Sammelbegriffe, 2) ethnisch/nationale Gruppen, 3) geographische Konzepte im weitesten Sinne, 4) Berufsgruppen, wozu auch politische, religiöse und wirtschaftliche Funktionen sowie Lebensstile zählen, 5) Religionsgemeinschaften und 6) soziale Gruppierungen. Aufgenommen werden als eigene Konzepte nicht nur Nomen, sondern auch Adjektive, wobei die entsprechenden Konzepte mit *skos:related* aufeinander bezogen sind. Abbildung 1 listet die *skos:topConcept(s)* und ihre Definitionen auf und gibt einige Beispiele der ihnen zugeordneten Konzepte und Labels sowie deren Beziehungen.

bdk:ConceptSchemeGroups skos:topConceptOf	bdk:Descriptor rdfs:label	skosxl:prefLabel literalForm	skosxl:altLabel literalForm
collective *generic term for a group of people*	bdk:Concept/00005 Ausländer @de bdk:Concept/00017 Bevölkerung @de bdk:Concept/00038 einheimisch @de	bdk:Term/00005 Ausländer @de bdk:Term/00017 Bevölkerung @de bdk:Term/00038 einheimisch @de	
ethnicNational *name of an ethnic or national community*	bdk:Concept/00113 Ägypter @de bdk:Concept/00115 ägyptisch @de ... 'Aeneze @de	bdk:Term/00113 Ägypter @de *hasTranslation* *hasVariant* bdk:Term/00113-en Egyptian @en *isTranslationOf* bdk:Term/00115 ägyptisch @de *hasVariant* ... 'Aeneze @de *hasVariant*	bdk:Term/00114 Aegypter @de *isVariantOf* bdk:Term/00116 aegyptisch @de *isVariantOf* ... 'Aenezebeduine @de *isVariantOf*
geographic *(more or less) geographical concept*	... Asiate @de ... Orientale @de ... orientalisch @de ... Damascener @de	... Asiate @de ... Orientale @de ... orientalisch @de ... Damascener @de	
profession *profession, political, religious, economic role, style of living*	... Kaufmann @de ... Gouverneur @de ... Priester @de ... Bauer @de ... Nomade @de	... Kaufmann @de ... Gouverneur @de ... Priester @de ... Bauer @de ... Nomade @de	... Kaufleute @de ... Pfarrer @de ... Farmer @de
religious *name of a religious community*	... Muslim @de ... buddhistisch @de	... Muslim @de *hasTranslation* ... Muslim @en *isTranslationOf* ... buddhistisch @de	... Mohammedaner @de ... Muselman @de
social *name of a social group*	... Adel @de ... Bettler @de ... Brahmane @de ... proletarisch @de	... Adel @de ... Bettler@de ... Brahmane @de ... proletarisch @de	

***Baedeker Group Taxonomy*. Farbliche Hervorhebungen Spalte 2: Konzepte, die in Beziehung zu anderen Konzepten stehen; farbliche Hervorhebungen kursiv Spalte 3 und 4: *Properties* für Übersetzungen und Varianten auf Ebene der Terme.**

Ein ähnlich vielfältiges Bild ergibt sich im Bereich der Monumente und Sehenswürdigkeiten, die als Gegenstand wertender Beschreibung und Einordnung eine zentrale Rolle in Reiseführern einnehmen. Aufgrund ihrer Vielzahl werden zunächst nur jene Sehenswürdigkeiten strukturiert, die mit Baedeker-Sternen als besonders sehenswert gekennzeichnet sind. Das Spektrum reicht von Architektur [6] und Kunst [7] bis zu Unterkünften, Landschaften und atemberaubenden Aussichten. Diese Kategorien strukturieren jeweils als *skos:topConcept(s)* das bdk:ConceptSchemeMonuments. Derzeit wird an einer Lösung gearbeitet, die es erlaubt, zwischen profaner

und sakraler Kunst und Architektur zu unterscheiden und, sofern sakral, die jeweilige Religion zuzuordnen.

Ausblick

Die linguistische Annotation des *Baedeker Corpus* und die Erstellung der vorgestellten kontrollierten Vokabulare stellen die Voraussetzung für eine weitergehende systematische Analyse der Textsorte Reiseführer dar. Neben der Identifikation der in den untersuchten Baedeker-Bänden genannten historischen Persönlichkeiten mithilfe der *Virtual International Authority Files* und der *Deutschen Biographie* erscheint insbesondere die Anbindung der hier beschriebenen Taxonomien an Ressourcen wie die *DBpedia*, den *GESIS Thesaurus Sozialwissenschaften*, den *AAT Art & Architecture Thesaurus* des *Getty Research Institute* oder den *UNESCO Thesaurus* erfolgversprechend, zumal die Verschränkung der lexikalischen Bestandsaufnahme mit der Kontextualisierung dieser externen Quellen neue Perspektiven auf den Text zu eröffnen vermag. Vor allem aber kann das hier vorgestellte Datenmodell die granulare Analyse jener semantischen Komponenten unterstützen, die das Sprechen über sowohl das ‚Fremde' als auch das ‚Eigene' bestimmen, und erschließt somit mit rezenten Technologien die Funktionsweise eines Diskurses, dessen Wirkungsgrad bis heute weit über das Feld der Reiseliteratur hinausreicht.

Die digitalen Texte inklusive der Faksimiles und die SKOS-Vokabulare werden zu Projektende in Form einer Web-Applikation [8] zur Verfügung gestellt. Die Navigations- und Abfragemöglichkeiten sowohl in den Volltexten als auch der linguistischen Annotation werden ergänzt durch Register der Wortformen, der Lemmata sowie der semantischen Komponenten. Die vorgestellten SKOS-Vokabulare zu *Menschen* und *Monumenten* fungieren zudem als Verbindung der konkreten Belegstellen im *Baedeker Corpus* und externen LOD-Ressourcen.

Notes

1. Das Projekt „travel!digital. Exploring *People* and *Monuments* in Baedeker Guidebooks (1875–1914)" wird im Rahmen der Plattform *Digital Humanities Austria* gefördert.
2. Dominique Maingueneau (2014: 437) nennt Reiseführer explizit als Vertreter sogenannter *Diskursgenres*, „[…] das heißt, soziohistorisch variierende Kommunikationsdispositive".
3. Gründe dafür finden sich u. a. darin, dass die Textwissenschaften in der Vergangenheit stets der Reiseliteratur den Vorrang gegenüber den Gebrauchstextsorten eingeräumt haben. Koshar, der das Fehlen einer allgemeinen Geschichte des Genres beklagt, verweist auf den schlechten Ruf der Reiseführer und bringt dies u. a. damit in Zusammenhang, dass die Textsorte äußerst variabel und daher konzeptuell schwer fassbar ist (Koshar 2000: 15-16). Eine Einschätzung, die eineinhalb Jahrzehnte später noch immer zutreffend ist. Als erwähnenswerte Ausnahme sei die Arbeit von Sabine Müller (2012) genannt. Dass die digitalen Geisteswissenschaften bisher kaum zur Verbesserung der Lage beigetragen haben, hängt vermutlich damit zusammen, dass die (Retro-)Digitalisierung der komplex strukturierten historischen Bände sehr aufwändig ist. Derzeit stehen wenige exemplarische Analysen kleiner analoger Korpora mit Fokus auf entweder Textsortenmerkmalen (vgl. Gorsemann 1995; Pretzel 1995; Ramm 2000; Mittl 2007) oder der Entwicklung einzelner Regionen (Gorsemann 1995: Island; Pretzel 1995: Rheinland; Forschungsgruppe Tüschau 16 1998: Polen; Epelde 2004: Indien; Bock 2010: Rheinland), noch wenigeren computerlinguistischen Arbeiten gegenüber. Die letzteren basieren zwar auf umfangreicheren digitalen Datenmengen, nachdem jedoch ausschließlich rezentes Material herangezogen wird, bleiben historische Aspekte unberücksichtigt (vgl. Neumann 2003; Lam 2007; Gandin 2013, 2014).
4. Für die breite Untersuchung kultureller Narrative war es naheliegend, Reiseführer zu außereuropäischen Destinationen in das Korpus aufzunehmen. Es war Fritz Baedeker, dritter Sohn des Verlagsgründers Karl Baedeker und seit 1869 Leiter des Hauses, der das Verlagsprogramm um außereuropäische Titel erweiterte: Palästina und Syrien (1875), Unter- (1877) und Ober-Ägypten (1891), Nordamerika und Mexiko (1893), Konstantinopel und Kleinasien (1905), die afrikanische Mittelmeerküste (1909) und Indien und Ceylon (1914). Die strukturelle und deskriptive XML-Annotation nach TEI-Richtlinien (Version P5) konnte bereits für das gesamte Korpus abgeschlossen werden.
5. Aktuell befindet sich der *OpenSKOS* Editor in Überarbeitung. Die neue um *SKOS-XL* Komponenten erweiterte Version soll mit Jahresende zur Verfügung stehen.
6. Unterschieden werden: Kapelle, Kirche, Kloster, Mausoleum, Friedhof, Bildungs- und Wissenschaftseinrichtung, Gesundheits- und Sporteinrichtung, Museum, Sammlung, Palast, Theater, Industriebau, Inneneinrichtung, Verkehrsbau, Ensemble, Park.
7. Unterschieden werden: Denkmal, Skulptur, Gemälde, anderes Kunstwerk, Sammlung.
8. Die Applikation basiert auf der *corpus_shell* (Ďurčo et al.), einem modularen Framework für die Publikation von Sprachressourcen. Zum FCS / SRU-Protokoll siehe CLARIN ERIC sowie Stehouwer et al. 2012.

Bibliographie

Bock, Benedikt (2010): *Baedeker & Cook — Tourismus am Mittelrhein 1756 bis ca. 1914*. Bern / Berlin / Frankfurt am Main / New York / Paris / Wien: Peter Lang.

CLARIN ERIC (o. J.): *Federated Content Search (CLARIN-FCS)*. https://www.clarin.eu/content/federated-content-search-clarin-fcs [letzter Zugriff 10.09.2015].

Digital Humanities Austria: http://clarin.arz.oeaw.ac.at/dha/ .

Ďurčo, Matej / Mörth, Karlheinz / Schopper, Daniel / Siam, Omar (o. J.): *corpus-shell*. https://clarin.oeaw.ac.at/corpus_shell [letzter Zugriff 10. September 2015].

Epelde, Kathleen R. (2004): *Travel Guidebooks to India*. A Century and a Half of Orientalism. PhD, University Wollongong http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1195&context=theses [10. September 2015].

Forschungsgruppe Tüschau 16 (1998): *Die Darstellung anderer Kulturen*. Ermittlung von Stereotypen in deutschen Polen-Reiseführern (der Jahre 1990-1996). Oberhausen: Athena Verlag.

Gandin, Stefania (2013): "Translating the Language of Tourism. A Corpus Based Study on the Translational Tourism English Corpus (T-TourEC)", in: *Procedia — Social and Behavioral Sciences* 95: 325-335.

Gandin, Stefania (2014): "Investigating loan words and expressions in tourism discourse: a corpus driven analysis on the BBC-Travel Corpus", in: *European Scientific Journal*10, 2: 1-17.

Gorsemann, Sabine (1995): *Bildungsgut und touristische Gebrauchsanweisung*. Produktion, Aufbau und Funktion von Reiseführern. Münster / New York: Waxmann.

Jaworski, Adam / Pritchard, Anette (eds.) (2005): *Discourse, communication and tourism*. Clevedon: Channel View Press.

Koshar, Rudy (2000): *German Travel Cultures*. Oxford: Berg.

Lam, Peter Y. W. (2007): "A corpus-driven lexico-grammatical analysis of English tourism industry texts and the study of its pedagogic implications in English for Specific Purposes", in: Hildalgo, Encarnación / Quereda, Luis / Santana, Juan (eds.): *Corpora in the Foreign Language Classroom*. Amsterdam / New York: Rodopi 71-88.

Maingueneau, Dominique (2014): „Diskurs und Äußerungsszene. Zur gattungsspezifischen Kontextualisierung eines Zeitungsartikels zum unternehmerischen Bildungsdiskurs", in: Angermüller, Johannes / Nonhoff, Martin / Herschinger, Eva / Macgilchrist, Felicitas / Reisigl, Martin / Wedl, Juliette / Wrana, Daniel / Ziem, Alexander (eds.): *Diskursforschung*. Ein interdisziplinäres Handbuch. Bielefeld: transcript 433-453.

Mittl, Katja (2007): *Baedekers Reisehandbücher* Funktionen und Bewertungen eines Reisebegleiters des 19. Jahrhunderts (= Alles Buch. Studien der Erlanger Buchwissenschaft 22). Friedrich-Alexander-Universität Erlangen-Nürnberg.

Müller, Sabine (2012): *Die Welt des Baedeker*. Eine Medienkulturgeschichte des Reiseführers 1830-1945. Frankfurt / New York: Campus Verlag.

Neumann, Stella (2003): *Textsorten und Übersetzen*. Eine Korpusanalyse englischer und deutscher Reiseführer. Bern / Berlin / Frankfurt am Main / New York / Paris / Wien: Peter Lang.

Pretzel, Ulrike (1995): *Die Literaturform Reiseführer im 19. und 20. Jahrhundert*. Untersuchungen am Beispiel des Rheins. Bern / Berlin / Frankfurt am Main / New York / Paris / Wien: Peter Lang.

Ramm, Wiebke (2000): "Textual Variation in Travel Guides", in: Ventola, Eija (ed.): *Discourse and Community*. Doing Functional Linguistics. Tübingen: Gunter Narr 147-165.

Schmidt-Brücken, Daniel (2015): *Verallgemeinerung im Diskurs*. Generische Wissensindizierung in kolonialem Sprachgebrauch. Berlin / München / Boston: Walter de Gruyter.

Stehouwer, Herman / Durco, Matej / Auer, Eric Auer / Broeder, Daan (2012): "Federated Search: Towards a Common Search Infrastructure", in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC 2012): 3255–3259.

Wierlacher, Alois (1997): „Verfehlte Alterität. Zum Diskurs deutschsprachiger Reiseführer über fremde Speisen", in: Teuteberg, Hans Jürgen / Neumann, Gerhard / Wierlacher, Alois (eds.): *Essen und kulturelle Identität*. Europäische Perspektiven. Berlin: Akademie Verlag 498-509.

Baustein statt Datenruine: Beitrag zu einer Forschungsumgebung mit Bild-Text-Annotationen

Decker, Eric
decker@asia-europe.uni-heidelberg.de
Universität Heidelberg, Deutschland

Volkmann, Armin
armin.volkmann@asia-europe.uni-heidelberg.de
Universität Heidelberg, Deutschland

Guth, Matthias
guth@asia-europe.uni-heidelberg.de
Universität Heidelberg, Deutschland

In dieser Posterpräsentation soll das am Exzellenzcluster Asien und Europa der Universität Heidelberg angesiedelten Projekt "Standardisierte Arbeitsabläufe zur Retrodigitalisierung am Fallbeispiel der Grabungsdokumentation Kastell Heidelberg-Neuenheim" (RetroDig) vorgestellt werden. Im Rahmen dieses einjährigen (04/2015 - 03/2016) Disziplin übergreifenden Forschungsprojekts entwickelt die JRG Digital Humanites and Digital Cultural Heritage zusammen mit der Heidelberg Research Architecture (HRA) eine Reihe von strikt generischen Komponenten für den Einsatz in einem interdisziplinären Forschungsverbund.

Außeruniversitärer Partner bei diesem Vorhaben ist das Kurpfälzische Museum der Stadt Heidelberg, das eine bisher unpublizierte archäologische Grabungsdokumentation aus den 1970er für einen exemplarischen Retrodigitalisierungs-Workflow bereitstellt.

Im Fallbeispiel werden sämtliche analogen Artefakte der vom Verfall bedrohten Dokumentation (handgezeichnete Pläne, handschriftliches Grabungstagebuch und Papierabzüge von Fotos) von Mitarbeitern der JRG Digital Humanities digitalisiert, erschlossen und annotiert. Dazu wird am Exzellenzcluster vorhandene digitale Infrastruktur verwendet und wo nötig erweitert.

Das Kernstück der digitalen Forschungsumgebung des Clusters ist das Metadaten-Ökosystem Tamboti[1] . Es handelt sich dabei um ein auf eXist DB basierendes Open Source System, dass von der HRA zusammen mit ihren Partnern bereits seit mehreren Jahren entwickelt und schrittweise ausgebaut wird. Größere Komponenten für Tamboti werden grundsätzlich an Forschungsfragen orientiert im Rahmen von kleinen thematischen Fallbeispielen entworfen und implementiert. Nach Projektabschluss werden die entwickelten Softwarekomponenten von der HRA in den Regelbetrieb in Forschung und Lehre überführt.

Im hier vorgestellten Projekt RetroDig sollen darüber hinaus auch Einsatzmöglichkeiten im Museumsbereich evaluiert werden. Die generierten Datensätze und Softwarekomponenten werden die Grundlage für eine inhaltliche Aufbereitung der RetroDig Ergebnisse in einem zukünftigen digitalen Editionsprojekt der JRG Digital Humanities and Digital Cultural Heritage sein.

Im derzeit laufenden Prokjekt wurden bereits die analogen Dokumente im Medialab der HRA digitalisiert. Im ersten Erschließungsschritt werden die materiellen Gesichtspunkte der Artefakte betrachtet und dabei die Objektmetadaten im VRA Core 4 Standard aufgenommen. Dazu wird der formbasierter Zizphus VRA Editor[2] verwendet, der die VRA-XML Datensätze direkt in einer Kollektion Tamboti speichert, wo die Daten bereits durchsucht, mit anderen Nutzern geteilt oder für Präsentationszwecke im integrierten Atomic-Wiki aufbereitet werden können.

Mitunter aus arbeitsökonomischen Gesichtspunkten werden beim Erstellen der Metadaten auch Beschriftungen transkribiert und vorerst im VRA <inscriptionSet> aufgenommen. Erst einmal ohne dabei qualitative Aussagen über deren Inhalt zu treffen.

Dieses Vorgehen erlaubt es der HRA parallel zur Erfassung der Daten an der Entwicklung einer mehrteiligen Annotationskomponente zu arbeiten. Diese besteht 1.) aus einem DOM-nahen SVG Editor, der auf OpenSeadragon aufsetzt, 2.) einem semantischen Verlinkungsmechanismus (zum Zeitpunkt der Einreichung des Posters werden mehrere Möglichkeiten der Informationsmodellierung und Umsetzung diskutiert und ausprobiert. Bei der finalen Posterpräsentation wird sowohl auf die Diskussion, als auch auf die Implementierung Bezug genommen werden) und 3.) der Integration von TEIAN[3] , einem Editor, der es erlaubt beliebige Subsets von XML Vokabularen über eine graphische Nutzeroberfläche auf einen Text anzuwenden. Mit diesen Komponenten wird es u. a. möglich sein: Annotationen im Sinne des Open Annotation Data Model (OADM) zu erstellen, das im ersten Erfassungsschritt bereits transkribierte Material z. B. in TEI auszuzeichnen und mit dem SVG-Editor Grabungspläne so nachzuzeichnen, dass sie im Detail annotiert und in ein Geoinformationssystem eingehängt werden können.

Im vorgestellten Projekt haben wir uns aufgrund der kurzen Projektlaufzeit bewusst gegen die Entwicklung einer projektspezifischen Präsentationsoberfläche entschieden. Stattdessen konzentrieren wir uns auf Arbeitsabläufe zum Erstellen standardisierter Daten und der Frage wie diese möglichst zukunftssicher modelliert werden können. Die Datenanzeige soll derweil über Tamboti oder per IIIF-P über Mirrador oder andere standardkonforme Viewer erfolgen können. Die Daten können über Tamboti für weitere Nutzergruppen freigegeben werden und so z. B. im Unterricht mit Atomic Wiki oder HyperImage[4] aufbereitet werden. Ein Tutorenprogramm[5] und eine ausführliche Dokumentation dafür wurden in den letzten Semestern bereits etabliert und regelmäßig in unterschiedlichen Fachbereichen eingesetzt.

Notes

1. Die Produktivinstanz des Exzellenzclusters Asien und Europa ist unter http://tamboti.uni-hd.de zu erreichen. Der Quellcode ist unter https://github.com/exc-asia-and-europe/tamboti veröffentlicht.
2. Ziziphus ist ein integrierter Bestandteil von Tamboti und wird von der HRA in Zusammenarbeit mit betterFORM und eXist solutions entwickelt. Der Quellcode ist unter https://github.com/exc-asia-and-europe/ziziphus/ veröffentlicht.

3. Der Quellcode ist unter https://sourceforge.net/projects/teian/ veröffentlicht.
4. Derzeit wird die Community Edition des HyperImage Authoring Environment in der Version 3.0.beta2 eingesetzt.
5. Das Tutorenprogramm wurde im Wintersemester 2013 / 14, unterstützt durch Mittel aus dem „Willkommen in der Wissenschaft“ Förderprogramm, vom Lehrstuhl für Visuelle und Medienanthropologie (Christiane Brosius) initiiert und gemeinsam mit der HRA implementiert. Der kontinuierliche Ausbau des Programms wird von der Abteilung Schlüsselkompetenzen und Hochschuldidaktik der Universität Heidelberg begleitet.

Ontologisierung vom Thompson Motif‘s Index Teilergebnisse eines Softwareprojektes zum Thema „Classification of Folktales“, bei Antónia Kostevá, Universität des Saarlandes

Declerck, Thierry
declerck@dfki.de
DFKI GmbH, Deutschland

Wir präsentieren eine Ontologisierung des Thompson‘s Motif Index Katalogs. Motive können als sich regelmäßig wiederholende Elemente einer Literaturgattung betrachtet werden, auch in verschiedenen Erzählungen. Es können darunter prototypische Charaktere, Objekte, Handlungen oder auch Ereignisse verstanden werden. Die Arbeit von Stith Thompson (Thompson 1955-58) ist ein Versuch, solche Motive in Märchen und ähnlichen Texttypen zu indizieren. Und dieser Index wird heute sehr oft für die Analyse von Märchen verwendet.

Basierend auf Arbeiten von Declerck et al. (2012) und Eisenreich et al. (2014), die die Bedeutung von Ontologien für die automatisierte Analyse von Märchen betonen, haben wir die Liste von Motiven, die Thompson bereitgestellt hat, in eine OWL [1] Ontologie überführt, die in Januar 2016 gemacht worden ist, nachdem das zugrundenliegende Softwareprojekt an der Universität des Saarlandes abgeschlossen war.

Wir verwenden für unsere Arbeit eine digitale Version des Indexes [2] . Vereinfachte Beispiele von Motiven sind unten angegeben:

- A. Mythological motifs
  - A0-A99. Creator
    - A21 Creator from above
      - A21.1. Male and female creators
  - A100-A499 Gods
  - A500-A599 Demigods and Culture Heroes
  - A500-A599 Cosmogony and cosmology
  - A900-A999 Topological
  - A1000-A1099 World calamitities
  - A1100-A1199 Establishment of natural order
  - A1100-A1699 Creation and ordering of human life
    - A1411 Theft of light
    - A1415 Theft of fire
  - A1700-A1799 Creation of animal life
  - A2200-A2599 Animal characteristics
  - A2600-A2699 Origins of trees and plants
- B. Animals
  - B0-B99 Mythical animals
  - B100-B199 Magic animals

….

Diese hierarchische Listenstruktur wurde in eine formale Sub-Kategorisierung unter Verwendung der OWL Repräsentationssprache überführt. Wir nennen die Hauptkasse dieser Hierarchie „TMI“.[3] Und alle Motive, die wir als „Blätter“ in dieser Hierarchie auftreten lassen, wurden einfach als Instanzen von einer flachen Struktur von Motivklassen eingeführt. Hier ist „Motif“ die Hauptkasse für alle Instanzen, wie es in Abbildung 1 und in den Kodebeispielen unten zu sehen ist.

Es sind somit 34.913 Motive kodiert worden, nach einer automatischen Konvertierung aus der oben genannten Internetquelle.

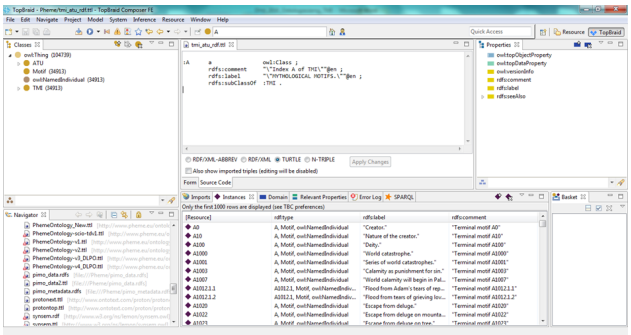


Abb. 1: Toplevel Stufe unserer Ontologie, dargestellt im TopBraid Editor.

Die Kodebeispiele unten zeigen erst eine Instanz eines Motivs, und dann wie die beinhaltenden Klasse in der Klassenhierarchie eingebettet ist.

```
:A104.1  a           :A104 , :Motif , owl:NamedIndividual ;
       rdfs:comment  "\"Terminal motif A104.1\""@en ;
       rdfs:label    "\"Living person becomes god.\""@en .
```

Und die Klasse, von der A104.1 eine Instanz ist, wird als Subklasse von A kodiert:

```
:A104   a               owl:Class ;
        rdfs:comment    "\"Index A104 of TMI\""@en ;
        rdfs:label      "\"The making of gods.\""@en ;
        rdfs:subClassOf :A .

:A      a               owl:Class ;
        rdfs:comment    "\"Index A of TMI\""@en ;
        rdfs:label      "\"MYTHOLOGICAL MOTIFS.\""@en ;
        rdfs:subClassOf :TMI .
```

Aktuelle Arbeit besteht darin, links zu einem anderen Referenzwerk in der Folkloristik zu etablieren: Die Aarne-Thompson-Uther (ATU, s. Uther 2004) Taxonomie von Märchentypen. Eine erste Ontologie Version von ATU ist von uns bereitgestellt worden, aber muss noch überprüft werden. Links von ATU zu TMI werden dann mit Hilfe von symmetrischen „Object Properties“ beschrieben. Wir arbeiten auch daran, die „labels“ der TMI Klassen und Instanzen in anderen Sprachen (Deutsch und Französisch) zu erweitern.

Notes

1. OWL steht für „ Ontology Web Language “. Siehe auch http://www.w3.org/TR/owl-features/
2. Siehe http://www.ruthenia.ru/folklore/thompson/index.htm
3. Siehe Abbildung 1 weiter unten, für einen Screenshot des von uns verwendeten Ontologie Editors “ TopBraid ” (cf. TopQuadrant).

Bibliographie

Eisenreich, Christian / Ott, Jana / Süßdorf, Tonio / Willms / Declerck, Thierry (2014): “From Tale to Speech: Ontology-based Emotion and Dialogue Annotation of Fairy Tales with a TTS Output”, in: *Proceedings of ISWC 2014, Riva del Garda, Italy, Springer.*

Declerck, Thierry / Lendvai, Piroska / Darányi, Sándor (2012): “Multilingual and Semantic Extension of Folk Tale Catalogues”, in: Jan Christoph Meister (ed.): *Digital Humanities 2012*. Conference Abstracts, Hamburg, Germany. Hamburg: Hamburg University Press.

TopQuadrant: *TopBraid Live* http://www.topquadrant.com/products/topbraid-live/ [letzter Zugriff 11. Februar 2016].

Stith Thompson (1955-1958): *Motif-index of folk-literature*. A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends. Revised and enlarged edition. Bloomington: Indiana University Press.

Uther, Hans-Jörg (2004): *The Types of International Folktales*. A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson (= FF Communications 284–286). Helsinki: Suomalainen Tiedeakatemia.

W3C = World Wide Web Consortium (2004): *OWL Web Ontology Language* http://www.w3.org/TR/owl-features/ [letzter Zugriff 16. Februar 2016].

Metaphern digital Auf dem Weg von der Annotation zur automatischen Detektion

Do Dinh, Erik-Lân
dodinh@kdsl.informatik.tu-darmstadt.de
UKP Lab, TU Darmstadt

Gerloff, Malte
gerloff@kdsl.informatik.tu-darmstadt.de
Institut für Philosophie, TU Darmstadt

Núñez, Alexandra
nunez@kdsl.informatik.tu-darmstadt.de
Institut für Sprach- und Literaturwissenschaft, TU Darmstadt

Das interdisziplinäre Forschungsteam *Natur & Staat* hat sich zum Ziel gesetzt, ein innovatives Computerprogramm für die (semi-)automatische Metapherndetektion zu entwickeln. Der didaktische Nutzen und wissenschaftliche Mehrwert des Tools für die geisteswissenschaftliche Forschung und Lehre lassen sich folgendermaßen umreißen: Synchrone und diachrone Textanalysen des kontextuellen Metapherngebrauchs können z. B. bei vorliegenden großen Textkorpora rascher durchgeführt und erste wissenschaftliche Hypothesen evaluiert und modifiziert werden. Es können zudem auch textsortenübergreifende Analysen qualitativer bis quantitativer Art durchgeführt werden. Des Weiteren soll das Tool sowohl auf verschiedene Metapherntheorien als auch auf die jeweiligen Forschungsabsichten anpassbar sein. Der Fokus des interdisziplinären Vortrags liegt auf der geisteswissenschaftlichen Methodik, der didaktischen Vermittlung der statistischen Modellierung des Tools und insbesondere auf den bis jetzt erzielten Zwischenergebnissen unserer Arbeit.

Der Untersuchungsgegenstand

Den namensgebenden Untersuchungsgegenstand bildet das neunbändige Volltextkorpus, das unter dem programmatischen Titel „Natur und Staat. Beiträge zur naturwissenschaftlichen Gesellschaftslehre“ (im Folgenden: *Natur & Staat*) u. a. von dem Zoologen Ernst Haeckel herausgegeben und im Zeitraum von 1903-1911 publiziert wurde. Die Bände liegen als Volltextdigitalisate in Frakturschrift vor.

Vorausgegangen war ein national und international stark rezipiertes Preisausschreiben, das 1900-1901 ausgeschrieben und vom bekannten Großindustriellen Friedrich Alfred Krupp (1854-1902) anonym finanziert wurde. Die Preisfrage lautete: „Was lernen wir aus den Prinzipien der Descendenztheorie in Beziehung auf die innerpolitische Entwicklung und Gesetzgebung der Staaten?“. Diskursanalytisch betrachtet, ist das Korpus u. a. eine Antwort auf die in verschiedenen wissenschaftlichen Disziplinen vorangegangene Rezeption der Theorien Charles Darwins im 19. Jahrhundert. Zusammen mit der sozialdarwinistischen Bewegung wurde ein neues Deutungsmodell des Menschen und seiner Stellung in der Welt vorbereitet und schließlich etabliert.

Natur & Staat bildet besonders aufgrund der Textsorte (vgl. Fix 2011) einen geeigneten Ausgangspunkt für die Metapherndetektion: Es handelt es sich um (populär-)wissenschaftliche Texte (vgl. Polenz 1999, 1981), die primär das Ziel verfolgen, Sachverhalte, Ideen, Theoreme auf der Basis einer sozialdarwinistischen Agenda zu reflektieren, argumentativ aufzubereiten und einem breiten Adressatenkreis zu vermitteln. Das Preisausschreiben öffnete den Weg für einen tiefgreifenden Wandel „ethisch“ genannter, auf eine Einflussnahme kollektiver „Entwicklungen“ abzielender Handlungsmaximen. Die Abhandlungen entwarfen Szenarien einer sozialdarwinistischen Governance von Bildung: einer Sitten- und Wertepolitik für das „Leben“ – mitsamt biotechnischen und eugenischen Implikationen (vgl. Gehring 2009).

Interaktionale Metaphernmodelle

Innerhalb des aufgezeigten Kontextes sollen Ausdrucksgestalt und semantische Funktion von sprachlichen Metaphern analysiert werden. Interaktionale Modelle mit ihrer binären Übertragungsstruktur fokussieren u. a. konventionalisierte, ubiquitäre Alltagsmetaphern. Eine wissenschaftliche Herausforderung in diesem Forschungsparadigma bilden jedoch die kühnen Metaphern.

Kühne Metapherntheorien

Metapherntheorien, die die kühne Metapher ins Auge fassen, verorten sich innerhalb des hermeneutischen Paradigmas und legen das Primat auf das Besondere. Nur Metaphern, die besonders seien, seien laut Max Black (1954) für die Philosophie relevant, da nur sie eine erkenntnistheoretische Funktion hätten. Des Weiteren stellt er in der Folge von I. A. Richards (1936) fest, dass die Metapher zwei Bestandteile aufweise: Fokus und Rahmen. Beide bestünden mindestens aus einem Wort und sie seien interaktional, derart, dass ein System von Implikationen, welches dem Rahmen unterliege, auf das Implikationssystem des Fokus‘ rückbezüglich wirke. Die Übertragung der Implikationen geschehe auf der Basis von Ähnlichkeiten. Sie speise sich somit aus der immanenten Semantik des Implikationssystems, welches durch statistische Verfahren z. B. über selektionale Präferenzen für den Computer abbildbar gemacht werden kann.

Gehring (2011) erweiterte dann Blacks Interaktionstheorie, insofern sie den Kontextbruch als eine notwendige Bedingung der Metapher einführt, da aus diesem und dem Fehlen des wörtlichen Sinns der Interaktion die Metapher entstehe. Da aber nicht jede semantische Übertragung eine Metapher sei, weil man ansonsten in die Beliebigkeit abdriften würde, ist die Interaktion der beiden Entitäten, auch eine Interaktion der besonderen Art. Die Größe des Rahmens respektive des Kontexts der Metapher ist sowohl bei Gehring als auch bei Black variabel, könne allerdings durch eine Weglassprobe evaluiert werden. Einigkeit besteht überdies auch darüber, dass die Metapher nur in der Gesamtheit von Fokus und Rahmen bestehe. Weder hinreichende noch notwendige Bedingungen der Metapher seien hingegen, laut Gehring (2011), sowohl Bildlichkeit als auch lexikalische sowie grammatikalische Indikatoren, weil die Metapher deutungsoffen sei; auch schließt sie aus, dass es ein Kontinuum zwischen Begriff und Metapher gebe.

Ubiquitäre Metapherntheorie

Metapher bezeichnet im Rahmen der konzeptuellen Metapherntheorie (Lakoff / Johnson 1980; Lakoff 1993; Goatly 2007; Kövecses 2015) zunächst ein zentrales kognitives Vermögen. Die Kernthese der konzeptuellen Metapherntheorie bildet die Annahme, dass eine unbekannte Erfahrung (*target domain*) in Analogie zu einer bereits bekannten Erfahrung (*source domain*) sprachlich konzeptualisiert wird. Dies betrifft beispielsweise vage Konzepte wie Emotionen (Kövecses 2003), Theorien und andere abstrakte Sachverhalte, Relationen (Johnson 1987) und Prozesse (Núñez 2014). Diese metaphorischen Übertragungsmuster lassen sich auch im Textkorpus *Natur & Staat* indexikalisch auf der Sprachoberfläche mit dem Fokus auf Lexemen, z. B. Genitivkonstruktionen ([[NP] der/des [NP]]), und weiteren usuellen Konstruktionen korpuslinguistisch eruieren und annotieren. Die Verteilung der metaphorischen Sprachphänomene ist dabei im Vergleich zu den bereits erwähnten kühnen Metaphern

als ubiquitär (Paul 1909, Bühler 1934, Paprotte / Dirven 1985) einzustufen. Die entlehnten konzeptuellen und sprachlich umgesetzten Domänen können schließlich in einem zweiten Schritt hinsichtlich regelmäßiger Konzeptübertragungen zwischen den beiden Domänen systematisiert werden.

Forschungslage

Seit einigen Jahren lässt sich eine verstärkte Entwicklung (zumeist überwachter) automatisierter Verfahren für die Identifikation und Interpretation von Metaphern beobachten. Diese Verfahren nutzen überwiegend als Grundlage für die Modellierung des Untersuchungsgegenstandes *Metapher* die bereits vorgestellte konzeptuelle Metapherntheorie von Lakoff und Johnson (1980). Dabei gibt es Unterschiede in der Zielsetzung einzelner Verfahren: Während einige Ansätze lediglich die metaphorische Verwendung bestimmter Konstruktionen (z. B. Subjekt-Verb-Objekt oder Adjektiv-Nomen) bewerten (Turney et al. 2011; Tsvetkov et al. 2013; Shutova 2013), weiten andere Verfahren eine solche Klassifizierung auf alle Inhaltswörter aus (Beigmann Klebanov et al. 2014; Dunn 2013). Darüber hinaus existieren Unterschiede in der Tiefe der Analyse. So wird bei einem Großteil der Verfahren die konzeptuelle Ebene ausgeblendet und Metaphern werden ausschließlich als Realisierungen auf der Sprachoberfläche identifiziert, wenn auch mit Methoden, die sich auf das Vorhandensein einer konzeptuellen Ebene stützen oder diese voraussetzen. Wenige Verfahren versuchen, metaphorische Abbildungen auf der konzeptuellen Ebene zu erkennen (Mason 2004; Shutova et al. 2013).

Neben der Wahl der untersuchten Metapherntheorie und der Ebene, auf der Metaphern erkannt werden, ist ein weiteres Merkmal bestehender automatischer Verfahren die Sprache der behandelten Texte. Diese ist – nicht zuletzt aus praktischen Gründen, wie dem Vorhandensein annotierter Korpora und weiterer Ressourcen wie vordefinierten Datenbanken zu Abstraktheits- und Konkretheitsbewertungen von Wörtern – üblicherweise Englisch. Für einige Verfahren existieren überwachte maschinelle Lernverfahren, die es ermöglichen, auf annotierten, englischen Daten ein Modell zu trainieren, welches durch zweisprachige Wörterbücher für die automatische Identifikation auf Texten anderer Sprachen anwendbar ist (Tsvetkov et al. 2013).

Annotationstool, Vorgehen und Ziele

Für die Annotation sowohl kühner als auch ubiquitärer Metaphern verwenden wir zunächst WebAnno (Yimam et al. 2013), ein Web-Annotationsprogramm für mehrere Benutzer mit frei definierbaren Annotationsarten. Dazu setzen wir jeweils unterschiedliche Annotationsebenen ein. Zwei Benutzer mit geisteswissenschaftlicher Expertise im Bereich der Metapherntheorien annotieren dieselben vereinbarten Textabschnitte, um Vergleichswerte für ein Inter-Annotator Agreement zu erhalten. Dabei nutzen wir die Exportmöglichkeiten von WebAnno, um die Annotationsdaten der Nutzer sowie weitere linguistische Merkmale wie Wortarten zu exportieren, mit denen dann ein Inter-Annotator Agreement berechnet, sowie die automatische Weiterverarbeitung für die Identifizierung von Metaphern ausgestaltet werden können.

Hierfür testen wir zunächst in einer Pilotstudie, wie bestehende state-of-the-art Verfahren für ubiquitäre Metaphern (Tsvetkov et al. 2013; Beigmann Klebanov et al. 2014) auf deutschen, insbesondere historischen, Texten abschneiden. Dafür benötigte Ressourcen werden erstellt oder erweitert. Außerdem wird die Weiterentwicklung bestehender Methoden sowie das Entwickeln neuer Methoden vorangetrieben, um auch kühne Metaphern zu identifizieren. Die Identifikation solcher Metaphern stand bislang nicht im Vordergrund automatischer Systeme. Durch die Wahl des Korpus‘ ergeben sich weitere Herausforderungen, zum Beispiel eine ungenügende Abdeckung von *Natur & Staat* durch bestehende manuell erstellte Ressourcen wie etwa GermaNet (Hamp / Feldweg, 1997; Henrich / Hinrichs 2010), sowie ein Mangel an Vergleichskorpora für statistische Verfahren, die beispielsweise mittels Kookkurrenzen oder selektionaler Präferenzen Unterschiede zur „Standardsprache“ feststellen können.

Ausblick/Fazit

Während mit dem ubiquitären Metaphernmodell somit eher implizite Mechanismen der sprachlichen Sachverhaltsperspektivierung, kurz: konventionalisierte Metaphern, in den Fokus rücken, wird das kühne Metaphernmodell besonders den strukturellen Rezeptions- und Gestaltungsprinzipien des Textes gerecht und vermag auf der Basis des vorgestellten Fokus- und Rahmenkonzepts die besonders markanten Sinnbezirke im Text in den Vordergrund zu rücken. Kühne Metaphern in theoretischen, argumentativen Texten heben sich deutlich von der diskursiv omnipräsenten Sprachstruktur ab: Sie weisen zugleich dadurch, dass sie einen Bruch mit der stilistisch homogenen Sprachstruktur erzeugen, erst auf eben diesen textuell gegebenen Sprachstandard hin und weisen zugleich aufgrund ihres semantischen Verdichtungspotenzials über diesen hinaus (Gehring 2011). Als Sinnbezirke auf der Sprachoberfläche rhetorisch inszeniert und umgesetzt, besitzen kühne Metaphern in *Natur & Staat* insbesondere epistemologisches Potenzial, indem sie neue Theoreme und Idee erst eine besondere sprachliche Gestalt zu geben vermögen.

Hinsichtlich der Annotation sowohl kühner als auch ubiquitärer Metaphern ist ein hohes Inter-Annotator Agreement notwendig; einerseits als Bestätigung für eine hinreichend gute Operationalisierung und Modellierung der verwendeten Theorie, andererseits um einen verlässlichen Goldstandard für das Training und die Evaluation automatischer Verfahren bereit zu stellen. Diesen Zwischenschritt und die daraus resultierenden Möglichkeiten für Geisteswissenschaften als auch Informatik werden wir in unserem Vortrag vorstellen.

Bibliography

Black, Max (1954): „Metaphor“, in: *Proceedings of the Aristotelian Society. New Series* 55: 273-294.

Black, Max (1977): „More about Metaphor“, in: *Dialectica* 31: 431-457.

Beigman Klebanov, Beata / Leong, Ben / Heilman, Michael / Flor, Michael (2014): „Different Texts, Same Metaphors: Unigrams and Beyond“, in: *Proceedings of the Second Workshop on Metaphor in NLP*, Baltimore, MD, USA 11-17.

Bühler, Karl (1934 / 1982): *Sprachtheorie*. Die Darstellungsfunktion der Sprache. Stuttgart / New York: Gustav Fischer.

Dunn, Jonathan (2013): „Evaluating the Premises and Results of Four Metaphor Identification Systems“, in: *Proceedings of CICLing 2013,* Samos, Griechenland: 471-486.

Fix, Ulla (2011): *Texte und Textsorten - sprachliche, kommunikative und kulturelle Phänomene*. Berlin: Frank & Timme.

Gehring, Petra (2006): „Vom Begriff zur Metapher. Elemente einer Methode der historischen Metaphernforschung“, in: Günter Abel (ed.): *Kreativität*. Kolloquiumsbeiträge des XX. Kongresses der Allgemeinen Gesellschaft für Philosophie in Deutschland. Hamburg: Meiner 800-815.

Gehring, Petra (2009): „Biologische Politik um 1900 – Reform, Theorie, Experiment?“, in: Griesecke, Birgit / Krause, Marcus / Pethes, Nicolas / Sabisch, Katja (eds.): *Kulturgeschichte des Menschenversuchs im 20. Jahrhundert*. Frankfurt am Main: Suhrkamp 48-76.

Gehring, Petra (2010): „Erkenntnis durch Metaphern? Methodologische Bemerkungen zur Metapherntheorie“, in: Junge, Matthias (ed.): *Metaphern in Wissenskulturen*. Wiesbaden: VS Verlag für Sozialwissenschaften 203-220.

Gehring, Petra (2011): „Metaphertheoretischer Visualismus – Ist die Metapher »Bild«?“, in: Kroß, Matthias / Zill, Rüdiger (eds.): *Metapherngeschichten – Perspektiven einer Theorie der Unbegrifflichkeit*. Berlin: Parerga Verlag 15-31.

Gehring, Petra / Gurevych, Iryna (2014): „Suchen als Methode? Zu einigen Problemen digitaler Metapherndetektion“, in: *Journal für Phänomenologie Schwerpunkt: Metaphern als strenge Wissenschaft* 41: 99-109.

Goatly, Andrew (2007): *Washing the Brain – Metaphor and Hidden Ideology*. Amsterdam/ Philadelphia: John Benjamin Publishing Company.

Hamp, Birgit / Feldweg, Helmut (1997): „GermaNet - a Lexical-Semantic Net for German“, in: *Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications,* Madrid, Spanien 9-15.

Henrich, Verena / Hinrichs, Erhard (2010): „GernEdiT - The GermaNet Editing Tool“, in: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta 2228-2235.

Johnson, Mark (1987): *The Body in the Mind*. The Bodily Basis of Meaning, Imagination, and Reason. Chicago: Chicago University Press.

Kövecses, Zóltan (2003): *Metaphor and Emotion*. Language, Culture, and the Body in Human Feeling. Cambridge: Cambridge University Press.

Kövecses, Zóltan (2015): *Where Metaphor Come From*. Reconsidering Context in Metaphor. Oxford: University Press.

Kohl, Katrin (2007): *Metapher*. Stuttgart: Sammlung Metzler.

Lakoff, George (2006): „Conceptual Metaphor. The Contemporary Theory of Metaphor [1993]“, in: Geeraerts, Dirk (ed.): *Cognitive Linguistics: Basic Readings*. Berlin: Mouton de Gruyter 185-238.

Lakoff, George / Johnson, Mark (1980): *Metaphors We Live by*. Chicago: Chicago University Press.

Mason, Zachary (2004): „Cormet: A Computational, Corpus-based Conventional Metaphor Extraction System“, in: *Computational Linguistics*, 30, 1: 23-44.

Núñez, Alexandra (2014): „Wenn das 'Embodiment' politisch wird: Das Image-Schema PATH und seine Realisierung im Mediendiskurs zum 'Arabischen Frühling '“, in: Polzenhagen, Frank / Kleinke, Sonja / Kövecses, Zoltán / Vogelbacher, Stefanie (eds.): *Cognitive Explorations into Metaphor and Metonymy*. Bern / Berlin / Frankfurt am Main / New York / Paris / Wien: Peter Lang 149-164.

Paprotté, Wolf / Dirven, René (ed.) (1985): *The Ubiquity of Metaphor*. Metaphor in Language and Thought. Amsterdam / Philadelphia: John Benjamin Publishing Company.

Paul, Hermann (1909): *Prinzipien der Sprachgeschichte*. Halle a. S.: Niemeyer.

Polenz, Peter von (1981): „Über die Jargonisierung von Wissenschaftssprache und wider die Deagentivierung“, in: Bungarten, Theo (ed.): *Wissenschaftssprache- Beiträge zur Methodologie, theoretische Fundierung und Deskription*. München: Wilhelm Fink Verlag 85-110.

Polenz, Peter von (1999): *Deutsche Sprachgeschichte*. Vom Spätmittelalter bis zur Gegenwart.

Bd. III. 19. und 20. Jahrhundert. Berlin / New York: Walter de Gruyter.

Ricoeur, Paul (1972): „La métaphore et le problème central de l'herméneutique", in: *Revue philosophique de Louvain* 70: 93-112.

Shutova, Ekaterina (2013): „Metaphor Identification as Interpretation", in: *Proceedings of *SEM 2013*, Atlanta, GA, USA 276-285.

Shutova, Ekaterina / Sun, Lin (2013): „Unsupervised Metaphor Identification using Hierarchical Graph Factorization Clustering", in: *Proceedings of NAACL*, Atlanta, GA, USA 978-988.

Tsvetkov, Yulia / Mukomel, Elena / Gershman, Anatole (2013): „Cross-lingual Metaphor Detection using Common Semantic Features", in: *Proceedings of the First Workshop on Metaphor in NLP*, Atlanta, GA, USA 45-51.

Turney, Peter D. / Neuman, Yair / Assaf, Dan / Cohen, Yohai (2011): „Literal and Metaphorical Sense Identification through Concrete and Abstract Context", in: *Proceedings of EMNLP 2011*, Stroudsburg, PA, USA 680-690.

Weinrich, Harald (1976): *Sprache in Texten*. Stuttgart: Ernst Klett Verlag.

Yimam, Seid Muhie / Gurevych, Iryna / Eckart de Castilho, Richard / Biemann, Chris (2013): „WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations", in: *Proceedings of ACL 2013, Demo Session*, Sofia, Bulgaria 1-6.

CLARIN-D: Ressourcen gesprochener Sprache und Webservices des Bayerischen Archivs für Sprachsignale

Draxler, Christoph
draxler@phonetik.uni-muenchen.de
Bayerisches Archiv für Sprachsignale, Institut für Phonetik und Sprachverarbeitung, Deutschland

Schiel, Florian
schiel@phonetik.uni-muenchen.de
Bayerisches Archiv für Sprachsignale, Institut für Phonetik und Sprachverarbeitung, Deutschland

Reichel, Uwe
reichelu@phonetik.uni-muenchen.de
Bayerisches Archiv für Sprachsignale, Institut für Phonetik und Sprachverarbeitung, Deutschland

Kisler, Thomas
kisler@phonetik.uni-muenchen.de
Bayerisches Archiv für Sprachsignale, Institut für Phonetik und Sprachverarbeitung, Deutschland

Das BAS Repository

Das Repository des BAS ist in einen öffentlich zugänglichen und einen zugangsgeschützten Bereich unterteilt. Der öffentliche Bereich enthält die Metadaten der im Repository enthalten Datenbanken gesprochener Sprache, der geschützte Bereich die Annotationen und Sprachsignaldaten.

Aktuell (Stand 31.12.2015) umfasst das Repository 33 Korpora mit den Schwerpunkten Sprachtechnologie, regionale Variation, Sprechermerkmale und Grundlagenforschung. Auf der obersten Repository-Ebene werden die wichtigsten Angaben prominent platziert: Name und Eigentümer der Ressource, Audio bzw. Video, verwendete Sprache, und die Zugangsrestriktionen (s. Kap. 6).

Von den 33 Korpora enthält eines arabische Sprachaufnahmen, und enthalten zwei italienische, drei englische und 26 deutsche Sprachaufnahmen; dazu kommt ein Korpus in Deutscher Gebärdensprache. 26 Korpora enthalten nur gesprochene Sprache, 6 sowohl gesprochene Sprache als auch Video, eines nur Video. Ein Korpus ist allgemein verfügbar, 30 sind für akademische Nutzer_innen frei zugänglich, zwei nur als lizensierte Korpora.

Der Datenumfang beträgt mehr als 2,86 TB an Audio-, Video- und Sensorsignaldaten, sowie ca. 17,4 GB an Annotations- und Metadaten.

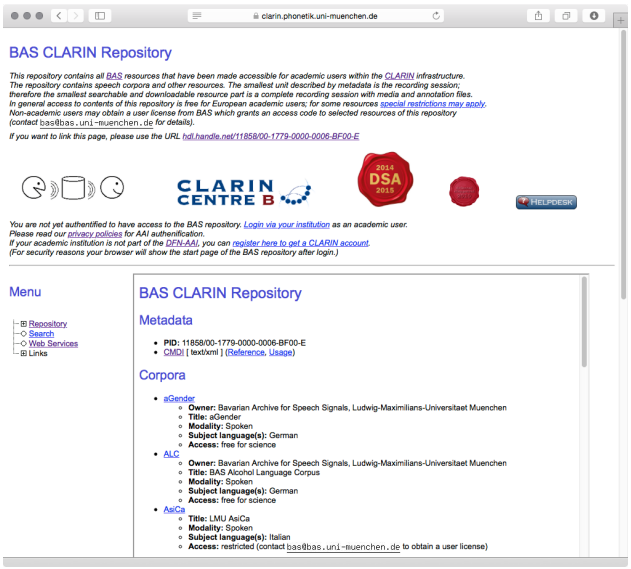


Abb. 1: Startseite des BAS Repository mit der Liste verfügbarer Korpora.

Inhalt des Repository

Die überwiegende Anzahl der Ressourcen wurde vom BAS selbst oder in Kooperation mit industriellen oder akademischen Partnern erstellt. Zunehmend werden weitere, von externen Partnern erstellte Ressourcen in das Repository aufgenommen.

Abb. 2: Repositoryansicht der Sessions im Korpus ALC.

Darüberhinaus sucht das BAS aktiv nach verwaisten Ressourcen, d. h. Sprachdatensammlungen, die im Rahmen von Projekten, Dissertationen, oder sonstigen Datenerhebungen erstellt wurden, deren Fortbestand aber aufgrund fehlender Finanzierung oder personeller Veränderungen gefährdet ist. Für diese Ressourcen bietet das BAS Unterstützung bei der Erstellung der Metadaten sowie auch die Archivierung im Repository an. Ein Leitfaden zur Aufbereitung externer Korpora ist in Vorbereitung.

Aktuell werden Metadaten für das schweizerische Jugendsprachekorpus (Tissot 2015), das WaSeP Korpus zur Analyse emotionaler Prosodie (Wendt 2007), das Sprechermerkmale-Korpus von Brüderpaaren von Hanna Feiser (Feiser 2015), das ASD-Corpus (Siebenbürger Deutsch) der Italianistik der LMU München sowie das VOYS Korpus mit Aufnahmen schottischer Jugendlicher (Dickie et al. 2009) für das Repository aufbereitet.

CMDI Metadatenformat

Grundlage des Repository sind zum einen das Metadatenformat CMDI (*Component Metadata Initiative*) (CLARIN-D AP 5 2012) , zum anderen ein in *perl* implementiertes und an die CLARIN-Anforderungen angepasstes Repositoryframework.

CMDI zeichnet sich dadurch aus, dass es in kontrollierter Form erweiterbar ist und selbstbeschreibend ist. Sämtliche in CMDI verwendeten Deskriptoren, Components genannt, müssen in einem öffentlich zugänglichen Register spezifiziert sein. Profile sind vordefinierte Listen von Deskriptoren für spezifische Datenbestände.

Das BAS hat zwei Profile für Ressourcen gesprochener Sprache definiert, das MediaCorpus und das MediaSession Profil. Diese Profile stellen sicher, dass das Repository

nur Daten enthält, die formalen Mindestanforderungen genügen,
von externen Diensten wie Suchmaschinen oder Informationsdiensten gelesen werden kann,
in alternativen Metadatenformaten ausgegeben werden kann, und
externe Suchanfragen nach Meta- und Inhaltsdaten unterstützt.

Um die Erstellung CLARIN-kompatibler Metadaten zu erleichtern bietet das BAS den Webservice COALA an (s. Kap. 7).

Programmierschnittstellen und Protokolle

Aktuell wird das BAS Repository regelmäßig von den Harvestern des Virtual Language Observatory (VLO) (Zinn et al. 2015) sowie vom Informationsdienst Reuters über die Standardschnittstelle OAI-PMH ausgelesen, die unterstützten Metadatenformate sind Dublin Core und OLAC, und für die Suche in den Annotationsdaten wird über die Schnittstelle SRU-CQL angeboten.

Persistent Identifiers

Das BAS Repository verwendet EPIC Handle Persistent Identifiers (PID) für die Korpus- und Session-Objekte. Zum Beispiel verweist die PID http://hdl.handle.net/11858/00-1779-0000-0006-BF00-E auf die Webadresse des Repository.

Unterschiedliche Versionen eines Korpus bzw. einer Session erhalten jeweils eine eigene PID. Auf Signal- oder Annotationsfiles in einer Session kann mithilfe von Part-Identifiern zugegriffen werden (vorausgesetzt der User ist authentifiziert), hier z. B. ist die Part-ID m_0000000001 http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3@partId=m_0000000001

Die PIDs werden von der Gesellschaft für wissenschaftliche Datenverarbeitung in Göttingen (GWDG) verwaltet.

Zugangskontrolle

CLARIN strebt eine Single Sign-On Authentifizierung an. Als Protokoll wird Shibboleth (Knight 2015) verwendet. Für den Zugriff auf geschützte Ressourcen gibt die Nutzerin ihr Login mit Passwort ein, dieses wird von der Heimatinstitution überprüft. Diese Institution teilt dem Server, der die Ressource verwaltet, mit, ob die Nutzerin bekannt ist – wenn ja, dann darf sie in CLARIN diese und alle weiteren Ressourcen nutzen, für die sie autorisiert ist.

Im Wesentlichen gibt es drei Autorisierungsstufen: ACA für akademische Nutzer, RES für beschränkten Zugriff, und PUB für den unbeschränkten öffentlichen Zugang. Im Repository sind diese Stufen gut sichtbar aufgeführt, so dass sofort klar wird, welche Ressource wie zugänglich ist.

Aktuell sind die akademischen Institutionen der 15 CLARIN Mitgliedsländer (AT, BG, CZ, DK, EE, DE, GR, LI, NL, NO, PL, PT, SL, SE, UK) sowie die Nederlandse Taalunie als länderübergreifende Einrichtung zugangsberechtigt. Für Nutzer außerhalb akademischer Einrichtungen unterhält CLARIN ein eigenes Nutzerverzeichnis, so dass ausgewählte Mitglieder ebenfalls Zugriff auf CLARIN Ressourcen haben.

BAS Webservices

Das BAS betreibt eine Reihe von phonetisch-linguistischen Webservices (Schiel 2013) sowie web-basierte Schnittstellen BAS (2011-2016) , die auf diesen Webservices basieren.

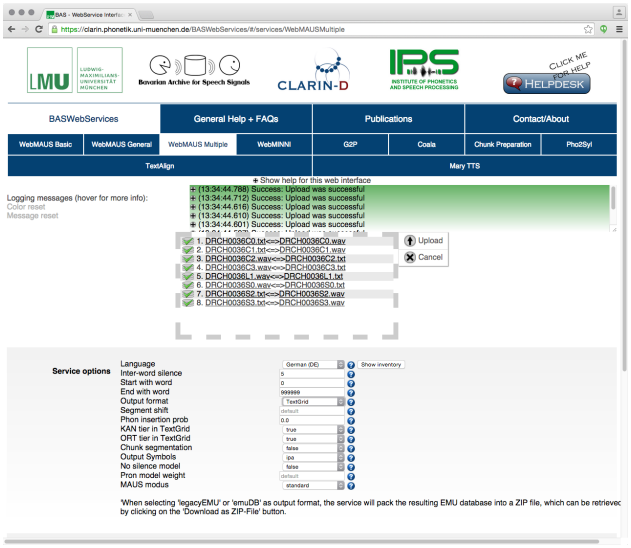


Abb. 3: WebMAUS mit 4 hochgeladenen Dateipaaren für die automatische Segmentation.

Ein Webservice erlaubt einem User oder einer Applikation mittels des REST-Protokolls automatische Verarbeitungen von Daten auf einem dedizierten Server durchzuführen. Zum Beispiel benutzt das Annotations-Tool ELAN (vgl. TLA: Elan) einen Webservices des BAS um während der Bearbeitung einer Sprachdatei eine vollautomatische Segmentierung anzustoßen (Kisler et al. 2012). Web-basierte Schnittstellen dagegen erlauben die benutzerfreundliche interaktive Verwendung solcher Webservices über einen Standard Web-Browser.

Das BAS hat in den vergangenen zwei Jahren seine vorhandenen Webservices erweitert, verbessert und neue Funktionalitäten bereitgestellt. Fast alle interaktiven Services erlauben jetzt auch das Batch-Processing von großen Datensammlungen.

Das bekannte System MAUS (vgl. Schiel 1999) zur vollautomatischen Segmentierung von Sprache wurde auf derzeit 17 Sprachvarianten erweitert. MAUS liest ein Sprachsignal-File und die dazugehörige Transkripton und berechnet daraus die Phonem/Wort-Segmentierung in Form einer hierarchischen Annotation.

Hervorzuheben sind dabei

die neue Verarbeitung von Schweizer Deutsch ausgehend von der sog. Dieth-Kodierung,
die vier Sprachvarianten amerikanisches, australisches, neuseeländisches und britisches Englisch,
sowie als neueste Sprachen Russisch und Französisch.

WebMAUS unterstützt seit kurzem auch das neue EMU Datenbank-System-Format und die sofortige Visualisierung der Ergebnisse mit Hilfe des Javascript-basierten Labeller-Tools EMU-webApp (Winkelmann et al. 2016) .

Ein neu entwickelter Webservice WebMINNI erlaubt die phonetische Segmentierung und Transkription von Medienfiles ohne vorhandene Verschriftung in sieben Sprachvarianten. Die stochastisch-phonologische Komponente des MAUS Systems wurde dabei ersetzt durch ein phontaktisches Bigram-Modell der jeweiligen Sprache.

Das BALLOON-Werkzeug G2P zur automatischen Generierung von Aussprache-Kodierung auf der Basis von orthographischem Text-Input wurde auf nunmehr 18 Sprachen erweitert (Reichel 2012).

Ein neu entwickeltes Tool Pho2Syl erlaubt die automatische Syllabifizierung von Transkriptionen. Syllabifizierungen werden in vielen Disziplinen der empirischen Linguistik benötigt. Pho2syl ist das erste verfügbare Werkzeug, das sowohl phonologische als auch phonetische Transkripte für 18 Sprachvarianten erlaubt.

Ein weitere Neuentwicklung ist TextAlign, ein allgemein einsetzbares Werkzeug zur optimalen symbolischen Alignierung (Reichel 2012).

Eine typische Anwendung ist die Alinierung von orthographischen zu phonologischen Symbol-Ketten. TextAlign bietet sowohl eine Vielzahl von

vortrainierten Kostenfunktionen für verschiedene Sprachvarianten als auch Mechanismen zur Abschätzung der optimalen Kostenfunktion aus den Input-Daten.

Für alle web-basierte Schnittstellen wurde die Benutzerfreundlichkeit kontinuierlich verbessert. Die BAS Webservices bieten dem Benutzer jetzt sowohl einen online Help Desk über die CLARIN Infrastruktur, eine FAQ-Sammlung als auch Tooltips und Usecase Beschreibungen direkt auf der Web-Seite.

Fazit und Ausblick

Das CLARIN Repository des BAS ist seit Ende 2012 in Betrieb und wird laufend erweitert. Der mit Abstand meistgenutzte Webservice ist WebMAUS - hier hat die Diskussion mit Anwendern auch dazu geführt, dass einzelne Bestandteile von MAUS mittlerweile als eigene Webservices verfügbar sind, z. B. G2P oder WebMINNI. Die zunehmende Erfahrung im Umgang mit Webservices hat dazu geführt, dass auch komplexe Arbeitsabläufe wie die Erstellung von Metadaten, oder früher wenig genutzte Dienste wie die Sprachsynthese, nun als Webservice zugänglich und damit wesentlich einfacher zu nutzen sind.

Der Wegfall der Notwendigkeit von Softwareinstallationen sowie die konsequent auf Nutzerfreundlichkeit ausgerichtete Gestaltung von Webservices hat dazu geführt, dass ganz neue Nutzerkreise erschlossen wurden: Ethnolog_innen lassen mit WebMAUS erste Rohsegmentationen bedrohter Sprachen erstellen, Toolentwickler_innen binden die Werbservices in ihre Tools ein, usw. Neben dem erwähnten ELAN nutzen in Studentenprojekten entwickelte innovative Prototypen von Sprachlerntools die MAUS-Segmentation für eine grafisch ansprechende Gegenüberstellung von Wort-, Silben- und Lautdauern von Muttersprachler- und Lerneräußerungen.

Zum Schluss ein Aufruf: das BAS sucht weiterhin Ressourcen gesprochener Sprache für das Repository. Insbesondere von Interesse sind Ressourcen gesprochener Sprache, deren Fortbestand gefährdet ist, oder die aus neuen, bislang unbekannten Forschungs- und Anwendungsbereichen stammen.

Bibliographie

BAS: Bavarian Archive for Speech Signals(2011-2016): *BAS WebService 2.06*. Institute of Phonetics and Speech Processing at the Ludwig-Maximilians-Universität München https://clarin.phonetik.uni-muenchen.de/BASWebServices/#/services [letzter Zugriff 08. Januar 2016].

CLARIN-D AP 5 (2012): *CLARIN-D User Guide: The Component Metadata Initiative (CMDI)*Clarin-D http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml [letzter Zugriff 08. Januar 2016].

Dickie, Catherine / Schaeffler, Felix / Draxler, Christoph (2009): "Speech Recordings via the Internet: An Overview of theVOYS project in Scotland", in: *Proc. Interspeech 2009* 1807-1810.

TLA:Elan (o. J.): *TLA (The Language Archive) Tools: Elan*. Nijmegen, Netherlands: Max Planck Institute for Psycholinguistics https://tla.mpi.nl/tools/tla-tools/elan/ [letzter Zugriff 08. Januar 2016].

Feiser, Hanna (2015): *Untersuchung auditiver und akustischer Merkmale zur Evaluation der Stimmähnlichkeit von Brüderpaaren unter forensischen Aspekten*. Frankfurt am Main: Verlag Polizeiwissenschaft

GWDG: *Gesellschaft für wissenschaftliche Datenverarbeitung mbH*. Göttingen: Georg-August-Universität Göttingen - Stiftung Öffentlichen Rechts, Max-Planck-Gesellschaft https://www.gwdg.de/ [letzter Zugriff 08. Januar 2016].

Kisler, Thomas / Schiel, Florian / Sloetjes, Han (2012): "Signal processing via web services: the use case WebMAUS", in: *Proceedings Digital Humanities 2012, Hamburg, Germany*: 30-34.

Knight, Justin (ed.) (2015): *Shibboleth* http://shibboleth.net/ [letzter Zugriff 08. Januar 2016].

Krefeld, Thomas / Stephan Lücke, Stephan / Mages, Emma(2009-2013): *Audioatlas Siebenbürgisch-Sächsischer Dialekte (ASD)*. Ludwig-Maximilians-Universität München [letzter Zugriff 08. Januar 2016].

Reichel, Uwe Dieter (2012): "PermA and Balloon: Tools for string alignment and text processing", in: *Proc. Interspeech. Portland, Oregon*: paper no. 346.

Schiel Florian (1999): "Automatic Phonetic Transcription of Non-Prompted Speech", in: *Proc. of the ICPhS 1999. San Francisco*: 607-610.

Schiel, Florian (2013): *BAS: Bavarian Archive for Speech Signals Webservices*. Universität Müchen http://www.phonetik.uni-muenchen.de/forschung/Bas/BasWebserviceseng.html [letzter Zugriff 08. Januar 2016].

Tissot, Fabienne (2015): *Gemeinsamkeit schaffen in der Interaktion* Diskursmarker und Lautelemente in zürichdeutschen Erzählsequenzen (= Sprache in Kommunikation und Medien 9). Bern / Berlin / Frankfurt am Main / New York / Paris / Wien: Peter Lang.

Wendt, Beate (2007): *Analysen emotionaler Prosodie*Hallesche Schriften zur Sprechwissenschaft und Phonetik. Bern / Berlin / Frankfurt am Main / New York / Paris / Wien: Peter Lang.

Winkelmann, Raphael / Raess, Georg / Jochim, Markus (2016):*EMU-webApp* Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität München https://github.com/IPS-LMU/EMU-webApp [letzter Zugriff 08. Januar 2016].

Zinn, Claus / Duin, Patrick / Stehouwer, Herman / Eckart, Thomas / Looij, Kees Jan van de / Goosen, Twan / Uytvan, Dieter van (2015): *Virtual Language Observatory 3.3.2* https://vlo.clarin.eu [letzter Zugriff 08. Januar 2016].

With a little help from my (HDC-)friends

Engelhardt, Claudia
claudia.engelhardt@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

Kurzawe, Daniel
kurzawe@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

Wuttke, Ulrike
uwuttke@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Buddenbohm, Stefan
buddenbohm@mmg.mpg.de
Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften

Der Einsatz digitaler Methoden in den Geisteswissenschaften bringt eine Fülle von digitalen Forschungsdaten und -ergebnissen hervor. Hierunter fallen zum einen “einfache”, dateibasierte Objekte, von denen die traditionell in den Geisteswissenschaften stark verbreiteten Textformate den größten Teil ausmachen, die aber unter anderem auch audiovisuelle Medien (Bild, Ton, Video) umfassen. Zum anderen spielen zunehmend auch komplexe digitale Objekte eine Rolle, die aus mehreren, miteinander verbunden Einzelobjekten bestehen, wie etwa Digitale Editionen oder verknüpfte Datenbanken.

Die Langzeitarchivierung und dauerhafte Bereitstellung dieser Vielfalt von Forschungsdaten und -ergebnissen stellt sowohl Datenzentren als auch Wissenschaftlerinnen und Wissenschaftler vor zahlreiche Herausforderungen. Neben der Entwicklung und Bereitstellung angemessener technischer Lösungen gehört hierzu das Wissen um den sorgfältigen und kundigen Umgang mit Forschungsdaten. Aspekte der Langzeitarchivierung und dauerhaften Bereitstellung von Forschungsdaten müssen idealerweise von Beginn an und über den gesamten Forschungsprozess hinweg beachtet werden, damit Hindernisse, die einer Archivierung bzw. Nachnutzung der Daten entgegenstehen, von vornherein ausgeschlossen und die Daten ohne vermeidbaren Mehraufwand ins Datenzentrum überführt werden können. Ein fachgerechtes Forschungsdatenmanagement, insbesondere komplexer Datentypen, erfordert jedoch entsprechendes Know-how und Ressourcen (hauptsächlich in Form von Arbeitszeit) - was sich im wissenschaftlichen Alltag oft als beträchtliches Hemmnis herausstellt. Untersuchungen zeigen entsprechend, dass sich Wissenschaftlerinnen und Wissenschaftler sachkundige Unterstützung beim Datenmanagement und der Langzeitarchivierung wünschen (vgl. bspw. Simukovic et al. 2013: 30ff.; Feijen 2011: 28).

In seinem in der Designphase (Mai 2014 - April 2016) entwickelten Konzept für ein geisteswissenschaftliches Forschungsdatenzentrum trägt das Humanities Data Centre (HDC 2014-2016) dieser Situation mit dem Entwurf eines stark nutzerorientierten Angebotsportfolios Rechnung. Das Portfolio kombiniert die technologischen Lösungen für die Langzeitarchivierung und Bereitstellung der Daten mit umfangreichen Angeboten zur Beratung, Unterstützung und Schulung von Wissenschaftlerinnen und Wissenschaftlern. Die tragende Säule des Beratungs- und Schulungskonzepts ist dabei ein verteiltes Netzwerk von Datenkuratorinnen und -kuratoren, die sowohl am Datenzentrum, als auch an assoziierten wissenschaftlichen Forschungseinrichtungen angesiedelt sind. Sie sollen zukünftig die Schnittstelle zwischen dem HDC und der Wissenschaft bilden und die Wissenschaftlerinnen und Wissenschaftler direkt vor Ort unterstützen und beraten.

Die Form und das Ausmaß der Zusammenarbeit der Datenkuratorinnen und -kuratoren mit den Forschenden wird, je nach Projektkonstellation, unterschiedlich sein. Grundsätzlich ist zwischen geplanten und abgeschlossenen Projekten zu differenzieren, da dieser Unterschied typische Implikationen für die (Beratungs-)Tätigkeit der Datenkuratorinnen und -kuratoren mit sich bringt.

Im Idealfall beginnt die Kooperation zwischen Wissenschaftlerinnen und Wissenschaftlern und Datenkuratorinnen und -kuratoren bereits in der Konzeptionsphase eines Projektes, in der gemeinsam der Umgang mit den Daten geplant, ein Datenmanagementplan erstellt, der Aufwand für das Datenmanagement abgeschätzt und idealerweise dafür auch Mittel beim Förderer beantragt werden. Das Projekt wird auch im weiteren Verlauf von der / dem Datenkurator/-in begleitet, z. B. in Form von Beratungsgesprächen an (für das Datenmanagement bzw. die spätere Langzeitarchivierbarkeit und Nachnutzbarkeit der Daten) neuralgischen Punkten bis hin zur Übergabe der Daten an das Datenzentrum. Da bereits während der Projektlaufzeit ein sorgfältiges Datenmanagement erfolgte, ist in diesem Fall mit einem vergleichsweise geringen Zusatzaufwand für die Aufbereitung der Daten für die Übergabe ins Datenzentrum und die Nachnutzung zu rechnen. Anders ist die Situation bei abgeschlossenen Projekten, insbesondere wenn die Forschungsdaten und -ergebnisse komplexer Natur sind. Wenn hier kein gutes Datenmanagement stattgefunden hat, ist die Gefahr groß, dass sich die Übergabe der Daten ins Datenzentrum und die Aufbereitung zur späteren Nachnutzung ungleich aufwendiger gestaltet. In diesem Fall muss zunächst einmal eine “Anamnese” durchgeführt werden, im Rahmen derer eruiert wird, ob und inwieweit Langzeitarchivierung und Nachnutzung der

Daten technisch und rechtlich möglich sind, eine Lösung für die Langzeitarchivierung entwickelt und die Daten anschließend entsprechend aufbereitet werden, bevor die Überführung ins Datenzentrum stattfinden kann.

Das Poster illustriert die Aufgaben der zukünftigen HDC-Datenkuratorinnen und -kuratoren am konkreten Beispiel.

Das HDC-Konsortium in der Designphase besteht aus der Akademie der Wissenschaften zu Göttingen (ADWG), der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW), der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), dem Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), dem Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften Göttingen (MPIMMG) und der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB). Assoziierter Partner ist das Max-Planck-Institut für Wissenschaftsgeschichte Berlin (MPIWG).

Das Projekt wird vom Niedersächsischen Ministerium für Wissenschaft und Kultur (Niedersächsisches Vorab) gefördert.

Bibliographie

Feijen, Martin (2011): *What researchers want*. A literature study of researchers' requirements with respect to storage and access to research data. Utrecht: Surffoundation https://www.surf.nl/binaries/content/assets/surf/en/knowledgebase/2011/What_researchers_want.pdf [letzter Zugriff 15. Oktober 2015].

HDC (2014-2016): *HDC - Projekt Humanities Data Centre*. Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen / Akademie der Wissenschaften zu Göttingen / Berlin-Brandenburgische Akademie der Wissenschaften / Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften / Niedersächsische Staats- und Universitätsbibliothek Göttingen / Konrad-Zuse-Zentrum für Informationstechnik / Max-Planck-Institut für Wissenschaftsgeschichte http://humanities-data-centre.org/ [letzter Zugriff 08. Januar 2016].

Simukovic, Elena / Kindling, Maxi / Schirmbacher, Peter (2013): *Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin*. Umfragebericht, Version 1.0. URN: urn:nbn:de:kobv:11-100213001, Computer- und Medienservice, Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin http://nbn-resolving.de/urn:nbn:de:kobv:11-100213001 [letzter Zugriff 15. Oktober 2015].

Kuration und Exploration des Korpus "Diskurs in der Weimarer Republik"

Fankhauser, Peter
fankhauser@ids-mannheim.de
IDS-Mannnheim, Deutschland

Auch in Zeiten von „Big Data“ haben relativ kleine, auf eine spezifische Fragestellung hin zugeschnittene und aufbereitete Korpora ihre Bedeutung. In diesem Beitrag beschreiben wir die Aufbereitung eines solchen Korpus für die nachhaltige Langzeitarchivierung und skizzieren die sich daraus ergebenden Möglichkeiten zur explorativen Analyse.

Das Korpus „Diskurs in der Weimarer Republik“ (DWR) wurde im Rahmen des Projektes „Demokratiediskurs 1918-1925“ (Kämper 2014) zur Dokumentation und Analyse des sprachlichen Wandels im Umbruch von der Monarchie zur Demokratie erstellt. Es umfasst 779 Dokumente im Zeitraum von 1912 bis 1933, davon 641 zwischen 1918 und 1925. 551 Dokumente sind (u. a.) nach Themenbereich und Textsorte klassifiziert (s. Tabelle 1).

Themenbereich	Abk	**#Docs**	#Wörter	Textsorte	Abk	**#Docs**	#Wörter
Politik	PT	231	364.077	Zeitungsartikel	Z	122	125.733
Frauen	FR	129	268.697	Manifest	I	117	94.430
Jugend	JU	70	89.727	Brief	B	78	57.542
Intellektuellendiskurs	IN	78	664.410	Rede	R	77	157.347
Zionismus	ZI	23	33.515	Essay	E	37	92.310
Kirche	KI	20	29.113	Kundgebung	K	18	11.893
Summe		551	1.449.539	Sonstige(9)		102	910.284

Tab. 1: Themenbereiche und Auswahl an Textsorten im DWR

Ursprünglich wurde das Korpus im Rich-Text-Format (RTF) bzw. MS-Office (DOC) erstellt, und die Metadaten in einer Oracle-Datenbank verwaltet. Im Rahmen des LIS-Projektes „Zentrum für germanistische Forschungsprimärdaten“ [1] wurde das Korpus für die Langzeitarchivierung aufbereitet. Im Einzelnen wurden folgende Schritte durchgeführt:

Alignierung und Bereinigung der Metadaten: Die Verknüpfung von Metadaten mit Dokumenten war über Dateinamen repräsentiert, die teilweise nicht einheitlich enkodiert waren. Diese wurden entsprechend normalisiert, um einen eindeutigen Bezug herzustellen. Darüber hinaus wurden die Wertebereiche der einzelnen Metadatenfelder von Tippfehlern (z. B. Poitik vs. Politik) und Enkodierungsproblemen weitestgehend bereinigt.

Validierung und Kuration der Datenformate: Die vorhandenen RTF-Versionen und DOC-Versionen wurden mithilfe von Open-Office-Macros in valides RTF transformiert. Zur besseren Nachnutzbarkeit wurde zusätzlich mit Hilfe des TEI Open-Office Pakets *teioop5* eine valide TEI-P5-XML-Version erstellt, die mit Metadaten für Autor, Titel und Erscheinungsjahr angereichert wurde. Zudem wurde auch eine PDF-Leseversion erzeugt.

Extraktion zusätzlicher Metadaten: Die in den Dokumenten vorhandenen bibliographischen Quellenangaben wurden mit Hilfe heuristischer Regeln extrahiert und in die Metadaten integriert.

Generierung von CMDI-Metadaten: Die Metadaten wurden in das CLARIN-Metadatenframework CMDI (Broeder et al. 2011) transformiert.

Das aufbereitete Korpus [2] ist im Langzeitarchiv des IDS [3] (Fankhauser et al. 2013) abgelegt.

Zur Exploration sprachlicher Variation im Korpus wurde das Korpus zudem für ein am Institut für Deutsche Sprache entwickeltes System zur kontrastiven Visualisierung von Korpora (Fankhauser et al. 2014a, 2014b) aufbereitet.

Dafür wurde das Korpus an Hand der Metadaten für Themenbereiche und Textsorten in Teilkorpora aufgeteilt, und für die einzelnen Teilkorpora Frequenzlisten aller Wörter (ohne Lemmatisierung oder Stopwortausschluss) erstellt. Diese Frequenzlisten, repräsentiert als multinomiale Verteilungen über das Vokabular, werden mit Hilfe der Kullback-Leibler Divergenz verglichen. Auf dieser Basis wird die Distanz zwischen Teilkorpora in Form von Heatmaps visualisiert, und der Beitrag einzelner Wörter zu der jeweiligen Distanz mit Hilfe von Wortwolken.

Zur Exploration sprachlicher Variation im Korpus wurde das Korpus zudem für ein am Institut für Deutsche Sprache entwickeltes System zur kontrastiven Visualisierung von Korpora (Fankhauser et al. 2014a, 2014b) aufbereitet.

Dafür wurde das Korpus an Hand der Metadaten für Themenbereiche und Textsorten in Teilkorpora aufgeteilt, und für die einzelnen Teilkorpora Frequenzlisten aller Wörter (ohne Lemmatisierung oder Stopwortausschluss) erstellt. Diese Frequenzlisten, repräsentiert als multinomiale Verteilungen über das Vokabular, werden mit Hilfe der Kullback-Leibler Divergenz verglichen. Auf dieser Basis wird die Distanz zwischen Teilkorpora in Form von Heatmaps visualisiert, und der Beitrag einzelner Wörter zu der jeweiligen Distanz mit Hilfe von Wortwolken.

Abbildung 1 zeigt die Distanz zwischen Themenbereichen sowie zwischen Textsorten innerhalb eines Themenbereichs (grün für geringe, purpur für große Distanz). Es wird deutlich, dass der Themenbereich *Kirche* (KI) sich am deutlichsten von den anderen Themenbereichen abhebt. Innerhalb der Themenbereiche zeigt sich, dass die Textsorten - soweit für einen Themenbereich mit Dokumenten belegt - im Themenbereich *Frauen* deutlich stärker ausdifferenziert sind als im Themenbereich *Politik*. Inbesondere die Textsorten *Stellungnahme* (S) und *Kundgebung* (K) heben sich deutlicher von den anderen Textsorten ab als im Themenbereich *Politik*.

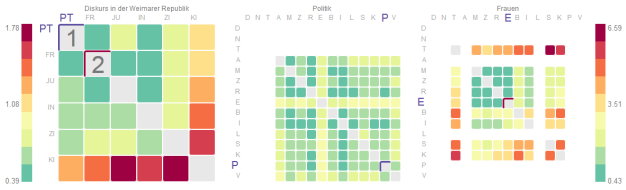


Abb. 1: Heatmaps für den Vergleich von Themenbereichen (links) und Textsorten innerhalb eines Themenbereichs (*Politik*: mitte, *Frauen*: rechts).

Abbildung 2 zeigt den Beitrag einzelner Wörter zu der Distanz zwischen Teilkorpora in Form von Wortwolken. Groß dargestellte Wörter sind hierbei besonders typisch für ein Teilkorpus, die Farbe korrespondiert mit der relativen Häufigkeit eines Wortes im Teilkorpus (blau für selten, purpur für häufig). Die Wortwolke links vergleicht *Frauen* mit dem restlichen Korpus. Sie wird sowohl auf begrifflicher Ebene (*Frau/Mann*) als auch auf grammatischer Ebene (*die, ihre, sie, ...*) vom allgemeinen Diskursgegenstand *Frauen* dominiert. Die Wortwolke in der Mitte zeigt die typischen Wörter von *Zeitungsartikeln* im Vergleich zu *Essays* innerhalb des Themenbereichs *Frauen*, die Wortwolke rechts typische Wörter im umgekehrten Vergleich. Hier wird deutlich, dass *Zeitungsartikel* sich im wesentlichen um die politisch/öffentliche Stellung der Frau drehen (*Wahlrecht, Frauenstimmrecht, politische*) und *Essays* um die private Welt der Frau (*Beziehung, Moral, Erotik*). Ein sehr deutlicher Unterschied zeigt sich auch im Numerus von *Frau:* Plural in *Zeitungsartikeln* und Singular in *Essays*.

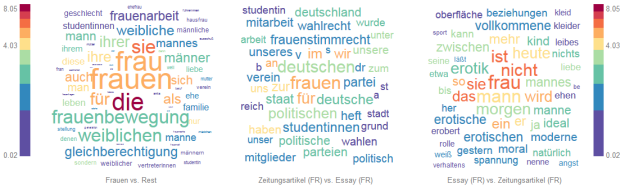


Abb. 2: Wortwolken für die typischen Wörter des Themenbereichs *Frauen* im Vergleich mit dem restlichen Korpus (links) und in den Textsorten *Zeitungsartikel* vs. *Essay* im Themenbereich *Frauen* (mitte und rechts).

Dieser kurze explorative Überblick kann natürlich nur einen kursorischen Eindruck über Inhalt und Vielfalt des Korpus geben. Technisch wurde er erst möglich durch die konsequente Kuration der Metadaten und Daten an Hand der generellen Richtlinien der CLARIN Infrastruktur.

Notes

1. Das Zentrum für germanistische Forschungsprimärdaten , wird gefördert von der DFG im Rahmen des Programms „Informationsinfrastrukturen für Forschungsdaten“.
2. Korpus: „Diskurs in der Weimarer Republik“ PID: http://hdl.handle.net/10932/00-01B9-43B3-1E1D-7B01-6
3. Siehe IDS-Repositorium .

Bibliographie

Broeder, Dan / Schonefeld, Oliver / Trippel, Thorsten / Van Uytvanck, Dieter / Witt, Andreas (2011): "A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI)", in: *Proceedings of Balisage*. The Markup Conference 2011 (= Balisage Series of Markup Technologies 7).

Fankhauser, Peter / Fiedler, Norman / Witt, Andreas (2013): "Forschungsdatenmanagement in den Geisteswissenschaften am Beispiel der germanistischen Linguistik", in: *Zeitschrift für Bibliothekswesen und Bibliographie (ZfBB)* 60, 6: 296-306.

Fankhauser, Peter / Knappen, Jörg / Teich, Elke (2014a): "Exploring and Visualizing Variation in Language Resources", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*

Fankhauser, Peter / Kermes, Hannah / Teich, Elke (2014b): "Combining Macro- and Microanalysis for Exploring the Construal of Scientific Disciplinarity", in: *Proceedings of the Digital Humanities 2014*.

Institut für Deutsche Sprache (IDS): *Zentrum für germanistische Forschungsprimärdaten* http://www1.ids-mannheim.de/fi/projekte/lis.html [letzter Zugriff 11. Februar 2016].

Institut für Deutsche Sprache (IDS): *IDS Repository* https://repos.ids-mannheim.de/ [letzter Zugriff 11. Februar 2016].

Kämper, Heidrun (2015): "Demokratiediskurs 1918-1925" http://www1.ids-mannheim.de/lexik/zeitreflexion18.html [letzter Zugriff 14. Oktober 2015].

„Bis zum Sankt(- \s)? [Nn]immerleins(- \s)?[Tt]ag“– der Datumserkenner „PDR-Dates“

Fechner, Martin
fechner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Körner, Fabian
fkoerner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Einleitung

Die Idee für einen Datumserkenner „PDR-Dates“ entwickelte sich 2009 während des Anfangsstadiums des von der DFG geförderten Personendaten-Repositoriums an der Berlin-Brandenburgischen Akademie der Wissenschaften . Im Zuge dieses Projektes wurde eine Lösung für die gemeinsame Speicherung und Bereitstellung von Informationen aus heterogenen historischen Personendatenbeständen geschaffen. Diese Bestände stammen vorrangig aus laufenden und abgeschlossenen Projekten der BBAW.

Neben der zu erstellenden Software-Umgebung lag also von Anfang an ein wesentlicher Anteil der Arbeit des PDR im Bereich der Migration bestehender Informationsmengen aus ihrem wie auch immer gearteten Ausgangsformat in das Format des PDR.

Dabei fiel auf, dass man bei dieser Gelegenheit zwei Fliegen mit einer Klappe schlagen und versuchen kann, die Strukturierung der Informationen nicht nur zu übernehmen, sondern zu verbessern. Dabei ist extrem hilfreich, dass man den genauen Kontext der Inhalte einer Informationsmenge nutzen kann, um das global extrem komplexe Problem der Erkennung von Datumsangaben soweit zu reduzieren, das ein realisierbares Werkzeug für die Automatisierung geschaffen werden kann: der Datumserkenner „PDR-Dates“.

Funktionsweise

Ziel der Datumserkenners „ PDR-Dates“ ist die Identifizierung von natürlich-sprachlichen Datumsangaben in verschieden-sprachigen Texten, um im Sinne des Data Retrieval einen Mehrwert zu erzielen. Natürlich-sprachliche Datumsangaben sollen hierfür in das Standardformat nach ISO 8601 (s. International Organization of Standardization und Wikipedia) umgewandelt werden, damit sie maschinenlesbar sind. Zu diesem Zweck wurde eine Java-Bibliothek programmiert, die in Forschungsumgebungen oder in Web Services integriert werden kann. Beispiele hierfür sind die Zeitraumangaben des Webservices „ correspSearch “ oder die Webservices des PDR , die mit „PDR-Dates“ arbeiten. Der Datumserkenner „PDR-Dates“ kann sowohl einzelne Zeitpunkte, als auch Zeiträume erkennen.

Der Datumserkenner baut auf der syntaktischen Mustererkennung durch reguläre Ausdrücke auf. Um komplexere Zeitangaben in Texten erkennen zu können, werden drei Schritte angewandt: (1) Der tokenisierte Text wird mit regulären Ausdrücken geprüft, ob die einzelnen Tokens für eine Datumsangabe relevante Informationen enthalten können. (2) Über mehrere klassifizierte Tokens hinweg wird nach definierten Mustern gesucht. Diese Mustererkennung wird mit einer Vielzahl von Mustern über dem gleichen Text wiederholt, so dass auch lange zusammengesetzte Datumsangaben erkannt werden können. (3) Schließlich werden alle erkannten Datumsangaben hinsichtlich ihrer Bedeutung interpretiert. Dafür wird auf alle zusammengetragenen Informationen zurückgegriffen.

(1) Die regulären Ausdrücke unterteilen die Tokens in einzelne Klassen, so wird: "Anfang", "Januar", "2016" als "approximation", "month01", "d4" erkannt. Neben Zahlen und Zahlausdrücken werden Feiertage, Monatsnamen, Jahreszeiten, Näherungsangaben, Jahrhundertangaben und Wörter mit Sonderfunktion erkannt.

(2) Die Mustererkennung gibt den einzelnen Tokens eine vorläufige Bedeutung für die spätere Interpretation (etwa "März 1800" als "month_yyyy"). Da im Text iterativ nach verschiedenen Mustern gesucht wird, ist es möglich schon erkannte Datumsangaben durch Konkretisierungen in Form von Prefix- oder Suffix-Mustern zu erweitern. Jede so erkannte Datumsangabe bezeichnet entweder einen Zeitpunkt ("1.1.1800" als "d_m_yyyy") oder es ist möglich, dass durch ungefähre Angaben ein Zeitraum bezeichnet wird ("Anfang März 1800" als "approximation_month_yyyy"). Auch kann erkannt werden, ob zwei schon erkannte Datumsangaben einen Zeitraum bezeichnen ("von Dezember 1800 bis Januar 1805" als "limit_month_yyyy_to_month_yyyy").

(3) Bei der Interpretation der Muster werden alle festgestellten Informationen für die Verarbeitung zum Format nach ISO 8601 genutzt. Einige Tokens erhalten je nach Positionierung im Muster eine andere Interpretation, so bezeichnet der Term "Anfang" in "Anfang März" und "Anfang 1800" jeweils unterschiedlich lange Zeiträume. Auch können feste Feiertage ("Mariä Empfängnis 1800"), sowie von Ostern und dem Jahr abhängige Feiertage ("Pfingstmontag 1800") interpretiert werden. Handelt es sich bei einem oder beiden Daten bereits um eine Zeitspanne, wird der volle Zeitraum als Zieldatum ausgegeben.

(Beispiel über den Web Service des PRD mit dem Text „Die nächsten Semesterferien dauern von Mitte Februar bis Mitte April 2016“ findet sich hier:

https://pdrprod.bbaw.de/pdrws/dates?lang=de&text=Die%20n%C3%A4chsten%20Semesterferien%20dauern%20von%20Mitte%20Februar%20bis%20Mitte%20April%202016&output=xml)

Erweiterungen und Begrenzungen

Mit Hilfe einer Konfigurationsdatei in XML ist es möglich eine eigene Java-Bibliothek zu erzeugen. Damit kann die Datumserkennung an einzelne Forschungskontexte angepasst und dort zwischen möglichen Kooperationspartnern ausgetauscht werden.

Mit dem geschilderten Vorgehen werden ausschließlich vollständige Datumsangaben erkannt. Für eine Interpretation von Datumsangaben, die sich nur relativ zu einem Bezugsdatum interpretieren lassen (etwa: "letzte Woche"), müsste die syntaktische Mustererkennung auch um eine semantische Mustererkennung erweitert werden. Die bereitgestellte Bibliothek (zu erreichen über den Web Service des PDR, für die APIs https://pdrprod.bbaw.de/pdrws/dates?doc=api) erkennt Datumsangaben in deutsch, englisch und italienisch. Mit der geschilderten Konfigurationsdatei ist eine Erweiterung aber auch ohne Programmierkenntnisse möglich.

Bibliographie

correspSearch (2015): *correspSearch – Verzeichnisse von Briefeditionen durchsuchen*. http://correspsearch.bbaw.de/search.xql?l=de [letzter Zugriff 15. Oktober 2015].

International Organization for Standardization (o. J.): *Date and time format – ISO 8601*. http://www.iso.org/iso/iso8601 [letzter Zugriff 15. Oktober 2015].

Personendaten-Repositorium (2012): *Webservices – Personendaten-Repositorium*. http://pdr.bbaw.de/software/webservices/ [letzter Zugriff: 15. Oktober 2015].

Wikipedia: *ISO 8601*. https://de.wikipedia.org/wiki/ISO_8601 [letzter Zugriff 15. Oktober 2015].

Distant-Reading-Showcase: 200 Jahre deutscher Dramengeschichte auf einen Blick

Fischer, Frank
frank.fischer@zentr.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Vogel, Andreas
andrea.uccello@gmail.com
Universität Leipzig

Göbel, Mathias
goebel@sub.uni-goettingen.de

Niedersächsische Staats- und Universitätsbibliothek Göttingen

Trilcke, Peer
trilcke@phil.uni-goettingen.de
Georg-August-Universität Göttingen

Kampkaspar, Dario
kampkaspar@hab.de
Herzog August Bibliothek Wolfenbüttel

Kittel, Christopher
c.kittel@edu.uni-graz.at
Karl-Franzens-Universität Graz

Der Terminus 'Distant Reading' wurde im Jahr 2000 von Franco Moretti geprägt. Ebenfalls auf Moretti zurück geht das Konzept der "Maps, Graphs, Trees" (2005), die Nutzbarmachung verschiedener Visualisierungsmethoden für literaturhistorische Daten. Die Beispiele für die Zusammenführung von Distant-Reading- und Visualisierungs-Ansätzen haben als Hintergrund jedoch fast ausschließlich englischsprachige Korpora.

Im Mittelpunkt unseres Distant-Reading-Showcase-Posters steht nun mit der "Digital Bibliothek" das größte deutschsprachige Korpus literarischer Texte. Auf einem einzigen A0-Poster werden die Figurennetzwerke von 465 deutschsprachigen Dramen aus den Jahren 1730 bis 1930 gezeigt. Dabei werden verschiedene semantischen Dimensionen vereint. Zum einen folgt die Positionierung der Dramen chronologisch. Neuralgische Punkte der deutschen Literaturgeschichte werden sofort sichtbar, etwa die Explosion des Figurennetzwerks in Goethes "Götz von Berlichingen" von 1773. Das bekannte Faktum der damals einsetzenden verstärkten Shakespeare-Lektüre wird so auf einen Blick erkennbar und erscheint im Kontext des zeitlichen Davor und Danach.

Eine weitere semantische Ebene ist die Abbildung der Figurennetzwerke mit dem clusternden Visualisierungsverfahren Fruchterman–Reingold. Dadurch werden jenseits der Chronologie (teils bisher unbekannte) Gemeinsamkeiten bei der Konstruktion von Dramen über zwei Jahrhunderte hinweg sichtbar. Eine zusätzliche semantische Ebene bilden die Namen tausender Figuren, wie sie die Bühnen und Dramenanthologien zweiter Jahrhunderte bevölkert haben. Dabei fällt nicht nur auf, dass darunter etwa 24 Faust-Figuren sind (inklusive einer weiblichen "Faustine"). Dieses Wimmelbild der deutschen Dramenliteratur ist zugleich ein möglicher Wiedereinsteig ins Close Reading, der zeigt, dass sich Close und Distant Reading nicht ausschließen, sondern ergänzen.

Die Vorarbeiten zu diesem Poster wurden bereits in Graz präsentiert (Trilcke et al. 2015). Der derzeitige Stand des Projekts erlaubt es uns, auf einem Schaubild die aktuellen Ergebnisse zu verschränken. Dabei möchten wir mit dem Genre 'Poster' (unter Beachtung der Data-Ink Ratio nach Edward Tufte) versuchen, digital betriebene Forschung so ins Zielformat zu übersetzen, dass das Ergebnis als wissenschaftlicher Showcase für das Feld des Distant Reading dienen kann.

Bibliographie

Moretti, Franco (2000): "Conjectures on World Literature", in: *New Left Review* 1: 54–68.

Moretti, Franco (2005): *Maps, Graphs, Trees*. Abstract Models for a Literary History. London / New York: Verso.

Trilcke, Peer / Fischer, Frank / Kampkaspar, Dario (2015): "Digitale Netzwerkanalyse dramatischer Texte (Vortrag)", in: *Tagung der Digital Humanities im deutschsprachigen Raum*, Graz. Slides: http://gams.uni-graz.at/o:dhd2015.v.040 .

Aufbau einer Korpusinfrastruktur für die Beobachtung des Schreibgebrauchs

Fischer, Peter M.
peter.fischer@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Diewald, Nils
diewald@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Kupietz, Marc
kupietz@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Witt, Andreas
witt@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Mit dem Ziel, eine systematische Beobachtung des Schreibgebrauchs unter Verwendung computerlinguistischer Methoden zu ermöglichen, wurde 2013 das vom BMBF geförderte Forschungsprojekt *Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen* ins Leben gerufen. An diesem beteiligen sich gemeinschaftlich das Institut für Deutsche Sprache, das Institut für Computerlinguistik der Universität des Saarlandes, sowie die Wörterbuchverlage Bibliographisches Institut GmbH (Dudenverlag) und Wahrig bei Brockhaus.

Das Projekt hat sich u.a. zur Aufgabe gemacht, eine zweckdienliche Datengrundlage (Fischer i.E.) und ein dazugehöriges Methodeninventar (Scholze-Stubenrecht 2013) aufzubauen.

Für die Erstellung von Korpusanalysen mit Auswertung nach eigens erarbeiteten Bewertungskriterien (Krome, 2013) ist das Projekt auf eine geeignete Korpusinfrastruktur angewiesen, die es den beteiligten Partnern erlaubt, entsprechende Suchanfragen auf den einerseits umfangreichen (über 10 Mrd. Tokens), andererseits aus datenschutz- und urheberrechtlichen Gründen mitunter verteilt liegenden Ressourcen effizient und zuverlässig durchzuführen. Dabei wird entsprechend Jim Grays (Gray 2003) Maxime "put the computation near the data" (Kupietz et al. 2014) der Ansatz verteilter virtueller Korpora bzw. Kollektionen (van Uytvanck 2010) verfolgt, der darauf abzielt, dedizierte, auf die spezifischen Suchanfragen ausgerichtete Subkorpora zu definieren und auf diesen rechtskonform zu operieren.

KorAP (Bański et al. 2013) ist eine Such- und Analyseplattform, die eine solche Infrastruktur zur Verfügung stellt. Sie wurde als Nachfolgesystem von COSMAS-II (Bodmer 1996) am Institut für Deutsche Sprache als primäre Schnittstelle für den Zugriff auf DeReKo (Kupietz / Lüngen 2014), das Deutsche Referenzkorpus, entwickelt. KorAP ermöglicht die Suche in sehr großen, mehrfach annotierten, und heterogen lizenzierten Korpora über eine Vielzahl von Suchoperatoren verschiedener Anfragesprachen. Die dynamische Erstellung virtueller Korpora wird dabei durch Kombination von Metadatenkriterien realisiert. Dies steht auch im Einklang mit dem Projektdesiderat, die Anbindung an die europäische Sprachressourceninfrastruktur CLARIN, die bereits eine Fülle von Werkzeugen anbietet, zu intensivieren und damit die Sichtbarkeit der Ressourcen auch im internationalen Kontext zu erhöhen.

Dieses Paper beleuchtet damit jene Arbeiten, die sich mit dem Prozess des Aufbaus der Korpusinfrastruktur, d.h. der Aufbereitung, Organisation und Bereitstellung der Datengrundlage befassen.

Als empirische Basis des Projektes dient die parallele Beobachtung und Auswertung von drei Zielgruppen und Ebenen der Textproduktion, nämlich die der professionellen Schreiber (in Zeitungen, Zeitschriften usw.), die den Schreibgebrauch der Schreibgemeinschaft heute entscheidend mitbestimmen, die der Schüler (in Klassenarbeiten, Abituraufsätzen, Literaturwettbewerben usw.), die als Repräsentanten der jungen Generation im schulischen Kontext an die amtlichen Regeln zur Rechtschreibung gebunden sind, und die der Internetnutzer (in E-Mails, sozialen Netzwerken, Meinungsportalen usw.), die in einer im Vergleich zu Druckerzeugnissen weniger kontrollierten Umgebung Entwicklungs- und Fehlertendenzen viel früher und deutlicher wiedergeben können als das beispielsweise in Zeitungstexten oder belletristischen Korpora der Fall ist. Dementsprechend steuern diese drei heterogenen Quellen auch unterschiedliche Informationen bei und stellen den Aufbau der Korpusinfrastruktur vor individuelle Herausforderungen.

Aus korpustechnologischer Sicht konnte das Projekt in Teilen auf bereits vorhandene, wohlstrukturierte und linguistisch aufbereitete Ressourcen wie das Deutsche Referenzkorpus DeReKo (Kupietz / Lüngen 2014), das WAHRIG Textkorpus $^{\text{Digital}}$ (Krome 2010) oder das Dudenkorpus (Münzberg 2011) zurückgreifen, während andere erst akquiriert, für eine maschinelle Verarbeitung vorbereitet und mit linguistischen Informationen angereichert werden mussten. Da entsprechende sprachtechnologische Verfahren (Tokenisierung, Lemmatisierung, Wortart-Tagging, flache syntaktische Analyse) jedoch überwiegend für stärker kontrollierte Texte entwickelt wurden und daher nicht auf alle diese drei Quellen gleichermaßen anwendbar sind, mussten überdies zunächst geeignete Werkzeuge (weiter-)entwickelt werden (Horbach et al. 2015), um einen für Vergleichsanalysen ausgewogenen Aufbereitungsstand zu erreichen.

Neben diesen linguistischen Merkmalen verfügen die Texte auch über gewisse Metadaten, die aber in Struktur und Ausprägung stark an den Ressourcenkontext gebunden sind und deshalb mitunter entsprechend heterogen ausfallen. Das Zurückgreifen auf diese Informationen stellt jedoch bei synchronen wie diachronen Auswertungen ein für die systematische Beobachtung des Schreibgebrauchs zentrales Nutzungsszenario dar, das eine ordentliche Zusammenstellung solcher Zusatzinformationen erfordert. Folglich ist für die Erstellung virtueller Korpora und damit für ihre anfrageoptimierte Bereitstellung innerhalb der Analyseinfrastruktur die Erfassung von Metadaten unerlässlich. Die folgende Aufstellung zeigt eine Übersicht der Ressourcentypen und ihrer Metadaten.

Texte professioneller Schreiber (am Beispiel Zeitschriftenkorpus)

- Name der Zeitung
- Nummer der Ausgabe
- Titel des Artikels
- Untertitel des Artikels
- Name des Autors
- Ort der Veröffentlichung
- Tag der Veröffentlichung
- Textklasse (z.B. Wirtschaft oder Sport)
- Textsorte (z.B. Gerichtsurteil oder Satire)

Schülertexte (am Beispiel Literaturwettbewerbskorpus)

- Name des Wettbewerbs
- Jahrgang (Einsendeschluss)
- Titel des Textes
- Altersklasse des Autors

- Geschlecht des Autors

Internettexte (am Beispiel Zeitungsleserkommentarkorpus)

- Name der Zeitung
- Titel des Artikels
- Teaser des Artikels
- Schlagwörter zum Artikel
- Tag der Artikelveröffentlichung
- Pseudonym des Kommentarautors
- Tag der Kommentarveröffentlichung
- Titel des Kommentars

Die Grundstrukturierung der Datenbasis samt aller Annotationen und Metadaten erfolgt einheitlich gemäß den Vorgaben von TEI P5 (TEI Consortium 2007), das als auf das Kodieren von Textkorpora ausgerichtetes und auf XML aufbauendes Datenformat einen langjährig etablierten Standard zur Strukturierung linguistischer Daten darstellt. Zur Auszeichnung der Wortartinformationen (POS) wurde das Stuttgart-Tübingen-Tagset STTS (Schiller et al. 1999) herangezogen, bzw. im Falle der nicht-professionellen Textsubstanzen um Elemente aus STTS 2.0 (Bartz et al. 2014), einer abwärtskompatiblen Weiterentwicklung, die speziell auf die Anwendung auf Ressourcen aus internetbasierter Kommunikation optimiert wurde, ergänzt. Die TEI-kodierten Daten werden daraufhin in die interne KorAP-Repräsentation überführt und indiziert. Für vorhandene Metadaten werden optimierte Indizierungsstrategien gewählt, um beispielsweise eine Kriterienwahl über reguläre Ausdrücke oder Zahlenbereiche zu ermöglichen.

Leider dürfen die von den Projektpartnern separat aufgebauten bzw. dort bereits vorliegenden Korpora aus datenschutz- und urheberrechtlichen Gründen jedoch nicht als solche an die jeweils anderen Partner weitergegeben, damit also auch nicht an einem Ort zentral zusammengetragen werden. Dieser Umstand verteilt liegender Ressourcen erfordert die Schaffung einer Möglichkeit, zentrale Anfragen parallel an die einzelnen real existierenden Korpora zu stellen und in einem zweiten Schritt die Resultate der jeweiligen Standorte konzertiert zusammenzuführen.

Dafür wurde die KorAP-Architektur um das Konzept entfernter, selbstverwalteter Knoten erweitert. Hierbei sind Korpuseigner für die technische Bereitstellung von Daten selbst verantwortlich. Auf diese Weise behalten sie die uneingeschränkte Kontrolle über den Zugriff auf ihre Daten, während gleichzeitig der zentrale Abruf über eine Web-Schnittstelle erhalten bleibt. Die Lokalität der Daten für die Suche und die Erstellung virtueller Korpora ist dabei ohne Bedeutung. Für die Aggregation der Suchresultate müssen bereitgestellte Daten lediglich zuvor mit ihren Metadaten an der zentralen Schnittstelle registriert werden. Dieses Vorgehen ist effizient, zuverlässig und rechtskonform durchführbar.

Bibliographie

Bański, Piotr / Bingel, Joachim / Diewald, Nils / Frick, Elena / Hanl, Michael / Kupietz, Marc / Pezik, Piotr / Schnober, Carsten / Witt, Andreas (2013): "KorAP: the new corpus analysis platform at IDS Mannheim." Präsentiert auf der *6th Conference on Language and Technology (LTC-2013), Poznan, Polen, Dezember 2013*.

Bartz, Thomas / Beißwenger, Michael / Storrer, Angelika (2014): "Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge", in: *Zeitschrift für germanistische Linguistik* 28, 1: 157-198.

Bodmer, Franck (1996): "Aspekte der Abfragekompononente von COSMAS-II", in: *LDV-INFO* 8. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung 112-122.

Fischer, Peter M. (i.E.): *Eine Datenbasis zur Beobachtung des Schreibgebrauchs im Deutschen*

Gray, Jim (2003): *Distributed Computing Economics*. Technical Report MSR-TR-2003-24. San Francisco: Microsoft Research.

Horbach, Andrea / Thater, Stefan / Steffen, Diana / Fischer, Peter M. / Witt, Andreas / Pinkal, Manfred (2015): "Internet Corpora: A Challenge for Linguistic Processing", in: *Datenbank-Spektrum*15, 1: 41-47 http://link.springer.com/article/10.1007/s13222-014-0172-z [letzter Zugriff 26. Februar 2016].

Krome, Sabine (2010): "Die deutsche Gegenwartssprache im Fokus korpusbasierter Lexikographie. Korpora als Grundlage moderner allgemeinsprachlicher Wörterbücher am Beispiel des WAHRIG Textkorpus Digital", in: Kratochvílová, Iva / Wolf, Norbert Richard (eds.): *Kompendium Korpuslinguistik*. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive. Heidelberg: Universitätsverlag Winter 117-134.

Krome, Sabine (2013): "Digitale Datenflut: Chancen und Tücken eines Textkorpus zur deutschen Gegenwartssprache. Anforderungsprofil, Methoden und Instrumentarien zur Beobachtung des aktuellen Sprach- und Schreibgebrauchs", in: Kratochvílová, Iva / Wolf, Norbert Richard (eds.): *Grundlagen einer sprachwissenschaftlichen Quellenkunde*. Tübingen: Narr Verlag 49-66.

Kupietz, Marc / Lüngen, Harald (2014): "Recent Developments in DeReKo", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* 2378-2385.

Kupietz, Marc / Lüngen, Harald / Bański, Piotr / Belica, Cyril (2014): "Maximizing the Potential of Very Large Corpora", in: *Proceedings of the LREC-2014-*

Workshop Challenges in the Management of Large Corpora (CMLC2) 1-6.

Münzberg, Franziska (2011): "Korpusrecherche in der Dudenredaktion. Ein Werkstattbericht", in: Konopka, Marek / Kubczak, Jacqueline / Mair, Christian / ticha, František / Waßner, Ulrich H.(eds.): *Grammatik und Korpora 2009*. Tübingen: Narr Francke Attempto 181–197.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textkorpora mit STTS*. Technical report. Tübingen / Stuttgart: Universität Stuttgart / Universität Tübingen http://www.ims.unistuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf [letzter Zugriff 26. Februar 2016].

Scholze-Stubenrecht, Werner (2013): "The World Wide Web as a resource for lexicography", in: Gouws, Rufus H. / Heid, Ulrich / Schweickard, Wolfgang /Wiegand, Herbert Ernst (Hrsg.): *Dictionaries*. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography (= HSK 5.4) 1365-1374. Berlin / New York: Mouton de Gruyter.

TEI Consortium (2007): *Guidelines for Electronic Text Encoding and Interchange (TEI P5)*. The TEI Consortium http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 26. Februar 2016].

van Uytvanck, Dieter (2010): *CLARIN Short Guide on Virtual Collections*. Technical report. CLARIN [letzter Zugriff 26. Februar 2016].

Ontologie-basierte Modellierung, Vernetzung und Visualisierung geschichtswissenschaftlichen, wirtschaftswissenschaftlichen und politikwissen-schaftlichen Wissens zur Unterstützung multiperspektivischer Konfliktforschung

Frank, Ingo
frank@ios-regensburg.de
Institut für Ost- und Südosteuropaforschung, Regensburg

Im Poster wird die Relevanz philosophisch informierter Ontologie-Modellierung für den Aufbau einer Informationsinfrastruktur und die Entwicklung von Analyse- und Visualisierungswerkzeugen am Beispiel des neuen Forschungsschwerpunkts „Frozen and Unfrozen Conflicts“ (Konfliktforschung) am Institut für Ost- und Südosteuropaforschung (IOS) in Regensburg gezeigt.

Zum Aufbau der Informationsinfrastruktur kommt Semantic Web-Technologie zum Einsatz. Die Modellierung bzw. Repräsentation des Wissens aus den Perspektiven der beteiligten Disziplinen Geschichtswissenschaft, Wirtschaftswissenschaft und Politikwissenschaft erfolgt durch geeignete Top-Level- und Fach-Ontologien. Die Vernetzung des Wissens erfolgt durch semantische Informationsintegration (Linked Data).

Das Ziel ist die Unterstützung der multiperspektivischen Erklärung von Territorialkonflikten durch die Integration des Wissens aus Sicht der verschiedenen Disziplinen auf den selben Untersuchungsgegenstand und die Nachnutzung dieses Wissens durch Werkzeuge zur visuellen Analyse. Entscheidend dabei ist gemäß dem erkenntnistheoretischen Perspektivismus (Nietzsche 2009), daß dadurch neue Erkenntnisse erlangt werden können.

Semantic Web-Ontologien ermöglichen zwar die Modellierung von Ereignissen, der beteiligten Akteure und des räumlichen und zeitlichen Kontexts (Simple Event Model (SEM) von van Hage et al. 2011) oder auch mereologischen und kausalen Relationen zwischen Ereignissen (Event-Model-F von Scherp et al. 2009), aber es können damit nur eingeschränkt die verschiedenen Perspektiven auf die Ereignisse und die Rollen der daran beteiligten Akteure repräsentiert werden. Goerz und Scholz (2009) behaupten, daß CRM ein Rahmenwerk für transdisziplinäre Forschung bereitstellen kann. Allerdings beschränkt sich ihr beschriebener Ansatz bisher nur auf die Integration von archäologischen und biologischen Wissensorganisationssystemen.

Das besondere Problem, die geisteswissenschaftliche Wirklichkeit (z. B. das Handeln verschiedener Akteure, die Konstitution von Institutionen und Staaten, Normen und Werte, verschiedene Kausalitäten) zu repräsentieren und den Zusammenhang und die gegenseitige Abhängigkeit des Wissens aus verschiedenen geisteswissenschaftlichen Disziplinen zu erfassen, wird bisher nur ansatzweise gelöst (Robinson 2011; Heller / Herre 2004; Semenova 2008; Krieger / Declerck 2014). Mein Ansatz greift daher zum Aufbau einer Ontologie zur Repräsentation des Bereichs der Geisteswissenschaften insbesondere auf Ideen aus der phänomenologischen Ontologie Husserls zurück. Seine Unterscheidung zwischen formaler und materialer Ontologie und seine Einführung der Grundkategorie ‚Sinn‘ (Husserl 2009) hilft dabei, die Top-Level-Ontologie und Fach-Ontologien zur Repräsentation des Wissens aus Sicht verschiedener Disziplinen aufzubauen. Die Theorie der Fundierung und Abhängigkeit (Husserl 2013) ermöglicht

die Repräsentation des Zusammenhangs der Entitäten aus verschiedenen Regionen und unterstützt dadurch die Multiperspektivität (z. B. bei der Modellierung der Konstitution von gesellschaftlichen Institutionen und deren gegenseitige Abhängigkeiten).

Durch die Anwendung von Ideen aus der philosophischen Ontologie wird ein explanatorisches Rahmenwerk für multiperspektivische Erklärung aufgebaut, in dem ontologische Fragen geklärt werden – z. B. nach dem ontologischen Status von Grenzen (Smith 1997) und deren Repräsentation in Semantic Web-Ontologien (Robinson 2009) oder die Frage, welche konstitutiven Elemente einen Staat ausmachen, d. h. ab wann ein Pseudo-Staat ein Staat ist oder was ein Staat überhaupt ist: eine Organisation oder ein völkerrechtliches Subjekt (Robinson 2010)?

Die Relevanz des Ansatzes wird mit einem Werkzeug zur visuellen Analyse demonstriert: Mit der ‚synchronoptischen Konfliktgeschichte' (in Anlehnung an die „Synchronoptische Weltgeschichte" von Peters) können historische, wirtschaftliche, politische und soziale Einflußfaktoren für die Entstehung bzw. das Einfrieren und Auftauen von *frozen conflicts* in Beziehung gesetzt werden. Ereignisse können am besten im Kontext anderer Ereignisse verstanden werden (Allen 2005). Deshalb werden mehrere parallel verlaufende Zeitleisten verwendet, um die Konflikte und die beinflussenden Faktoren aus verschiedenen Perspektiven darzustellen und in Beziehung zu setzen. Die Forscher werden so bei der explorativen Suche und Analyse der ‚unknown unknowns' der „Wirkungszusammenhänge" (Dilthey 1992) zwischen den an der Entwicklung der Konflikte beteiligten Akteuren (z. B. die Rolle der EU-Außenpolitik oder der Politik Russlands) unterstützt.

Territorialkonflikte wie der gegenwärtige Ukraine-Konflikt haben die Eigenart, daß sie einfrieren, aber jederzeit wieder auftauen können. Die vergleichende Analyse der Fluktuationsdynamik des Einfrierens und Auftauens in Abhängigkeit der Einflußfaktoren über längere Zeiträume ist daher ein naheliegender Anwendungsfall für das Werkzeug. Ein anderer Anwendungsfall ist der synchrone oder auch diachrone Vergleich. So lassen sich über einen historischen Vergleich womöglich Territorialkonflikte zur Zeit des zaristischen Russland finden, die ähnliche Konstellationen aufweisen, wie die aktuellen Konflikte.

Das Werkzeug kann über SPARQL auch Information aus der DBpedia oder anderen Informationsbeständen einbeziehen. Ein Vorteil dabei ist, daß zunächst der zeitliche und räumliche Kontext genügt, um Konfliktereignisse und die beteiligten Akteure in Beziehung zu bringen. Über Vokabulare wie das Data Cube vocabulary oder SDMX können auch Forschungsdaten wie Statistiken in die visuelle Analyse einbezogen werden (Atemezing / Troncy 2014), um auch sozio-ökonomische Faktoren zu berücksichtigen.

Die Bedeutung der Digital Humanities als fächerübergreifendes Forschungsparadigma, das geisteswissenschaftliches *Verstehen* im Sinne Diltheys unterstützt, wird an diesem Beispiel für multiperspektivische Konfliktforschung deutlich: Die Digital Humanities können zur multiperspektivischen Erklärung von Konflikten beitragen, indem sie den beteiligten Disziplinen eine gemeinsame virtuelle Forschungsumgebung und digitale Werkzeuge zur Analyse und Visualisierung ontologisch aufbereiteter digitaler Ressourcen bereitstellen.

Bibliographie

Allen, Robert B. (2005): "A Focus-context Browser for Multiple Timelines", in: *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)*, New York 260–261.

Atemezing, Ghislain / Troncy, Raphaël (2014): "Towards a Linked-Data based Visualization Wizard", in: *5th International Workshop on Consuming Linked Data (COLD'14)*.

Dilthey, Wilhelm (1992): *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften* (= Gesammelte Schriften 7). Göttingen: Vandenhoeck & Ruprecht.

Goerz, Günther / Scholz, Martin (2009): "Content Analysis of Museum Documentation with a Transdisciplinary Perspective", in: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, LaTeCH-SHELT&R '09: 1–9.

Hage, Willem R. van / Malaisé, Véronique / Segers, Roxane / Hollink, Laura / Schreiber, Guus (2011): "Design and use of the Simple Event Model (SEM)", in: *Web Semantics: Science, Services and Agents on the World Wide Web* 9, 2: 128–136.

Heller, Barbara / Herre, Heinrich (2004): "Ontological Categories in GOL", in: *Axiomathes* 14, 1: 57–76.

Husserl, Edmund (2009): *Ideen zu einer reinen Phänomenologie und phänomenologischen Philososophie* (= Philosophische Bibliothek 602). Hamburg: Meiner.

Husserl, Edmund (2013): *Logische Untersuchungen*. (= Philosophische Bibliothek 601). Hamburg: Meiner.

Krieger, Hans-Ulrich / Declerck, Thierry (2004): "TMO—The Federated Ontology of the TrendMinder Project", in: *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*.

Nietzsche, Friedrich (2009): "Genealogie der Moral", in: Paolo D'Iorio (ed.): *Digitale Kritische Gesamtausgabe Werke und Briefe*. Nietzsche Source.

Robinson, Edward H. (2010): "An Ontological Analysis of States: Organizations vs. Legal Persons", in: *Applied Ontology* 5, 2: 109–125.

Robinson, Edward H. (2011): "A Theory of Social Agentivity and Its Integration into the Descriptive Ontology for Linguistic and Cognitive Engineering", in:

International Journal on Semantic Web & Information Systems 7, 4: 62–86.

Robinson, Edward H. (2012): "Reexamining Fiat, Bona Fide and Force Dynamic Boundaries for Geopolitical Entities and Their Placement in OLCE", in: *Applied Ontology* 7, 1: 93–108.

Scherp, Ansgar / Franz, Thomas / Saathoff, Carsten / Staab, Steffen (2009): "F–a Model of Events Based on the Foundational Ontology Dolce+DnS Ultralight", in: *Proceedings of the Fifth International Conference on Knowledge Capture*, K-CAP '09, New York 137–144.

Semenova, Elena (2008): "Ontologie als Begriffssystem: Theoretische Überlegungen und ihre praktische Umsetzung bei der Entwicklung einer Ontologie der Wissenschaftsdisziplinen", in: *Konferenzen der Deutschen ISKO 2008 – Wissensspeicher in digitalen Räumen: Nachhaltigkeit, Verfügbarkeit, semantische Interoperabilität*.

Smith, Barry (1997): "The Cognitive Geometry of War", in: Koller, Peter / Puhl, Kuhl (eds.): *Current Issues in Political Philosophy: Justice in Society and World Order*. Wien: Hölder-Pichler-Tempsky.

Biblissima - Semantic Web Application für Handschriften, Inkunabeln und historische Sammlungen - Zwischenbericht

Gehrke, Stefanie
stefanie.gehrke9@gmail.com
Campus Condorcet, Frankreich

Charbonnier, Pauline
pauline.charbonnier@biblissima-condorcet.fr
Campus Condorcet, Frankreich

Eduard, Frunzeanu
eduard.frunzeanu@biblissima-condorcet.fr
Campus Condorcet, Frankreich

Portale zu mittelalterlichen Handschriften sind bereits oft auf nationaler Ebene zu finden. So bieten beispielsweise " Manuscripta Mediaevalia " für Deutschland, " http://manuscripta.at/ " bezogen auf Österreich und " e-codices - Virtual Manuscript Library of Switzerland" in der Schweiz einen Einstieg in die jeweiligen Bestände.

Das französische DH-Projekt *Biblissima* (Bibliotheca Bibliothecarum Novissima) zielt unter anderem darauf, Ende 2016 der breiten Öffentlichkeit einen übergreifenden Einstieg in über 40 Datenbanken zu Handschriften des Mittelalters und der Renaissance, Inkunabeln und historischen Sammlungen zu bieten. *Biblissima* wurde von neun Partnerinstitutionen ins Leben gerufen, darunter der *Campus Condorcet* im Norden von Paris, das *IRHT* (Institut de recherche et d'histoire des textes), die *BnF* (Bibliothèque nationale de France), die *EPHE* (Ecole pratique des hautes études) und das *CESR* (Centre d'Etudes Supérieures de la Renaissance). Projektstart war der Beginn des Jahres 2013.

Finanziert wird unser Projekt von der Forschungsförderung *ANR* (Agence Nationale de la Recherche) und es befasst mit der Geschichte historischer Sammlungen sowie der Geschichte der Überlieferung von Texten. In gewisser Weise sieht es sich sogar in der Tradition von *Antoon Sanders* und *Bernard de Montfaucon* , die mit der *Bibliotheca Belgica Manuscripta* (1641-44) beziehungsweise der *Bibliotheca bibliothecarum manuscriptorum nova* (1739) bedeutende bibliotheksübergreifende Inventare veröffentlichten.

Unsere technische Lösung hin zu einer Realisierung einer "Online-Bibliothek des 21. Jahrhunderts zu den heutigen und historischen Sammlungen Frankreichs" ist die Entwicklung einer Semantic Web Application. So haben wir bereits im Sommer 2015 einen Prototypen (http://demos.biblissima-condorcet.fr/prototype) veröffentlicht, der einen vereinigten Zugang zu zwei bedeutenden ikonographischen Datenbanken ermöglicht: *Initiale* und *Mandragore* . Ein Export der Metadaten zu all jenen Illuminationen, die Geografika als Schlagwort tragen, war unser Ausgangspunkt. Personen (Autoren, Übersetzer und / oder Illuminatoren), Körperschaften (zeitgenössische Bibliotheken), Werktitel und Ortsnamen haben wir mit Normdateien der *BnF* verknüpft sowie, wenn möglich, mit externen Vokabularen verlinkt (*GeoNames*, *VIAF*). Der Prototyp bot zum Zeitpunkt der Veröffentlichung Zugang zu ca. 20.000 Illuminationen in 2000 Handschriften, die heute in 70 verschiedenen Bibliotheken aufbewahrt werden. Über 5000 Geografika wurden mit Hilfe von Normdaten der *BnF* und/oder *GeoNames* eindeutig identifiziert. 115 Werktitel und ein Großteil der 359 Personennamen konnten mit Normdaten der *BnF* verbunden werden.

Dank dieser Verknüpfungen werden nun dynamische Webseiten generiert, die jeweils eine Person, eine Handschrift, einen Handschriftenteil etc. beschreiben und Links zu den entsprechenden Einträgen in den ursprünglichen Datenbanken bereitstellen. Auch können wir nun – für unseren Datenbestand - beispielsweise über zwei verschiedene Karten bezogen auf die Geografika darstellen, in welchen Handschriften und durch welche Illuminationen diese dargestellt sind oder aber welche (Anzahl von) Illuminationen am jeweiligen Ort aufbewahrt werden.

Derzeit arbeiten wir an einer breitflächigen Erweiterung dieses Prototypen, der auf dem Semantic Web Application Framework *CubicWeb* basiert. Neben Handschriften binden wir Inkunabeln ein und legen einen Fokus auf die Provenienz, indem wir Daten aus den vier Datenbanken *Esprit des Livres* (Ecole nationale des Chartes, Datenbank zu Verkaufskatalogen), *Bibale* (IRHT, Datenbank zu historischen Büchersammlungen), *Europeana Regia* (BnF in Kooperation mit u. a. der Bayerischen Staatsbibliothek und der Herzog August Bibliothek, Datenbank zu drei bedeutenden historischen Sammlungen des Mittelalters und der Renaissance) und *CRII* (CNRS, Catalogues Régionaux des Incunables Informatisés) beziehungsweise *CRIICO* (CNRS, Catalogues Régionaux des Incunables Informatisés Centre-Ouest) integrieren.

Hinter dieser Applikation liegt ein Datenmodell, dass auf CIDOC-CRM und FRBRoo basiert. Erste Bereitstellungen unserer Daten in RDF liegen bereits als Test vor. Wir hoffen, in Zukunft nicht nur Werks-, Manifestations- und Exemplarsebene (unter Verwendung der Klassen *F4 Manifestation Singleton* für Handschriften sowie *F3 Manifestation Product Type* und *F5 Item* für Inkunabeln) zu behandeln, sondern die Texte der Handschriften und frühen Drucke auch einer Textvariante (Klasse *F2 Expression*) zuordnen zu können, wie beispielsweise von *Perseus* durch die Bereitstellung von URIs pro Text und Sprache begonnen.

Ein eingebundener Viewer zeigt zudem das Digitalisat des Exemplars beziehungsweise des entsprechenden Folios. Als Viewer wird derzeit Mirador favorisiert. Die Informationen zur Darstellung beruhen auf dem IIIF (International Image Interoperability Framework). Diese werden in JSON-LD Manifesten, erzeugt aus den Metadaten- und Imagesfiles der Partnerinstitution, an den Viewer übergeben. Der Viewer selbst arbeitet im Client und beherrscht unter anderem Zoomfunktionen der eingebundenen Bilder, Anzeige von Metadaten oder auch Überlagerung verschiedener Ebenen, wie zum Beispiel Bild und transkribierter Text.
Im Rahmen unseres Posters möchten wir die Vorgehensweise und technische Umsetzung ausgehend vom neuesten Stand des Portals Biblissima veranschaulichen.

Digitales Arbeiten in den Geisteswissenschaften stärken – wissenschaftliche Begleitforschung in DARIAH-DE

Gnadt, Timo
gnadt@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Stiller, Juliane
jstiller@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Deutschland

Thoden, Klaus
kthoden@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Deutschland

DARIAH bietet eine soziale und technische Forschungsinfrastruktur für digital arbeitende Geistes- und KulturwissenschaftlerInnen. Der deutsche Partner DARIAH-DE befasst sich neben konzeptionellen technischen und fachlichen Entwicklungen unter anderem mit wissenschaftlicher Begleitforschung. Dies bedeutet im konkreten Fall der 2016 endenden zweiten Förderphase von DARIAH-DE die Erforschung von Nutzerverhalten und -erwartungen, Aspekten der Usability sowie Impactfaktoren und Erfolgskriterien bei digitalen Werkzeugen und Forschungsinfrastrukturen in den Geisteswissenschaften.

Als Voraussetzung für die strukturierte Erforschung des Nutzerverhaltens wurde in DARIAH-DE ein Modell des geistes- und kulturwissenschaftlichen Forschungsprozess erstellt, dessen Phasen bestimmte Arbeitsschritte umfassen. Die wissenschaftliche Begleitforschung konzentrierte sich im Folgenden darauf, die zur Abdeckung dieser Arbeitsschritte erforderlichen digitalen Tools bzw. die in dieser Hinsicht noch bestehenden Lücken zu identifizieren. Dies wurde u. a. durch eine breit angelegte Umfrage unter FachwissenschaftlerInnen realisiert. Die Auswertungen und Ergebnisse dieser Umfrage dienen dazu, die Erwartungen der AnwenderInnen an die Forschungsinfrastruktur aufzufangen und noch besser umzusetzen. So galt es zu verstehen, warum bestimmte digitale Tools häufig eingesetzt werden und andere weitestgehend ungenutzt bleiben. Daru#ber hinaus sollten Lu#cken in der Abdeckung des geisteswissenschaftlichen Forschungsprozesses durch digitale Dienstleistungen und Tools aufgedeckt und Erwartungen der NutzerInnen an digitale Tools strukturierter erfasst werden. Das übergeordnete Ziel dieser Arbeiten ist eine bessere Integration von Software in den geisteswissenschaftlichen Forschungsprozess.

Unerlässlich für die Akzeptanz von Tools ist deren Bedienbarkeit und Nützlichkeit beim Ausführen bestimmter wissenschaftlicher Tätigkeiten. Hierzu zählt unter anderem auch der durch eine gewisse visuelle und funktionale Vereinheitlichung erreichbare Wiedererkennungswert, welcher sowohl einen Grad von

Vertrautheit wie auch eine verbesserte Erlernbarkeit neuer Tools und Funktionen schafft. Um einen Mindeststandard für die Anforderungen an geisteswissenschaftliche Tools in der DARIAH-Infrastruktur zu setzen, wurde ein Style Guide zur Steigerung der Usability der angebotenen Dienste und der Umsetzung gewisser Qualitätsstandards erarbeitet (cf. Romanello et al. 2015). Dieser Styleguide umfasst verschiedene Aspekte, die von lizenzrechtlichen Bedingungen, über ausreichende Dokumentation des Tools bis hin zu erforderlichen Funktionen (z. B. Export) reichen. Das Ziel dieses Teils der wissenschaftlichen Begleitforschung in DARIAH-DE ist es, Tools, die oft unabhängig voneinander entwickelt wurden, in eine gemeinsame Forschungsinfrastruktur überführen zu können und diese Überführung auch den NutzerInnen sichtbar zu machen. Zu diesem Zweck war es auch erforderlich, die Vorteile einer solchen Integration gegenüber den erforderlichen Aufwänden herauszuarbeiten und sowohl den NutzerInnen wie auch den Diensteanbietern anschaulich zu machen.

Ein zentrales Ergebnis der wissenschaftlichen Begleitforschung bildet der erarbeitete Katalog von Erfolgskriterien, der viele Aspekte der Nutzeranforderungen an Tools und auch deren Qualitätsmerkmale wieder aufgreift. Der Katalog basiert sowohl auf von DARIAH-DE durchgeführten Umfragen unter verschiedenen Stakeholdergruppen als auch auf vorangegangen Analysen und Modellen, wie z. B. dem 2014 abgeschlossenen DFG-Projekt zu Erfolgskriterien virtueller Forschungsumgebungen (Buddenbohm et al. 2014). Konkret wurden hierbei die zusammengetragenen möglichen Kriterien wie von Buddenbohm et al. (2014) vorgeschlagen zur Erstellung eines disziplinspezifischen Katalogs verwendet, indem die für die jeweiligen Stakeholdergruppen in den Geisteswissenschaften relevanten Eigenschaften abgefragt bzw. bewertet wurden.

Hierbei wurden folgende Stakeholdergruppen unterschieden:

DH-Anwender/Nutzer, vorwiegend “digital affine”, d. h. mit digitalen Tools bzw. Methoden vertraute GeisteswissenschaftlerInnen
Dienste-Entwickler, insbesondere Software-EntwicklerInnen und InformatikerInnen innerhalb von Infrastrukturen
Diensteanbieter, also Infrastrukturdienstleister wie Rechenzentren, Bibliotheken etc.
Fördererinstitutionen

Die über Umfragen erhaltenen Ergebnisse wurden nach diesen Gruppen aufgeschlüsselt und analysiert, um verschiedene Schwerpunkte und Ausprägungen von Kriterien herauszuarbeiten. Somit wird einerseits den jeweiligen Stakeholdern ein Überblick über die innerhalb der eigenen Gruppe bestehenden Anforderungen gegeben, andererseits aber auch deren Perspektive für die anderen Gruppen geöffnet. Hierdurch soll ein nachhaltiger Austausch angestoßen werden, um letztendlich eine höhere Effektivität von Entwicklungs-, Angebots- und Förderungsprozessen zu bewirken.

Der geplante Posterbeitrag stellt die Ergebnisse in den genannten drei Bereichen Nutzeranforderungen, Usability und Erfolgskriterien dar und zeigt hierbei mögliche Konsequenzen sowohl für Dienste-Entwickler und Anbieter, als auch für Fördererinstitutionen auf.

Bibliographie

Buddenbohm, Stefan / Enke, Harry / Hofmann, Matthias / Klar, Jochen / Neuroth, Heike / Schwiegelshohn, Uwe (2014): *Erfolgskriterien für den Aufbau und nachhaltigen Betrieb Virtueller Forschungsumgebungen*. DARIAH-DE Working Papers Nr. 7. Göttingen: DARIAH-DE.

Romanello, Matteo / Stiller, Juliane / Thoden, Klaus (2015): *Usability Criteria for External Requests of Collaboration (R 1.2.2/R 7.5)*. DARIAH-DE Aufbau von Forschungsinfrastrukturen für die eHumanities. DARIAH-DE https://wiki.de.dariah.eu/download/attachments/14651583/R1.2.2_Usability_Criteria_for_External_Requests_of_Collaboration.pdf [letzter Zugriff 15. Oktober 2015].

Erschließung digitaler Ressourcen und Forschungsdaten in den Digital Humanities: Der Digitale Wissensspeicher an der Berlin-Brandenburgischen Akademie der Wissenschaften

Grabsch, Sascha
grabsch@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

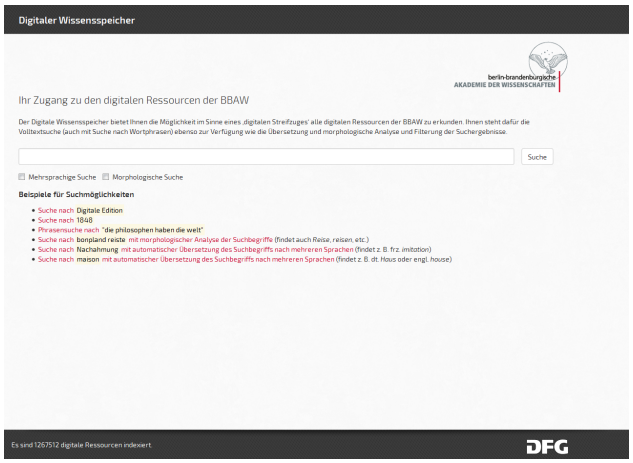
Jürgens, Marco
juergens@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

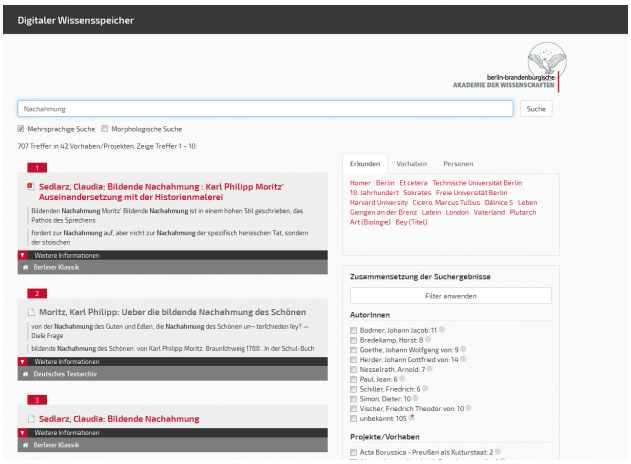
Im Rahmen des DFG-geförderten Projektes „Digitaler Wissensspeicher“ wurde an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) von 2012 bis April 2015 ein zentraler Zugang für sämtliche digitalen Forschungsdaten und Ressourcen der Akademie geschaffen. Damit sind über 170 Projekte mit insgesamt mehr als 1 Mio. digitaler Ressourcen im Volltext und mit Metadaten erfasst. Diese in Inhalt, Format und Sprache äußerst vielfältigen und heterogenen Ressourcen werden erstmals vollständig über ein einheitliches Interface zugänglich gemacht.[1] Den Kern der Ressourcen stellen digitale Editionen und Übersetzungen (in XML, HTML, PDF), elektronische Kataloge, Dokumentationen und Datenbanken, digitale Volltextsammlungen sowie Wörterbücher dar. Durch den Einsatz von Sprachtechnologien (Bing, DONATUS) kann der Digitale Wissensspeicher mehrsprachig (deutsch/englisch/französisch) und morphologisch normalisiert durchsucht werden. Neben manuell erfassten, fließen auch automatisiert erstellte Metadaten in den Datenbestand des Wissensspeichers ein und können von den Nutzerinnen und Nutzern abgefragt werden. Die Metadaten für den gesamten Bestand der digitalen Ressourcen werden dabei auch über eine maschinenlesbare Schnittstelle (OAI-PMH) zur Verfügung gestellt. Damit werden die digitalen Forschungsdaten der BBAW Teil der Linked Open Data Cloud. Neben der Einbeziehung von Semantic-Web-Ressourcen zur Anreicherung der Suchergebnisse stellt der Wissensspeicher so auch Ressourcen und Forschungsdaten der BBAW zur Verfügung.

Die größte Herausforderung beim Aufbau einer disziplinübergreifenden Forschungsdateninfrastruktur wie dem Wissensspeicher stellte die große Heterogenität der an der BBAW in den letzten 20 Jahren entstandenen digitalen Ressourcen dar. Sowohl inhaltlich als auch technisch bilden diese ein unübersichtliches Feld, das für den Digitalen Wissensspeicher strukturiert und zugänglich gemacht werden musste. Gestützt auf einen Volltextindex (Apache Lucene) und ein an die Bedingungen der Akademie angepasstes Metadatenschema (basierend auf OAI-ORE) wurden für die einzelnen Projekte und Ressourcensammlungen Importmodule entwickelt, welche die sehr unterschiedlichen Datenstrukturen der jeweiligen Projekte in den (Meta-)Datenbestand des Wissensspeichers integrieren. Über die Einbindung von Semantic-Web-Technologien (u. a. von DBpedia) und Text-Mining-Tools werden für die Nutzerinnen und Nutzer semantisch an die Suchanfrage gebunden Vorschläge für die Erweiterung und Erkundung des Datenbestandes angeboten.

Die zweite Projektphase des Wissensspeichers wird bis Ende 2017 die Erweiterung der Nutzungsmöglichkeiten und besonders die Nachhaltigkeit des Projektes zum Schwerpunkt haben. Es besteht ein erheblicher Bedarf von wissenschaftlichen Institutionen an einer nachhaltigen Forschungsdateninfrastruktur, welche die spezifische Situation der Geisteswissenschaften – heterogene Ressourcen – adäquat abbilden kann. Ein wichtiges Ziel der nächsten Projektphase ist es daher, den Wissensspeicher für weitere Nutzergruppen zu öffnen. Dafür werden die Softwarekomponenten des Wissensspeichers als installierbares Paket zur Verfügung gestellt. So werden externe Partnerinstitutionen in die Lage versetzt, einen eigenen Wissensspeicher mit eigenen digitalen Ressourcen zu betreiben. Für die Abstimmung mit zukünftigen externen Nutzern und die koordinierte Weiterentwicklung als Open Source Software wird an der BBAW im April 2016 ein offener Workshop stattfinden.

Ein weiterer Hauptpunkt der Arbeit in dieser Projektphase bildet die Erstellung von Guidelines mit strukturellen und inhaltlichen Mindestanforderungen, die Ressourcen und Metadaten erfüllen müssen, um effektiv und zweckmäßig in den Wissensspeicher aufgenommen werden zu können. Damit werden Zielvorgaben für die (technische) Qualität der in den Wissensspeicher aufzunehmenden Ressourcen mit ihren Metadaten formuliert. Diese Best-Practice-Empfehlungen können auch über den konkreten Anwendungsfall des Digitalen Wissensspeichers hinaus einen Empfehlungscharakter für den Aufbau und Betrieb von Ressourcensammlungen in den Digital Humanities bekommen und sollen von den Partnerinstitutionen sukzessive optimiert und an die eigenen Gegebenheiten angepasst werden. Ebenso werden Workflows für die manuelle Erhebung und Einspeisung von Metadaten entwickelt. Weitere Entwicklungsziele stellen der Ausbau von Visualisierungskomponenten sowie die automatisierte Auswertung und Einbindung von Nutzerfeedback in den Suchprozess dar.





Notes

1. Der derzeitige Entwicklungsstand des Digitalen Wissensspeichers kann unter http://wspdev.bbaw.de eingesehen und getestet werden. Je nach Gegebenheiten vor Ort soll dies über einen Laptop auch bei der Posterpräsentation ermöglicht werden.

Dissertation: Der Berner Chorherr Heinrich Wölfli (1470-1532) und die Beschreibung seiner Heiligland-Wallfahrt von 1520/21 - Erschliessung und Darstellung durch klassisch-literaturwissenschaft-liche und digital-moderne Methoden

Habicht, Stephanie
stephanie.odok@uzh.ch
Universität Zürich, Schweiz

Gegenstand meiner Dissertation ist die Reise nach Jerusalem von 1520/21 von Heinrich Wölfli, erhalten in der deutschen Übersetzung von Johannes Haller. Einerseits soll - ganz klassisch im literaturwissenschaftlichen Sinne - eine narratologische Analyse der Beschreibung sowie ein Handschriftenvergleich, inkl. Stemma und Jahresschätzung der 6 bekannten Handschriften, anhand von inhaltlichen Abweichungen und Wasserzeichen vorgenommen werden. Zudem wird ein kodikologischer Befund, die Biographien des Autors sowie des Übersetzers, ein Stellenkommentar und ein kleines Lexikon der für diesen Text speziellen und ansonsten eher unbekannten Wörter erstellt.

Andererseits wird mithilfe von *TUSTEP* ebenfalls ein Stemma erstellt, welches mit den Ergebnissen der klassischen Vorgehensweise verglichen wird, und es werden abweichende Textpassagen der Handschriften dadurch genauer analyisert. Der kritische Apparat dazu soll ebenfalls mittels *TUSTEP* erstellt werden, mithilfe des Satz-Moduls.

Zusätzlich - und das ist der für das Poster relevanteste Punkt - ist angedacht, die teilweise sehr detailliert beschriebenen und auf das Datum und die kleinste Ortschaft genau festgehaltenen Reisestationen Wölflis in XML zu erfassen, und die Reise im *DARIAH-Geobrowser* zu visualieren. Dafür müssen zuerst die Ortschaftsnamen verifiziert und auf unsere heutige Namensgebung hin gemappt werden, sowie mit Koordinaten versehen werden. Angedacht ist eine interaktive Karte, in der schließlich nicht nur die Route dargestellt, sondern auch die geschilderten Begebenheiten der einzelnen Reisestationen sowie die zeitliche Komponente der Reise sichtbar gemacht werden sollen. Dafür soll *TUSCRIPT* genutzt werden. Ziel ist es, diese lebendig und humoristisch geschriebene Reisebeschreibung als eine ebenso lebendige und greifbare Erfahrung auf digitalem Weg zugänglich zu machen und dem Leser näher zu bringen; somit wage ich einen Brückenschlag über die zeitliche Distanz von knapp 500 Jahren und die gegebene räumliche Distanz, sowohl die der Reise selbst, als auch (unter Umständen) die von Rezipient und Autor. Ideologische Vorlage dieses Unterprojektes der Dissertation, welches ganz im Zeichen der Digital Humanities steht, ist die epigraphische Datenbank *epidat* des Steinheim Instituts.

Das vorgestellte Teilprojekt dient sicher der Visualisierung geisteswissenschaftlicher Daten und vielleicht auch der Vernetzung in einem ganz speziellen Sinne. Zudem kann es, weiter angereichert, auch für andere Fächer nutzbar gemacht werden, so zum Beispiel für die Fächer Geschichte, Geographie und Religionswissenschaften.

Dieses Teilprojekt der Dissertation ist derzeit in Arbeit. Wieviel davon im März 2016 zur DHd-Tagung bereits abgeschlossen ist, ist derzeit nicht abschätzbar. Sicher ist aber, dass bis dahin ein Teil der Daten schon aufbereitet ist und auf dem Poster vorgestellt werden kann, sowie, dass die generelle technische Umsetzung und die theoretische Überlegung dargestellt und erörtert werden kann. Dasselbe gilt für den Handschriftenvergleich, welcher sowohl mit klassisch-literaturwissenschaftlichen als auch mit digital-modernen Methoden angegangen werden soll, wobei der Fokus

für den Beitrag doch primär auf der Visualisierung der Reiseroute liegt.

Für die Zukunft muss (urheberrechtspolitisch) noch eruiert werden, inwiefern die digitalisierte Version der vermuteten Originalhandschrift, welche mit aufwändigen Bildern versehen ist, im Netz (wenigstens eingeschränkt) zugänglich gemacht werden kann. Falls dem nichts entgegensteht, ist als Projekt nach der Dissertation eine digitale Edition von Wölflis Reise nach Jerusalem angedacht, im Stile des *Parzival-Projektes* der Universitäten Bern, Berlin und Erlangen (Stolz / Haustein / Glauch 2016).

Abb. 1: Beschreibung der Rückreise: Schiff im Sturm vor Griechenland, Dezember 1520

Bibliographie

Heinrich Woelfli's Syrische Reise 1520, aus dem Lateinischen übersetzt von Johann Haller 1582, Burgerbibliothek Bern, Signatur Mss.h.h.XX.168

Steinheim Institut / Kollatz, Thomas (2006-2016): *Epidat*. Epigraphische Datenbank http://www.steinheim-institut.de/cgi-bin/epidat [letzter Zugriff 21. Januar 2016].

Stolz, Michael / Haustein, Jens / Glauch, Sonja (2016): *Parzival-Projekt*. Ein Projekt des Schweizerischen Nationalfonds und der Deutschen Forschungsgemeinschaft http://www.parzival.unibe.ch/home.html [letzter Zugriff 21. Januar 2016].

Mit der FinderApp durch Goethes Faust: Treffer im Faksimile visuell hervorgehoben und multimediale Ausgabe in Videoaufführung und Hörbuch.

Hadersbeck, Maximilian
maximilian@cis.uni-muenchen.de
Ludwig Maximilians Universität München, Deutschland

Eder, Elisabeth
e.eder@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

Capsamun, Roman
r.capsamun@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

Eichfeldt, Nora
nora.eichfeldt@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

Herteis, Simeon
s.herteis@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

Lindinger, Matthias
m.lindinger@live.de
Ludwig Maximilians Universität München, Deutschland

Höps, Raphael
R.Hoeps@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

Schweter, Stefan
stefan@schweter.eu
Ludwig Maximilians Universität München, Deutschland

Einleitung

In unserem Poster möchten wir unsere neueste FinderApp GoetheFind vorstellen, die mit computerlinguistischen Methoden die Originalausgabe von Goethes Faust durchsucht. GoetheFind hat einen neuen browser-basierten Faksimile-Viewer, der die Schaltstelle der multimedialen Ausgabe der Suchtreffer darstellt: Die Treffer werden im Faksimile der Originalausgaben gehighlighted dargestellt und mit einer gesprochenen Faust-Hörbuchausgabe verlinkt. Bei Faust I werden die Treffer mit der entsprechenden Szene des Videos der Bühnenaufführung vom Hamburger Schauspielhaus mit Gustav Gründgens (1960) verlinkt. GoetheFind entstand aus unserer FinderApp WiTTFind (Hadersbeck / Pichler; Hadersbeck et al. 2014), die den öffentlich zugänglichen Teil des Nachlasses von Ludwig Wittgenstein durchsucht und mit der wir im Sommer 2014 den EU-AWARD des EU-Projekt Digitised Manuscripts to Europeana (cf. Ploeger 2014) gewannen.

In unserer neuen FinderApp GoetheFind, setzen wir Ideen des „Standoff-Markups“ um, damit „overtagged“-XML vermieden wird. Wir entwickelten eine reduzierte „XML-TEI-P5 anchor-key“ Edition und speichern die Metainformatione in einer „NoSQL-mongo“-Datenbank. Alle relevanten Editions-, OCR- und Transkriptionsinformationen zur multimedialen Trefferausgabe sind in der Datenbank gespeichert.

Modellierung der Editionsdaten

Anstatt „overtagged“ XML ein reduziertes „anchor-key“ XML-TEI-P5 mit „NoSQL“-Database

Grundlage unserer FinderApp GoetheFind ist die XML-TEI-P5 Textedition im DTABf Format vom Deutschen Textarchiv (DTA) der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW 2013; Haaf et al. 2015), dazu die Bilddigitalisate der Staatsbibliothek zu Berlin (BBAW 2013) und dem freiem Hochstift Frankfurt des Frankfurter Goethe-Hauses (Signatur: III B / 23). Da wir zur multimedialen Ausgabe der Suchtreffer zahlreiche Metainformationen benötigen und "overtagged" XML des Editionstexts vermeiden wollten, verwenden wir Ideen des „Standoff-Markups“ und lagern alle notwendigen Meta-Informationen in der „NoSQL“ mongo-Datenbank GoetheDB aus. Eine eindeutige Referenz der Datenbankeinträge zum Editionstext lösen wir über den XML-TEI-P5 Tag <anchor/>, der an Seitenanfängen in die Edition eingefügt ist. Die Trefferpositionen werden über ein XML-Attributtrippel (Seite, Zeile, Token) genau spezifiziert.

Vorverarbeitung der Edition, Faksimile, Videoaufführung und Audioaufnahme

Da unsere FinderApp die Suchanfragen regelbasiert mit Hilfe von lokalen Grammatiken im Kontext eines Satzes realisiert, verwenden wir als wichtigste Strukturierungsebene Sätze. Goethes Faustdrama bettet Sätze in Rede und Gegenrede, sogenannte Repliken ein, die die zweite Strukturierungsebene darstellt. Zur visuellen Hervorhebung und multimedialen Ausgabe der gefundenen Textstellen im Faksimile ermitteln wir für die Replike geometrische Informationen mit Hilfe eines von uns entwickelten semiautomatischem OCR-Correction Tools. Die Bühnen- und Audioaufnahme werden mit Hilfe des Clarin-Tools: „Munich Automatic Segmentation System WebMAUS“ (CLARIN) semiautomatisch transkribiert.

Computerlinguistische Methoden zur Textinterpretation

Mit Hilfe unseres Speziallexikons GoetheLEX, angereichert um historische Sprachvarianten, Part of Speech Tagging und lokalen Grammatiken implementierten wir eine Partikelverb- und Semantische Suche. Bei der Eingabe von Suchanfragen verwenden wir eine symmetrische Autovervollständigung mit Informationen zur Häufigkeit des Auftretens im Text.

Treffer im Faksimile-Viewer und multimediale Ausgabe

Ähnlich wie bei Google-Docs entwickelten wir einen Browser basierter Faksimile-Viewer mit dem man in einem doppelseitigen Buchlesemodus durch das Dokument blättern kann und die gefundenen Textstellen farblich hervorgehoben werden. GoetheFind vernetzt alle Treffer mit der entsprechenden Replik in der Bühnenaufführung und der Hörbuchausgabe. Sobald der Nutzer auf die Mulitmediabuttons des Treffers drückt, startet im Browser ein Videoviewer oder eine Ausdioausgabe ab dieser Stelle.

Danksagung

Wir danken dem Deutschen Textarchiv für die gute Zusammenarbeit und die freundliche Verfügungsstellung der Editionsdaten von Goethes Faust (BBAW 2013). Der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz und dem Freien Deutschen Hochstift, in Frankfurt danken

wir für die Wiedergaberechte der Bilddigitalisate der Originalausgabe Goethes Faust.

Bibliographie

BBAW (2013): *Das DTA-Basisformat DTABf*. Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) http://www.deutschestextarchiv.de/doku/basisformat [letzter Zugriff 09. Januar 2016].

CLARIN (o. J.): *CLARIN-D WebMAUS*. Automatic Segmentation of Speech. https://www.clarin.eu/movies/clarin-d-webmaus-automatic-segmentation-speech [letzter Zugriff 08. September 2015].

Haaf, Susanne / Geyken, Alexander / Wiegand, Frank (2015): "The DTA 'Base Format': A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources", in: *Journal of the Text Encoding Initiative* 8 https://jtei.revues.org/1114 [letzter Zugriff 09. Januar 2016].

Hadersbeck, Max / Bruder, Daniel / Capsamun, Roman / Conforti, Costanza / Eder, Elisabeth / Eichfeldt, Nora / Fink, Florian / Herteis, Simeon / Höps, Raphael / Lindinger, Matthias / Ling, Jennifer / Mittelhammer, Katharina / Schmidt, Katharina / Schweter, Stefan *FinderApp GoetheFind*. Centrum für Informations- und Sprachverarbeitung (CIS), Ludwig Maximilians Universität München http://goethefind.cis.uni-muenchen.de/ [18.02.2016].

Hadersbeck, Max / Pichler, Alois (eds.) (o. J.) *FinderApp WiTTFind*. Centrum für Informations- und Sprachverarbeitung (CIS), Ludwig Maximilians Universität München & Wittgenstein Archives, University of Bergen http://wittfind.cis.uni-muenchen.de/ [letzter Zugriff 09. Januar 2016].

Hadersbeck, Max, Pichler, Alois, Fink, Florian , Gjesdal, Øyvind Liland (2014): "Wittgenstein's Nachlass. WiTTFind and Wittgenstein advanced search tools (WAST)", in: *Digital Access to Textual Cultural Heritage (DaTeCH 2014)*, Madrid 91-96 http://dblp.uni-trier.de/db/conf/datech/datech2014.html#HadersbeckPFG14 [letzter Zugriff 09. Januar 2016].

Hadersbeck, Max / Pichler, Alois / Fink, Florian / Bruder, Daniel / Arends, Ina / Baiter, Johannes (2015): "Wittgensteins Nachlass. Erkenntnisse und Weiterentwicklung der FinderApp WiTTFind", in: *2. Tagung der Digital Humanities im deutschsprachigen Raum 23.-27.2.2015*, Graz.

Ploeger, Lieke (2014): "Open Humanities Awards round 2 – Winners announced", in: *DM2E. Digitised Manuscripts to Europeana* http://dm2e.eu/open-humanities-awards-round-2-winners-announced/ [letzter Zugriff 09. Januar 2016].

TEI (2015): "Stand-off Markup", in: *TEI Guidelines for Electronic Text Encoding and Interchange*, Version 2.9.1, 16.9 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SASO [letzter Zugriff 09. Januar 2016].

TInCAP – ein interdisziplinäres Korpus zu Ambiguitätsphänomenen

Hartmann, Jutta
jutta.hartmann@uni-tuebingen.de
Universität Tübingen, Deutschland/ Universität Stuttgart, Deutschland

Sauter, Corinna
corinna.sauter@uni-tuebingen.de
Universität Tübingen, Deutschland

Schole, Gesa
gesa.schole@uni-tuebingen.de
Universität Tübingen, Deutschland

Wagner, Wiltrud
wiltrud.wagner@uni-tuebingen.de
Universität Tübingen, Deutschland

Gietz, Peter
peter.gietz@daasi.de
DAASI International

Winkler, Susanne
susanne.winkler@uni-tuebingen.de
Universität Tübingen, Deutschland

Ambiguitäten allerorten: Ambiguität ist ein integraler Bestandteil menschlicher Kommunikation. Sie kann unbeabsichtigt produziert werden, wie in (1a), oder zu strategischen Zwecken eingesetzt werden, z. B. für komische Effekte wie in (1b).

1a. William isn't drinking because he's unhappy (vgl. Hirschberger / Avesani 1997).

1b. …men who can sheer sheep and women with long hair… (vgl. Cutting from BBC News Website).

Ambiguität findet sich vornehmlich in sprachlichen Ausdrücken, kann jedoch auch in der Interaktion mit Bildern oder in Bildern selbst zu finden sein, sowie auf nicht-sprachliche Kommunikation übertragen werden. Daher eröffnet das Thema Ambiguität ein interdisziplinäres Forschungsfeld, an dem neben Sprach- und Literaturwissenschaft auch Rhetorik, Psychologie, Theologie, Rechtswissenschaft und Medienwissenschaften größtes Interesse bekunden (siehe beispielsweise Klein / Winkler 2010; Winkler 2015).

Das Datenbankprojekt TInCAP (Tübingen Interdisciplinary Corpus of Ambiguity Phenomena), das im Rahmen des interdisziplinären Graduiertenkollegs

1808 „Ambiguität: Produktion und Rezeption“ entsteht, zielt darauf ab, Belege von Ambiguität verschiedener Provenienz zu sammeln und diese interdisziplinär zu annotieren und nachhaltig zu speichern. Dabei stehen drei Ziele im Vordergrund: (a) die interdisziplinäre Auseinandersetzung mit dem Thema Ambiguität durch die Erstellung eines gemeinsamen Annotationsschemas, (b) die Nachhaltigkeit der gesammelten Daten und (c) die Zugänglichkeit der Datensammlung für die nationale und internationale Forschungsgemeinschaft.

ANNOTATION. Jeder Beleg zur Ambiguität kann interdisziplinär im Hinblick auf fünf verschiedene Aspekte annotiert werden. Die (i) Kommunikationsebene legt fest auf welcher Ebene die Ambiguität in der Kommunikation eine Rolle spielt, beispielsweise auf der Ebene der fiktiven Charaktere vs. Erzähler-Leser vs. konkrete Kommunikation. Wir unterscheiden außerdem (ii) zwischen strategischem vs. nicht-strategischem Einsatz der Ambiguität in Produktion und Rezeption. Darüber hinaus wird (iii) sowohl die Ebene des Auslösers der Ambiguität (z. B. auf Wortebene, Phrasenebene etc.) annotiert als auch ihre Reichweite, d. h. bis zu welcher Ebene sie für die Interpretation relevant ist. Als weiteren Punkt kennzeichnet (iv) eine qualitative Annotation wie sich die unterschiedlichen Lesarten zueinander verhalten: Sind sie voneinander abgeleitet oder komplett unabhängig? Spielt Vagheit eine Rolle? Nicht zuletzt sieht die Datenbank die Möglichkeit vor, (v) disziplininterne Begrifflichkeiten zur Beschreibung des behandelten Phänomens zu verwenden. Dieses interdisziplinär erarbeitete und anwendbare Schema erlaubt es uns langfristig Korrelationen in den Daten zu finden, die über die unterschiedlichen Modi (unterschiedliche Typen von Ambiguität in unterschiedlichen Texttypen / Bildern) hinweg gelten, sodass wir damit ein genuin interdisziplinäres Forschungsergebnis erreichen können.

NACHHALTIGKEIT. Ein wichtiges Ziel der Datenbank ist es, die gesammelten Daten langfristig und nachhaltig zu speichern. Dazu haben wir ein XML-Schema entwickelt, das weitestgehend TEI-konform [5] ist. Diese XML-Dateien können im Rahmen der universitären Infrastruktur langfristig gespeichert, katalogisiert und zugänglich gemacht werden. Bei Video-, Audio- und Bilddateien halten wir uns an die üblichen Standards für nachhaltige Datenformate.

INTERFACE. Für die aktive Arbeitsphase mit der Datenbank im Rahmen des GRK 1808 und für die Zugänglichkeit für die (inter)nationale Forschergemeinschaft haben wir eine Datenbankanwendung spezifiziert, die von einem externen Dienstleister implementiert wird. Dabei setzen wir auf die objektorientierte hierarchische Datenbanktechnologie LDAP (vgl. Zeilenga 2006), die bereits im BMBF-Projekt RiR eingesetzt wurde. So lässt sich nicht nur die XML-Hierarchie bestens abbilden, sondern es wird auch eine sichere und feingranulare Zugriffskontrolle ermöglicht. Mittels einer entsprechend angepassten Synchronisierungssoftware konnte die Datenbank während der XML-Erfassungsphase ständig aktualisiert werden. Eine webbasierte Benutzeroberfläche ermöglicht u.a. komplexe Suchen in den verschiedenen Hierarchieebenen, wobei mehrere Einträge in der Hierarchie (also z. B. ein Haupteintrag sowie mehrere Annotationseinträge und bibliographische Einträge) als ein virtueller Eintrag zusammengezogen werden. Über die Benutzeroberfläche können neue Einträge erstellt und vorhandene Einträge modifiziert werden, wobei die Zugriffskontrolle erlaubt, auch nur Teile eines solchen virtuellen Eintrags sichtbar/bearbeitbar zu machen. Gleichzeitig erlaubt der Export einzelner bzw. aller Datensätze, im XML-Format die Nachhaltigkeit der eingegebenen Daten auch über einen längeren Zeitraum hinweg sicherzustellen.

Damit zeigt das Datenbankprojekt, wie sich interdisziplinäre inhaltliche Arbeit innovativ mit den Zielen der Nachhaltigkeit verknüpfen lässt, ohne die Benutzerfreundlichkeit in der aktiven Arbeitsphase zu vernachlässigen.

Bibliographie

Cutting from BBC News Website, quoted BBC 4, Friday Night Comedy, the News Quiz, Series 82, Episode 2; Broadcasted: 15.Nov 2013.

Hirschberg, Julia / Avesani, Cinzia (1997): „The role of prosody in disambiguating potentially ambiguous utterances in English and Italian“, in: Botinis, Antonis / Kouroupetroglou, Georgios / Carayannis, George (eds.): *Intonation*. Theory, Models and Applications 189–192.

Klein, Wolfgang / Winkler, Susanne (eds.) (2010): *Ambiguität*. Zeitschrift für Literaturwissenschaft und Linguistik 40, 158. Stuttgart: Metzler.

RiR (2012-2015): *Relationen im Raum* http://www.steinheim-institut.de/wiki/index.php/RiR [letzter Zugriff 15. Februar 2016].

TEI Consortium (eds.): *Guidelines for Electronic Text Encoding and Interchange* http://www.tei-c.org/P5/ [letzter Zugriff 15. Februar 2016].

Winkler, Susanne (ed.) (2015): *Ambiguity*. Language and Communication. Berlin / New York: de Gruyter.

Zeilenga, Kurt (2006): *Lightweight Directory Access Protocol (LDAP)*. Directory Information Models, IETF RFC 4512, June 2006.

Annotation natürlichsprachlicher Texte aus Onlineforen zur Entwicklung domainspezifischer Ontologien

Hastik, Canan
hastik@linglit.tu-darmstadt.de
TU Darmstadt, Deutschland

Annotation natürlicher Sprachdaten aus sozialen Medien zur Erforschung zeitgenössischer Szenen, zur Sprach- und Trendanalyse und zur Weiterentwicklung von Sprachtechnologien gewinnt mit der zunehmenden Verfügbarkeit großer Datenbestände weiter an Bedeutung (Farzindar / Inkpen 2015). Zeitgenössische Kommunikation in sozialen Medien verfügt über inhaltliche und strukturelle Besonderheiten und ist von umgangssprachlicher Ausdrucksform geprägt. Beiträge, die im Kontext internetbasierter Diskussionskulturen in Foren entstehen, stellen eine wichtige Forschungsquelle dar. Diese nutzergenerierten Texte, in Form von semi- oder unstrukturierten Kommentaren, repräsentieren Meinungen und Bewertungen einer Gemeinschaft zu einem Thema, Produkt oder Werk und beziehen sich in der Regel auf inhaltliche, technische oder ästhetische Aspekte. Die Autoren verwenden dabei Sprachmittel wie Metaphern, Analogien, Ambiguität, Humor und Ironie sowie metalinguistische bildhafte Mittel wie Emoticons oder andere graphische Zeichen (Reyes et al. 2012).

Vor diesem Hintergrund adressiert dieses Projekt Herausforderungen, die bei der linguistischen und statistischen Verarbeitung von realen web-basierten Daten entstehen. Es wird ein Ansatz semi-automatischer Annotation zur Extraktion von Begriffen für die ontologiebasierte Beschreibung von computergenerierten audiovisuellen Kunstwerken einer digitalen Kunstszene präsentiert. Forschungsgegenstand ist die Diskussionskultur der Demoszene, einer spezialisierten Computerkunstszene. Bisher sind die zahlreichen Beiträge der Gemeinschaft, die sich auf ästhetische und technische Aspekte der Kunstwerke beziehen, nicht erschlossen. Bei diesen Beiträgen handelt es sich um informelle, emotionale, kurze und unstrukturierte Kommentartexte. Das verwendete Vokabular ist mehrsprachig und beinhaltet fachspezifische Terminologien, exklusive Neologismen und einen eigenen szenespezifischen orthographischen Stil. Diese Beiträge bieten detaillierte Einblicke in die Charakteristika der Werke, weshalb ihre Erschließung deren Verständnis fördert und eine gezielte Recherche einzelner Werke ermöglicht. Das Projekt befasst sich mit der Fragestellung, in wieweit sich aktuelle Verfahren der natürlichen Sprachverarbeitung (NLP), die auf grammatikalisch korrekte Schriftformen optimiert und auf Zeitungskorpora trainiert sind, anwenden lassen. Somit leistet das präsentierte Projekt einen Beitrag im Bereich der Entwicklung von Ansätzen zur Aufbereitung großer textbasierter Datenbestände sowie der Erforschung des Sprachgebrauchs zeitgenössischer digitaler Kunstszenen, aber auch hinsichtlich Nutzung semantischer Technologien.

Die Anwendung von NLP-Verfahren für textbasierte Kommunikation in soziale Medien bedarf einiger Anpassungen an die sprachlichen Besonderheiten (Maynard 2012). Die Nutzung standardisierter Techniken ist bisher nur wenig erfolgversprechend (Gimpel 2011; Finin 2010). Bestehende Frameworks, wie das Natural Language Toolkit (NLTK, vgl. Bird et al. 2015), bieten die Möglichkeit der Implementierung eines individuellen NLP-Prozesses, bei dem verschiedene Verarbeitungsschritte modular integriert und miteinander kombiniert werden können. Für das vorliegende Projekt wurde eine Pipeline konzipiert und implementiert, die die Generierung von Annotationsebenen, begonnen mit der Tokenisierung und Part-of-Speech Tagging bis hin zur Extraktion von relevanten werkbeschreibenden Begriffen umfasst. Zur Evaluation des entwickelten Ansatzes wird ein regelbasiertes überwachtes Experiment mit einer definierten Teilmenge von 1255 Kommentaren durchgeführt. Es lässt sich feststellen, dass Emoticons und Partikeln falsch verarbeitet werden. Darüber hinaus werden auch Nomen, Verben und Adjektive, insbesondere Gerundien häufig falsch annotiert. Das Experiment zeigt, dass die konzipierte Pipeline für das vorliegende Kommentarkorpus iterativ optimiert werden muss. Der generierte Index werkbeschreibender Terminologie wird ferner für die Erweiterung einer domainspezifischen Ontologie zur Unterstützung semantischer Annotation verwendet. Hierfür wird ein Ansatz für das Lernen von Ontologien aus Texten verfolgt, wobei die ermittelten Begriffe als Kandidaten für Instanzen beschrieben werden. Als Referenzontologie wird eine auf CIDOC CRM-basierte Adaption verwendet (Hastik et al. 2013).

Dieses Projekt präsentiert einen innovativen Ansatz, um mit NLTK Kommentartexte aus Onlineforen der Demoszene zu annotieren. Das Standard-Tagset muss jedoch angepasst werden. Die Erweiterung der CIDOC CRM-basierten Ontologie auf Basis des generierten Indexes ermöglicht die semantische Beschreibung der Werke.

Bibliographie

Bird, Steven / Klein, Ewan / Loper, Edward (2015): *Natural Language Processing with Python*. NLTK Book http://www.nltk.org/book/ [letzter Zugriff 15. Februar 2016].

Farzindar, Atefeh / Inkpen, Diana (2015): *Natural Language Processing for Social Media*. San Francisco: Morgan & Claypool.

Finin, Tim / Murnane, Will / Karandikar, Anand / Keller, Nicholas / Martineau, Justin (2010): "Annotating Named Entities in Twitter Data with Crowdsourcing", in: *Proceedings of the NAACL HLT* 80–88.

Gimpel, Kevin / Schneider, Nathan / O'Connor, Brendan / Dipanjan, Das / Mills, Daniel / Eisenstein, Jacob / Heilman, Michael / Yogatama, Dani / Flanigan, Jeffrey / Smith, Noah A. (2011): "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments", in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* 42-47.

Hastik, Canan / Steinmetz, Arnd / Thull, Bernhard (2013): "Ontology based Framework for Real-Time Audiovisual Art", in: *IFLA World Library and Information Congress*. 79th IFLA General Conference and Assembly: Audiovisual and Multimedia with Cataloguing http://library.ifla.org/87/1/124-hastik-en.pdf [letzter Zugriff 15. Februar 2016].

Maynard, Diana / Bontcheva, Kalina / Rout, Dominic (2012): "Challenges in Developing Opinion Mining Tools for Social Media", in: *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at International Conference on Language Resources and Evaluation (LREC 2012)* 8.

Reyes, Antonio / Rosso, Paolo / Buscaldi, Davide (2012): "From Humor Recognition to Irony Detection: The Figurative Language of Social Media", in: *Data Knowledge Engineering*. Applications of Natural Language to Information Systems 74: 1-12.

Romantik im Wandel der Zeit – eine quantitative Untersuchung

Hellrich, Johannes

johannes.hellrich@uni-jena.de
Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Deutschland

Hahn, Udo

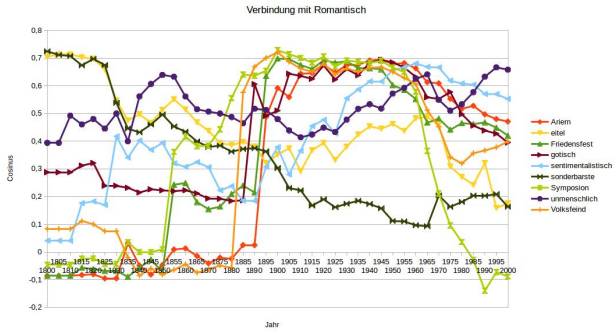
udo.hahn@uni-jena.de
Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Deutschland

„Romantik“ und „romantisch“ bezeichnen heute Profanes, wie ein privates Abendessen in edlem Ambiente bei Kerzenschein oder ein idyllisch gelegenes Hotel mit Butzenglasscheiben, meinte aber ursprünglich Ästhetisches, insbesondere Literarisches (DWB). Dieser Bedeutungswandel kann durch die automatische Analyse eines großen Korpus' quantifiziert werden; etwa unter Verwendung des Google Books N-Gram Korpus, das 4% aller je gedruckten Bücher enthält (Michel et al. 2011). Methodische Grundlage für die Analyse sind Verfahren aus der Computerlinguistik (distributionelle Semantik), mittels derer die Bedeutung von Wörtern über den für sie typischen Kotext (also ihren direkten textuellen Kontext) approximiert wird. Ein aktuelles und mächtiges Verfahren ist word2vec (Mikolov et al. 2013), das auf Forschungsarbeiten zu künstlichen neuronalen Netzen basiert (LeCun et al. 2015). Damit gewonnene Repräsentationen können sowohl synchron als auch diachron auf ihre Ähnlichkeit hin verglichen werden, wodurch Bedeutungswandel quantifiziert werden kann. Dies wurde bereits am Beispiel des Englischen durch Kim et al. (2014) demonstriert, die für einen Zeitraum von 100 Jahren Wortrepräsentationen erzeugten und verglichen. Ein Ergebnis war die Quantifizierung des Bedeutungswandels von „gay“ hin zu einer Bezeichnung für (männliche) Homosexualität. Zwischen den word2vec-Repräsentationen einzelner Wörter sind zudem semantisch sinnvolle arithmetische Operationen möglich (Mikolov et al. 2013). Beim Vergleich der modernen und der historischen Bedeutung von „romantisch“ und „Romantik“ könnte somit ermittelt werden, ob sie einander ähnlicher sind, wenn beispielsweise sexuelle Aspekte ignoriert werden.

Alternative Verfahren zur Quantifizierung von Bedeutungswandel nutzen die Kookkurrenz von Wörtern in Bi-Grammen (Gulordava / Baroni 2011), oder einen auf Nachbarwörtern trainierten Klassifikator (Mihalcea / Nastase 2012). Dabei kann der erste Ansatz lediglich lokale Zusammenhänge erfassen, während der zweite vordefinierte Zeiträume erfordert, zwischen denen ein Bedeutungsunterschied gesucht werden soll. Riedl, Steuer und Biemann (2014) entwickelten einen distributionellen Thesaurus, der für vordefinierte Zeiträume ähnliche Wörter gruppiert. Nachteil dieser Methoden ist wiederum die Notwendigkeit, den untersuchten Zeitraum in Abschnitte zu unterteilen. Vorteilhaft gegenüber word2vec ist die Möglichkeit, die einzelnen Bedeutungen polysemer Wörter getrennt zu erfassen – statt einer veränderlichen Gesamtbedeutung, liegen Teilbedeutungen mit unterschiedlicher Frequenz vor. Eine semantisch schwächere Form des quantitativen Zugangs, aber nützlich für spätere qualitative Interpretationen, ist die Visualisierung von Kollokationen im Zeitverlauf (Jurish 2015).

Die bei unseren Untersuchungen erwarteten Ergebnisse umfassen nicht nur die zunehmende Trivialisierung der Wörter „romantisch“ und „Romantik“ während der letzten 200 Jahre, sondern auch eine Reflexion des deutschen Nationalismus im 19. und 20. Jahrhundert, von dem die Epoche der Romantik instrumentalisiert wurde (Kremer 2007: 50-58). Dem

entspricht etwa die erhöhte Frequenz von „Romantik“ in deutschen Texten der Zwischenkriegszeit und direkten Nachkriegszeit, die mit einer Verdrängung des Bi-Gramms „romantische Liebe“ einhergeht, das dafür Ende des 20. Jahrhunderts seine maximale Popularität erreicht. [1] Auch erste Ergebnisse mit einer an Kim et al. (2014) angelehnten Untersuchung, für die ein word2vec-Modell auf dem deutschen Google Books 5-gram Korpus trainiert wurde, entsprechen dieser Erwartung – Abbildung 1 zeigt, dass „romantisch“ während der ersten Hälfte des 20. Jahrhunderts eine hohe Ähnlichkeit (als Cosinus im Vektorraum) zu „Ariern“ und „Volksfeind“ hatte.



Wörter mit hoher Ähnlichkeit zu „romantisch“ im Zeitverlauf (hoher Cosinus entspricht hoher Ähnlichkeit).

Geplante Folgearbeiten beinhalten, neben der geisteswissenschaftlichen Einordnung der quantitativen Ergebnisse, den Vergleich über mehrere europäische Sprachen hinweg und die Einbeziehung von Sentiment Analysis-Technologien, um die emotionale Ladung der Wörter im Verlauf der Zeit abzubilden (Acerbi et al. 2013).

Die beschriebenen Arbeiten sind Teil eines Promotionsvorhabens am von der Deutschen Forschungsgemeinschaft finanzierten Graduiertenkolleg „Modell 'Romantik'“ der Friedrich-Schiller-Universität Jena.

Notes

1. Am 24.8.2015 ermittelt über https://books.google.com/ngrams/ .

Bibliographie

Acerbi, Alberto / Lampos, Vasileios / Garnett, Philip / Bentley, R. A (2013): "The Expression of Emotions in 20th Century Books", in: *PLoS ONE* 8, 3: e59030.

DWB: *Deutsches Wörterbuch von Jacob und Wilhelm Grimm*. 16 Bde. in 32 Teilbänden. Leipzig 1854–1961. Quellenverzeichnis Leipzig 1971. http://dwb.uni-trier.de/de/ [letzter Zugriff 30. Juli 2015].

Gulordava, Kristina / Baroni, Marco (2011): "A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus ", in: *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics @ EMNLP 2011* 67–71.

Jurish, Bryan (2015): "DiaCollo: On the trail of diachronic collocations", in: *Proceedings of the CLARIN Annual Conference 2015* 28–31.

Kim, Yoon / Chiu, Yi-I / Hanaki, Kentaro / Hegde, Darshan / Petrov, Slav (2014): "Temporal Analysis of Language through Neural Language Models ", in: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* 61–65.

Kremer, Detlef (2007): *Romantik*. Lehrbuch Germanistik. 3. Auflage. Stuttgart: Metzler.

LeCun, Yann / Bengio, Yoshua / Hinton, Geoffrey (2015): "Deep learning", in: *Nature* 521, 7553: 436–444.

Michel, Jean-Baptiste / Shen, Yuan K. / Aiden, Aviva P. / Veres, Adrian / Gray, Matthew K. / The Google Books Team / Pickett, Joseph P. / Hoiberg, Dale / Clancy, Dan / Norvig, Peter / Orwant, Jon / Pinker, Steven / Nowak, Martin A. / Aiden, Erez L. (2011): "Quantitative Analysis of Culture Using Millions of Digitized Books ", in: *Science* 331, 6014: 176–182.

Mihalcea, Rada / Nastase, Vivi (2012): "Word Epoch Disambiguation: Finding How Words Change over Time", in: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* 259–263.

Mikolov, Tomas / Yih, Wen-tau / Zweig, Geoffrey (2013): "Linguistic Regularities in Continuous Space Word Representations", in: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*. Human Language Technologies – NAACL-HLT 2013 746–751.

Riedl, Martin / Steuer, Richard / Biemann, Chris (2014): "Distributed Distributional Similarities of Google Books over the Centuries ", in: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)* 1401–1405.

Erstellung und Visualisierung von Topic-Modellen in WebLicht

Hinrichs, Marie
marie.hinrichs@uni-tuebingen.de
Universität Tübingen, Deutschland

Coltekin, Cagri
cagri.coeltekin@uni-tuebingen.de

Universität Tübingen, Deutschland

Die neueste Erweiterung von besteht aus einer Funktion zur Generierung von Topic-Modellen auf Basis von Nutzerinput. Auch die Visualisierung und Analyse des generierten Modells sind mit WebLicht möglich.

WebLicht (Web-Based Linguistic Chaining Tool) stellt eine virtuelle Forschungsumgebung zur Verfügung, in der Nutzer_innen Verarbeitungsketten zur linguistischen Annotation erstellen und ausführen können und die generierten Annotationen in der Folge visualisieren können. Die WebLicht Webapplikation ist das Frontend eines breiten Frameworks, das darauf abzielt, verbreitete state-of-the-art Annotationswerkzeuge den Forschenden gut zugänglich zur Verfügung zu stellen, ohne dass einzelne Anwendungen heruntergeladen oder installiert werden müssen. Die in WebLicht verfügbaren Werkzeuge werden von den CLARIN Zentren entwickelt und als Webservices zur Verfügung gestellt. Neue Werkzeuge können jederzeit integriert werden, sofern einige Richtlinien befolgt werden. Im Kern muss ein neues Werkzeug hierzu in der CLARIN Center Registry beschrieben werden, als Webservice zur Verfügung gestellt werden und gut definierte Input- und Outputformate benutzen.

Topic-Modelle sind statistische Modelle, die ein Inputdokument in ein Set abstrakter Themen kategorisieren und mit verschiedenen Gewichten oder Prioritäten versehen. Topic-Modelle werden typischerweise automatisch aus einem Set von Dokumenten abgeleitet, ohne dass eine manuelle Annotation notwendig ist. Die resultierenden Modelle können in verschiedensten Aufgaben im Bereich des Natural Language Processing genutzt werden, z. B. zum automatischen Klassifizieren von Dokumenten oder zur Bestimmung verschiedener Wortbedeutungen. Alternativ können sie auch als “Data Mining” Tool genutzt werden, vor allem in Kombination mit passenden Visualisierungen. Auch in den digitalen Geisteswissenschaften ist Topic-Modellierung eine gängige Methode. Insbesondere kann Topic-Modellierung dabei helfen, Muster in großen Textkollektionen zu erkennen. Die Nutzung von Topic-Modellierung in den digitalen Geisteswissenschaften wird unter anderem beschrieben in Jockers (2010) Arbeiten zum Klassifizieren von Blogs beim ‘Day of DH 2010’, in Drouin (2011) Arbeiten zu Proust, in Griffiths und Steyvers' (2004) Arbeiten zu Themen in der Wissenschaft im Verlauf der Zeit und in einer weiteren diachronen Studie von Riddell (2012) zum Thema Topics in der Germanistik.

Da es eines der vorrangigen Ziele von WebLicht ist, Natural Language Processing Tools Geisteswissenschaftlern gut zugänglich zur Verfügung zu stellen, haben wir einen Webservice zur Erstellung und Visualisierung von Topic-Modellen in die WebLicht-Umgebung eingefügt. Das aktuelle Modell nutzt ein Topic-Modell, das im KobRA Projekt entwickelt wurde und auf der weithin bekannte Latent Dirichlet Allocation (LDA) Technik wie von Blei et al. (2003) beschrieben, basiert ist. Das resultierende Topic-Modell wird mit der weit verbreiteten Visualisierungssoftware DFR-browser (cf. Goldstone 2013-2015) visualisiert. DFR-browser bietet vielfältige Visualisierungen, unter anderem von Listen der “Topwörter” zu jedem Topic und von Topic-übergreifenden Worträngen.

Bei der Topic-Modellierung sieht ein üblicher Ablauf wie folgt aus: Der Nutzer lädt eine Textsammlung zu WebLicht hinauf; WebLicht berechnet mit dem oben genannten Webservice ein Topic-Modell mit einer vorgegebenen Anzahl an Topics; die resultierenden Topic-Wort- und Topic-Dokument-Verteilungen werden in das vom DFR-browser benötigte Format konvertiert und im Webbrowser des Nutzers visualisiert. Abbildung 1 enthält eine der Visualisierungsansichten und zeigt die am höchsten eingestuften Wörter für sechs abstrakte Topics in einem Topic-Modell berechnet auf Basis eines großen Zeitungskorpus. Die identifizierten Topics korrespondieren grob mit den Themenfeldern Kultur, Finanzen, Reisen, Politik und Familie. Solch ein Modell kann beispielsweise verwendet werden um Artikel aus einem bestimmten Themenfeld auszuwählen bevor weitere automatische oder manuelle Analysen erfolgen.

Der vorgestellte Webservice arbeitet zur Zeit mit mehreren Dokumenten, die als einzelne Textdatei ohne Metadaten formatiert sein müssen. Der Webservice erlaubt es zur Zeit die gewünschte Anzahl von Themen einzustellen, sowie die häufigsten (zum Beispiel Funktionswörter) oder seltensten (zur Vermeidung von statistischen Störeinflüssen) Wörter herauszufiltern. Die Ergebnisse werden als statische HTML Seite, welche für begrenzte Zeit auf dem Server gespeichert wird, dargestellt. Es ist geplant, zukünftig auch die Metadaten der Dokumente (Autor, Veröffentlichungsdatum usw.) zu verwenden, sowie die Visualisierungen weiter zu verbessern. Abschließend möchten wir bemerken, dass die Integration von Topic-Modellierung in die WebLicht-Umgebung eine Anzahl von Möglichkeiten für die Verarbeitung schafft, wie z.B. das Erstellen von Modellen, die Annotationen (z. B. Part-of-Speech Tags oder syntaktische Relationen) der WebLicht Umgebung nutzen.

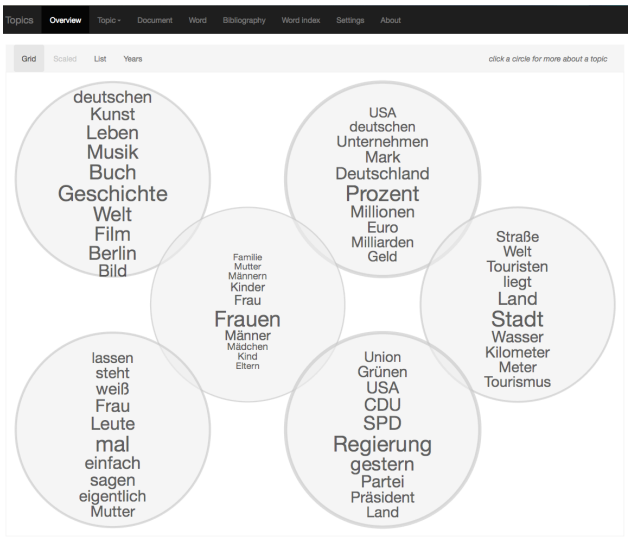


Abb. 1: Beispiel Visualisierung mit DFR-browser.

Bibliographie

Blei, David M. / Ng, Andrew Y. / Jordan, Michael I. (2003): "Latent dirichlet allocation", in: *The Journal of machine Learning research* 3: 993-1022.

CLARIN-D / SfS-Uni. Tübingen (2012): *WebLicht*. Web-Based Linguistic Chaining Tool https://weblicht.sfs.uni-tuebingen.de [letzter Zugriff 11. Februar 2016].

Drouin, Jeff (2011): "Foray Into Topic Modeling", in: *Ecclesiastical Proust Archive* http://www.proustarchive.org/?q=node/35 [letzter Zugriff 11. Februar 2016].

Goldstone, Andrew (2013-2015): *DFR-browser* http://agoldst.github.io/dfr-browser [letzter Zugriff 11. Februar 2016].

Griffiths, Thomas / Steyvers, Mark (2004): "Finding scientific topics", in: *Proceedings of the National Academy of Sciences* 101 (Suppl 1): 5228–5235 [letzter Zugriff 11. Februar 2016].

Jockers, Matthew (2010): "Who's your DH Blog Mate: Match-Making the Day of DH Bloggers with Topic Modeling", in: Matthew Jockers: *Homepage* http://www.matthewjockers.net/2010/03/19/whos-your-dh-blog-mate-match-making-the-day-of-dh-bloggers-with-topic-modeling [letzter Zugriff 11. Februar 2016].

Storrer, Angelika et al. (2012-2015): *KobRA*. Korpus-basierte Recherche und Analyse mit Hilfe von Data-Mining http://kobra.tu-dortmund.de/mediawiki/ [letzter Zugriff 11. Februar 2016].

Das erste dynamische Stemma, Pionier des digitalen Zeitalters?

Hoenen, Armin
hoenen@em.uni-frankfurt.de
Goethe Universität Frankfurt, Deutschland

Einleitung

Ein Stemma codicum ist ein Stammbaum, der jedoch nicht die Verwandtschaft von Wesen oder Sprachen, sondern von Manuskripten darstellt. Dabei symbolisiert ein Knoten in der Terminologie der Graphtheorie ein Manuskript und eine Kante einen Kopierprozess. Bei jedem handschriftlichen Kopierprozess passieren Fehler, die dazu führen, dass alle Manuskripte unterschiedliche Textfassungen aufweisen . [1]

Geschichte der Stemmatologie

Laut Timpanaro (2005) ist das Stemma von C. Schlyter aus dem Jahre 1827 eines der frühesten Stemmata im wissenschaftlichen Sinne überhaupt, s. Abbildung 1. Nachdem Editoren zu Beginn des Printzeitalters für handschriftlich tradierte Texte entscheiden mussten, welche Version eines Textes sie drucken und damit einer breiten Masse zugänglich machen, entstand langsam ein Bewußtsein für die Notwendigkeit einer genauen Analyse der Manuskriptgenealogie, um möglichst präzise den Autorentext wiederzugeben. Im Printzeitalter bedeutete dies die Reise an verschiedene teils weit auseinanderliegende Aufbewahrungsorte der Manuskripte, den händischen Vergleich der Wortlaute der Manuskripte und später den Vergleich durch Photographien (Faksimiles). Gleichzeitig wurde die theoretische Auseinandersetzung mit der richtigen Art und Weise, Übereinstimmungen und Divergenzen in Manuskripten zu interpretieren immer genauer geführt. Als Beispiele sind besonders Bedier (1928) und Maas (1937) zu nennen.

Der Übergang ins digitale Zeitalter erfolgte in den 90er Jahren des letzten Jahrhunderts, vgl. z. B. Robinson and O'Hara (1996); van Reenen et al. (1996); Robinson et al. (1998). Hierbei wurde aus der bio-informatischen Phylogenie viel Methodik entlehnt und die meisten Programme, die bis heute zur Berechnung und / oder Visualisierung in der automatisierten Stemmatologie eingesetzt wurden, stammen aus der Bio-Informatik. Dort herrschen andere Foci, die zu bis dato in der Stemmatologie nicht bekannten neuen Visualisierungen geführt haben (s. u.). Der vorliegende Artikel bezieht sich im Kern nicht auf die digitalen Methoden der Berechnung von Similaritäten in Stemmata, sondern auf die dem auch in der manuellen Stemmatologie folgende Umsetzung in eine geeignete Visualisierung.

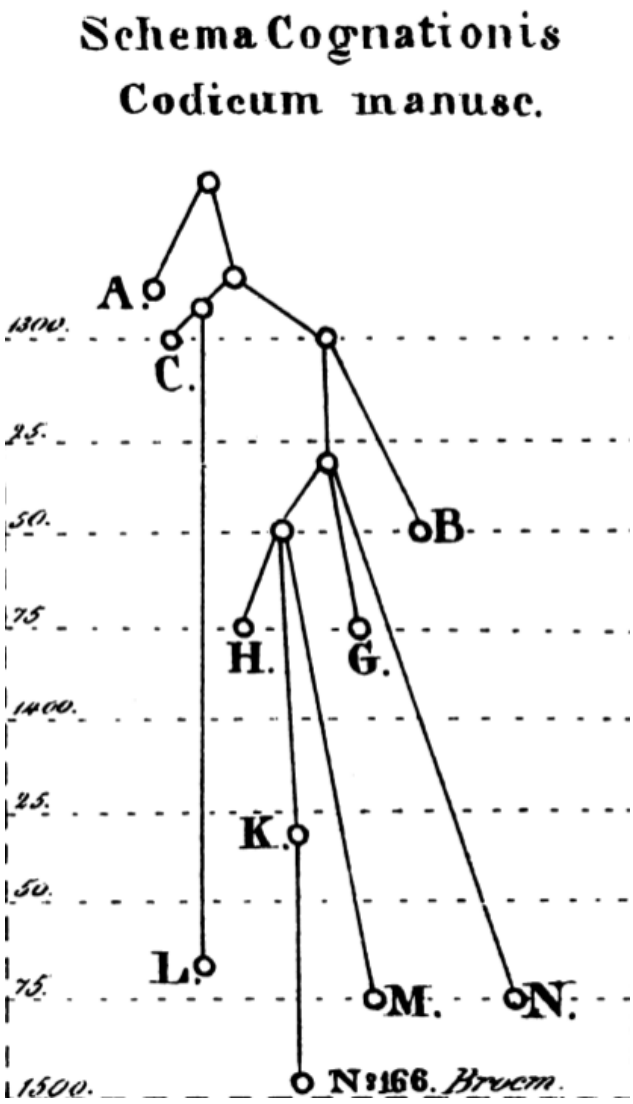


Abb. 1: Frühes Stemma, nach O'Hara (1996), aus H. S. Collin und C. Schlyter, 1827, Corpus Iuris Sueo-Gotorum Antiqui, Z Haeggstrom, Stockholm,1.

Geschichte der Visualisierung

Die Geschichte der Visualisierung eines Stemmas oder von Stammbäumen läuft parallel zur allgemeinen Geschichte der Stemmatologie, weist jedoch Unterschiede zu den anderen Trees-of-history, wie sie O'Hara (1996) getauft hat, auf. Zum Beispiel wurde der buchstäbliche Baum selber zwar sowohl in der Sprachevolution, als auch in der Phylogenie zur Illustration herangezogen, nach bestem Autorenwissen nicht (wenn überhaupt wesentlich seltener) jedoch in der Stemmatologie. Bereits Schlyter benutzt 1827 Kreise als Manuskriptrepräsentatoren und unterlegt den Baum mit einer Zeitskala, die jedoch für das Gros der Stemmata nicht üblich ist. Wenige Konventionen sind in der Philologie für Stemmata allgemein verbindlich. Griechische Buchstaben repräsentieren verlorengegangene Manuskripte; Kontamination, der Prozess zwei Vorlagen für eine Kopie zu benutzen wird durch eine gestrichelte Kante von einem zweiten anzestralen Manuskript her symbolisiert. Seit dem Übergang ins digitale Medium werden Visualisierungen publiziert, die parallel zu biologischen Darstellungen die Wurzel ins Zentrum, aber nicht an den Kopf der Darstellung rücken. Diese aus der Phylogenie entlehnte Darstellungsweise beruht zugegebenermassen auf anderen Foci der Phylogenie, für die die Similarität der Blätter im Mittelpunkt steht, während die Stemmatologie ebenso an den Zwischenknoten interessiert ist. Die Darstellung in Abbildung 2 zeigt ebenfalls eine farbliche Kennzeichnung von Manuskriptgruppen. Die vielfältigen Möglichkeiten der digitalen Darstellung, wie z. B. dieses farbliche Unterlegen sind jedoch noch nicht konventionalisiert.

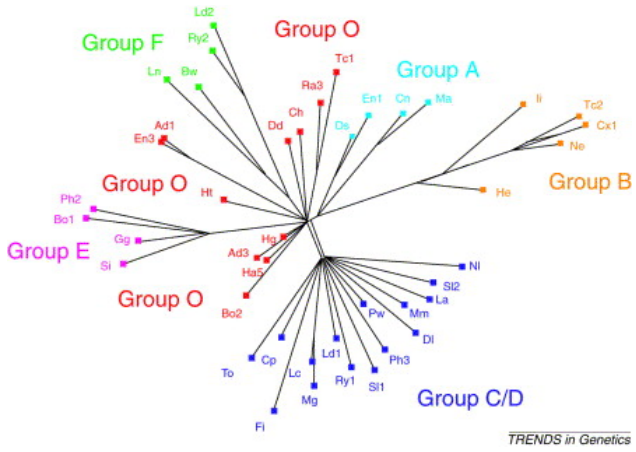


Abb. 2: Ein Stemma, das von bio-informatischer Software (SplitsTree) generiert wurde, Howe et al. (2001).

Dynamisches Stemma

Im digitalen Medium sind Visualisierungen und Berechnungen möglich, die im Printzeitalter schlichtweg nicht realisierbar gewesen sind. Was die Stemmatologie angeht, bezieht sich dies auf die Möglichkeit dynamischer Stemmata. Dynamische Stemmata sind z. B. solche Stemmata bei denen das dynamische Stemma sich aus mehreren Unterstemmata S_0 , .., S_n zusammensetzt, wobei jedes dieser Unterstemmata einen Zeitpunkt ti in der Genese des endgültigen Stemmas repräsentiert. Auf diese Weise können auch Zyklen, wie sie von Andrews and Mace (2013) angesprochen wurden genetisch analysiert werden. Jedes Mal, wenn also eine Manuskriptkopie angefertigt wurde, entspricht dies einem neuen Zeitpunkt ti und damit einem neuen Unterstemma S_i , welches de facto dem Stemma S_{i-1} plus einer Kante (dem Kopierprozess) sowie eines Knotens (des neuen Manuskriptes) entspricht. Entsprechend lassen sich Phänomene wie der Verlust von Manuskripten, z. B. durch eine Bücherverbrennung oder Kontamination einbinden.

Die dynamische Darstellung, s. beispielhaft Abbildung 3 ist eine wichtige Innovation, die im Printzeitalter für jedes Teilstemma eine eigene Abbildung erfordert hätte, wobei diese Abbildungen bis auf ein einziges Detail einander geglichen hätten. Eine solche Darstellung wurde nach bestem Wissen des Autors daher nicht, definitiv nicht als Standard produziert. Eine Posterpräsentation als Printmedium kann jedoch mit Farben oder aufeinanderfolgenden Einzelbildern die Animation simulieren und digital literaten Betrachtern so einen Eindruck des digitalen Erscheinungsbildes vermitteln. Epistemologisch ist der Übergang von einem statischen Stemma zu einem dynamischen gleichzeitig ein Desideratum und eine Aufgabe der digitalen Geisteswissenschaften. Die digitalen Geisteswissenschaften leisten hier den Übergang vom Print ins digitale Zeitalter in der Weise, dass die Möglichkeiten des Digitalen erfasst und bestmöglich zur Erweiterung oder auch Transformation der disziplinenspezifischen Grundlagen, Methoden und Werkzeuge für den hermeneutischen Prozesses umgesetzt werden. Gleichzeitig emanzipiert sich das digitale Stemma hierbei vom statischen Printgegenpart, indem es die bloße Nachahmung des alten im neuen Medium überkommt. Die Dynamisierung von Bäumen ist eine digitale Bereicherung nicht nur für die Stemmatologie, sondern auch für die Phylogenie und Sprachevolution und bietet auf Anhieb mehr visuelle Information als ein statisches Stemma. Durch die Identifikation der Wurzel eines Stemmas, die z. B. durch die Nutzung paläographischer Gegebenheiten forciert werden könnte, kann die dynamische Animation operationalisiert werden oder aber man legt eine manuell erarbeitete Genese zu Grunde. Das Stemma in Abbildung 3 wurde mit der Software LaTeX (Packete tikz und animate) erstellt und durch GIMP in ein gif übertragen. Der Vorteil eines animierten Stemmas ist es dank des Überganges von einem statischen Bild des Endzustandes des Stemmas zu einer animierten Sequenz aufeinanderfolgender Zustände des Stemmas im selben visuellen Raum das intuitive Verständnis der Stemmagenese zu fördern oder vielleicht sogar erst zu ermöglichen.

Notes

1. Auch Prozesse wie willentliche Veränderungen, falsche Rückverbesserungen usw. spielen in der Manuskriptgenese eine Rolle.

Bibliography

Andrews, Tara L. / Macé, Caroline (2013): "Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmas", in: *Literary and Linguistic Computing* 28, 4: 504–521.

Bédier, Joseph (1928): "La tradition manuscrite du 'Lai de l'Ombre': Réflexions sur l'Art d'Éditer les Anciens Textes", in: *Romania* 394: 161–196.

Howe, Christopher J. / Barbrook, Adrian C. / Spencer, Matthew / Robinson, Peter / Bordalejo, Barbara / Mooney, Linne R. (2001): "Manuscript evolution", in: *TRENDS in Genetics* 17, 3: 147-152.

Huson, Daniel H. (1998): "Splitstree: analyzing and visualizing evolutionary data", in: *Bioinformatics* 14: 68–73.

Maas, Paul (1937): "Leitfehler und Stemmatische Typen", in: *Byzantinische Zeitschrift* 37, 2: 289–294.

O'Hara, Robert J. (1996): "Trees of history in systematics and philology", in: *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano* 27, 1: 81–88.

Robinson, Peter / Barbrook, Adrian C. / Blake, Norman / Howe, Christopher J. (1998): "The Phylogeny of The Canterbury Tales", in: *Nature* 394: 839.

Robinson, Peter / O'Hara, Robert J. (1996): "Cladistic Analysis of an Old Norse Manuscript Tradition", in: *Research in Humanities Computing* 4: 115-137.

Timpanaro, Sebastiano (2005): *The Genesis of Lachmann's Method*. Chicago: University of Chicago.

van Reenen, Pieter T. van / van Mulken, Margot (1996): *Studies in Stemmatology* I. Amsterdam / Philadelphia: John Benjamins Publishing Company.

EbnerOnline. Konzept – Realisierung – Probleme. Erfahrungen aus der Praxis Digitaler Editionsarbeit

Hörmann, Richard
richard.hoermann@sbg.ac.at
Universitätsbibliothek Salzburg, Österreich

Weiss, Romedius
me@romediusweiss.com
Universität Innsbruck, Österreich

Konzept

EbnerOnline ist eine im www frei zugängliche Digitale Edition von Gesammelten Werken des österreichischen Dialog- und Sprachphilosophen Ferdinand Ebner (vgl. Universitätsbibliothek Salzburg).

In der aktuellen Version beinhaltet die Edition eine Auswahl aus den Texten Ebners, in erster Linie aus seinen philosophischen Werken, den Tagebüchern und

Aphorismenbänden. Die Auswahl besteht aus einer Reihe von Schlüsseltexten des Autors, die einen repräsentativen Querschnitt aus den verschiedenen Phasen seines Denkens bieten.

Konzipiert war die Edition von Beginn als kommentierte, textkritische Ausgabe. „Textkritisch" heißt, dass es eine diplomatische Fassung gibt, in der der Text mit allen Bearbeitungen des Autors so wiedergegeben wird wie er im Original vorliegt, ergänzt durch eine normalisierte Fassung, die den Text lesbar macht und alle dazu vorgenommenen Änderungen durch die Herausgeber kennzeichnet.

Der Kommentar sollte zusätzliche Informationen bereitstellen, um zu einem besseren Verständnis der Texte zu führen und das Denken Ebners in den größeren Zusammenhang zeitgenössischer Entwicklungen und der europäischen Geistesgeschichte im Allgemeinen einordnen zu lassen.

Basierend auf den Erfahrungen des Brenner-Archivs in Innsbruck, an dem der Nachlass liegt, wurden für die Kommentierung ein Stellenkommentar, eine Biographien-Sammlung und ein Werkverzeichnis ausgewählt. Vor- bzw. Nachworte, Editorische Berichte zu jedem Text, ein Sach- und Personenregister sollten den Kommentar vervollständigen.

Angelegt wurde die Ebner-Edition als Hybridedition, bestehend aus einer Druckausgabe und einer elektronischen Ausgabe im Hintergrund, deren Veröffentlichung im www erfolgen sollte. Das Ziel war, EbnerOnline so aufzubauen, dass auf die Quelltexte und die Metainformationen in der gleichen Weise stabil und dauerhaft zugegriffen werden kann wie das beim analogen Pendant der Fall ist.

Realisierung

Zur Umsetzung dieses Konzeptes wurden die transkribierten Texte in XML-TEI ausgezeichnet. Das markup umfasst dabei sowohl die textkritischen Aspekte, d. h. die Bearbeitungen des Autors und der Herausgeber als auch den Kommentarteil. Zu jedem Text gibt es eine XML-TEI Datei, die den Einzelstellenkommentar enthält. Die Kommentare sind mit fortlaufenden, automatisiert vergebenen ID's eindeutig mit den Lemmata verlinkt. Gleiches gilt für die biographischen und bibliographischen Verzeichnisse.

Für die Biographien wurde mit den entsprechenden TEI-Elementen eine Biographien-Datenbank angelegt, die Kurzbiographien zu den in den Texten genannten Persönlichkeiten enthält, eine Angabe wichtiger Werke und wo möglich und sinnvoll den Bezug der Person zu Ferdinand Ebner. Im Falle von bekannten Personen entfällt die Kurzbiographie zugunsten eines Links auf die entsprechende Wikipedia Seite, die umfassende, inzwischen auch wissenschaftlich anerkannte Informationen bereitstellt.

Trotz der fortgeschrittenen Digitalisierung und stabilen Identifizierung von Werktiteln in den Bibliothekssystemen entschieden sich die Herausgeber auch hier, auf den angebotenen Satz von TEI-Regeln zurückzugreifen und eine bibliographische Datenbank aufzubauen. So können Informationen zu den Werken, die in Bibliothekssystemen (noch) nicht enthalten sind, einheitlich angezeigt und eine stabile Referenz garantiert werden.

In der editionsphilologischen Bearbeitung und Kommentierung musste ein Kosten-Nutzen Kalkül beachtet werden. Das heißt, es wurden nicht alle möglichen TEI-Elemente für das markup verwendet und ein „over tagging" vermieden. Um dennoch zukünftig einen Ausbau des markups zu ermöglichen, werden den Texten Faksimiles beigestellt, die stets eine Überprüfung des tatsächlich Ausgezeichneten an dem potentiell Auszeichenbaren erlauben.

Zur Darstellung und Auswertung der XML-TEI Dokumente im www wurde auf einem Server der Universität Salzburg mit eXistdb eine native XML-Datenbank eingerichtet, die open source ist und in der die Suchmaschine Lucene mit frei konfigurierbaren XQuery Routinen implementiert ist. Die Server-Client Verbindung ist so gestaltet, dass clientseitig unter Einsatz von CSS, HTML und JavaScript (+ Bibliotheken) die stabilen und dynamischen Seitenelemente generiert und über asynchrone http-Anfragen vom Server die Seiteninhalte geliefert werden. Letztere werden am Server mittels xQuery Abfragen und XSLT-Transformationen aus den XML-TEI Dokumenten aufgebaut. Auf diese Weise gelingt es, die Metainformationen nicht nur strukturell, sondern auch visuell mit ihren Lemmata zu verbinden. Auch die Verschachtelung von Metainformationen ist so möglich, ohne dass es zu einem Orientierungsverlust kommt.

Die Suche nach Begriffen und Phrasen erfolgt über eine Suchmaske, in der die zu suchenden Ausdrücke eingegeben werden. In der aktuellen Version wird immer über den gesamten Corpus gesucht, die Ergebnisse werden zunächst in der Maske in Form einer Liste von Werken zurückgeliefert und dann bei Klick auf einen Werktitel in Form der hervorgehobenen und abspringbaren Treffer in dem jeweiligen Text selbst. Eine feinkörnigere Suche, die die Möglichkeiten der umfangreichen Auszeichnung ausnützt (z. B. die Suche nach einer bestimmten Person in bestimmten Werken) musste bisher aufgrund der beschränkten Projekt-Ressourcen unterbleiben.

Probleme

Die während der Arbeit an EbnerOnline aufgetretenen Probleme sind in der Hauptsache auf fehlende oder ungenügende Standards im Bereich Digitaler Editionen zurückzuführen. Die Verwendung von XML-TEI konnte daran nur bezüglich des markups etwas ändern. Und auch hier ist festzuhalten, dass der Umfang der von der TEI angebotenen Regeln so groß ist, dass es dem einzelnen

Projektteam überlassen bleibt, einen eingeschränkten Satz daraus auszuwählen. Das fördert die Flexibilität, macht aber die Intention der TEI, editionswissenschaftliche Standards zu schaffen, ein Stück weit zunichte.

Viel auffallender ist das Fehlen von Standards, wenn es um die Auswahl der anderen Komponenten geht, die für den Aufbau einer Online Edition erforderlich sind. Die Wahl von eXistdb als Datenbank etwa war eine projektinterne Entscheidung, für die es gute Argumente, aber keine Empfehlung a la TEI gab. Dass diese Entscheidung zum Glück die richtige gewesen sein dürfte, scheint die zunehmende Verbreitung dieses Produktes innerhalb Digitaler Editionsprojekte zu belegen.

Die Probleme aus dem Einsatz dieser Datenbank mussten aus den Erfahrungen im Umgang mit ihr ermittelt werden. Beispiele dazu, die im Vortrag genauer beschrieben werden, sind die zu langen Ladezeiten beim Öffnen eines Textes oder die nicht korrekt angezeigten Treffer bei der Suche nach Phrasen, die über mehrere TEI-tags gehen. Problematisch erwies sich das Konzept, die Texte über JavaScript-Funktionen vom Server zu holen. Dadurch können ihnen keine persistenten Identifier zugeordnet werden, die ihre bibliothekarische Aufnahme und ihre Durchsuchbarkeit über globale Suchmaschinen (Google) ermöglichen.

Die derzeit größten Schwierigkeiten bereitet die Sicherstellung der Langzeitarchivierung. EbnerOnline ist das Ergebnis eines Projektes, das 2014 ausgelaufen ist. Die Frage, die sich seitdem stellt, ist, wie die Seite erreichbar bleiben kann, wenn die am Projekt Beteiligten nicht mehr oder nur zum Teil die Betreuung übernehmen können. Eine ausbleibende Integration der Seite in die IT-Infrastruktur der Universität Salzburg würde über kurz oder lang dazu führen, dass EbnerOnline vom Netz verschwindet und die eingesetzten Mittel verpuffen.

Ein effizientes Hosting der Seiten Digitaler Editionen durch Institutionen wie Universitäten hängt auch davon ab, ob es insgesamt gelingt, das Standardisierungsproblem zu beheben, etwa indem ein Baukastensystem geschaffen wird, das die zum Aufbau einer Digitalen Edition benötigten Komponenten frei zugänglich macht. Die gegenwärtig häufig noch sehr hohe Hemmschwelle, eine Online-Edition eines Autors zu beginnen, könnte damit gesenkt und die Anzahl an derartigen Projekten gesteigert werden, was wiederum die Leistungsfähigkeit Digitaler Geisteswissenschaften im allgemeinen stärken würde.

Notes

1. Die Kommentare sind in der aktuellen Version nur für das Hauptwerk Ebners *Das Wort und die geistigen Realitäten* eingeblendet.

Bibliographie

Universitätsbibliothek Salzburg (o. J.): *Ferdinand Ebner Online Edition* http://www.uni-salzburg.at/index.php?id=202485 [letzter Zugriff 15. Februar 2016].

Über die Nutzung von TagPies zur vergleichenden Analyse von Textdaten

Jänicke, Stefan
stjaenicke@informatik.uni-leipzig.de
Leipzig University, Deutschland

Efer, Thomas
efer@informatik.uni-leipzig.de
Leipzig University, Deutschland

Blumenstein, Judith
blumenst@rz.uni-leipzig.de
Leipzig University, Deutschland

Wöckener-Gade, Eva
woeckener-gade@uni-leipzig.de
Leipzig University, Deutschland

Schubert, Charlotte
schubert@uni-leipzig.de
Leipzig University, Deutschland

Scheuermann, Gerik
scheuermann@informatik.uni-leipzig.de
Leipzig University, Deutschland

Motivation

Vor allem durch die häufige Nutzung in den sozialen Medien sind Tag Clouds heutzutage weit verbreitete Visualisierungen um den Inhalt textbasierter Daten zu veranschaulichen. Im Forschungsbereich der Informatik wurden zahlreiche Verfahren zur Berechnung von Tag Clouds entwickelt, unter anderem *Wordle* (Viégas et al. 2009). Abbildung 1 zeigt eine durch Wordle generierte Tag Cloud für die häufigsten Schlagworte aus den fünf Shakespeare Werken *As You Like It*, *Macbeth*, *Othello, Richard III* und *Romeo and Juliet*. Wie für Tag Clouds üblich, wird die Häufigkeit eines Wortes mit Schriftgröße kodiert. Leider tragen die weiteren visuellen Eigenschaften – Farbe, Position und Orientierung – der Darstellung keine Informationen. In unserer Posterpräsentation möchten wir TagPies vorstellen, ein im Rahmen des Digital Humanities Projektes eXChange

entwickeltes Tag Cloud Layout. Während das Design von TagPies bereits vorgestellt wurde (Jänicke et al. 2015), sollen hier repräsentative Anwendungsszenarien illustriert werden. Ein TagPie kann als Hybrid aus Tag Cloud und Pie Chart gesehen werden, bei dem mehrere einzelne Tag Clouds miteinander zu einer Einheit verschmelzen. Im Gegensatz zu herkömmlichen Tag Clouds verwenden TagPies neben Schriftgröße auch Farbe und Position als Informationsträger und ermöglichen damit den Vergleich von Schlagworten verschiedener Kategorien. Abbildung 2 zeigt einen TagPie für die häufigsten Schlagworte aus den oben genannten Shakespeare Werken. Wie in den folgenden Beispielen bestimmt die Größe einer Textklasse im Vergleich zu den anderen die Größe des zugehörigen Kuchenstückes im TagPie. Weitere Beispiele in den Abbildungen 3 und 4 veranschaulichen die sprachunabhängige, vielseitige Anwendbarkeit von TagPies (Jähnicke 2015) – welche als Web-basierte Open Source JavaScript Bibliothek implementiert und damit projektunabhängig einsetzbar sind – für verschiedene geisteswissenschaftliche Fragestellungen.

Abb. 1: Wordle Tag Cloud vereint Schlagworte aller Texte.

Abb. 2: TagPie ermöglicht den Vergleich von Schlagworten zwischen den Texten.

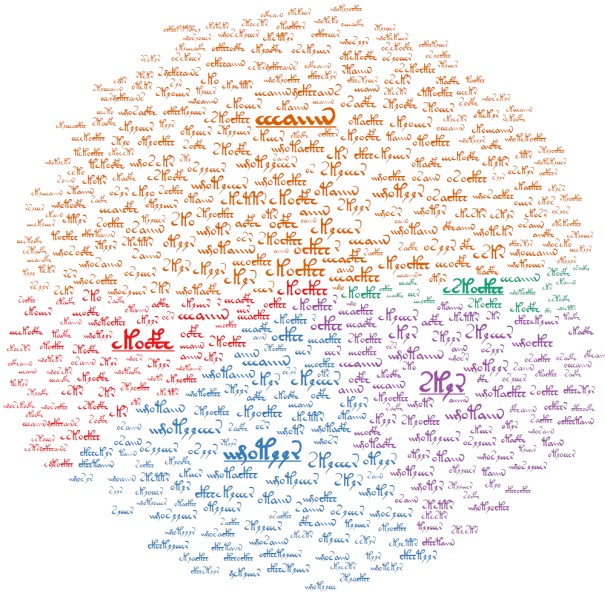


Abb. 3: Das Voynich-Manuskript ist ein mehr als einhundert Seiten umfassendes, in einer Geheimschrift verfasstes Werk (wahrscheinlich aus dem frühen 15. Jahrhundert), welches bis heute nicht entschlüsselt wurde und damit viele Wissenschaftler fasziniert. Die TagPie-Visualisierung ermöglicht die visuelle Exploration von Wort-Kontexten, welche neueren Analysen zufolge (Montemurro / Zanette 2013) interessante Muster aufweisen.

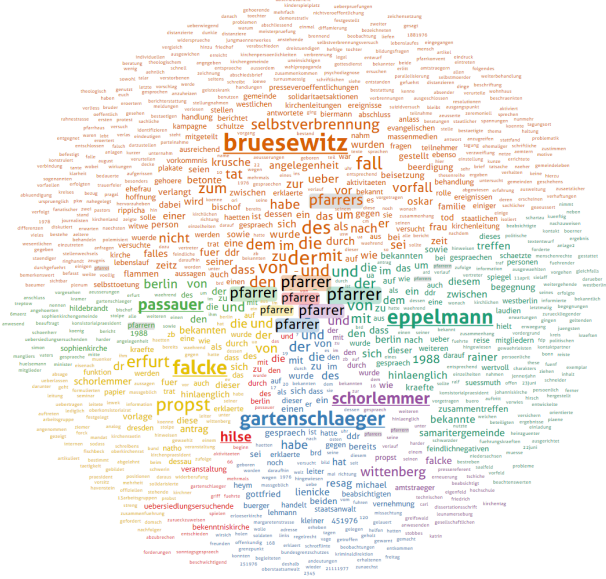


Abb. 4: Das Editionsprojekt “Die DDR im Blick der Stasi” (BStU 2016) stellt die geheimen Stimmungs- und Lageberichte der “Zentralen Auswertungs- und Informationsgruppe” (ZAIG) des MfS einer breiten Öffentlichkeit zur Verfügung. Der TagPie zeigt die Kookkurrenzwörter für die Nachnamen ausgewählter, besonders in den Fokus der Stasi geratener kirchlicher Amtsträger. Weiterführende Informationen in (Kuras et al. 2014).

Über die Nutzung von TagPies in eXChange

Das Digital Humanities Projekt eXChange untersucht den Gebrauch medizinischer und politischer Fachtermini in antiken griechischen und lateinischen Texten. Zur explorativen, vergleichenden Analyse sind TagPies in eine Rechercheplattform eingebunden, die den Geisteswissenschaftlern einen dynamischen Zugriff auf Textstellen ermöglicht, die eingegebene Suchterme enthalten. Jeder Suchterm wird innerhalb des TagPies durch einen Sektor repräsentiert, der seine häufigsten Kookkurrenzen anzeigt. Im Vergleich zu traditionellen Suchergebnislisten geben TagPies damit einen schnellen Überblick über die Kontexte, in denen die Suchterme verwendet wurden. Der TagPie ist interaktiv explorierbar und ermöglicht Close Reading durch die Selektion eines Schlagwortes via Mausklick, welches Textpassagen anzeigt, die sowohl den Suchterm als auch die gewählte Kookkurrenz enthalten. Im Folgenden sind zwei typische Anwendungsszenarien aus dem eXChange Projekt erläutert.

Vergleich von gibbus **und** gibbosus

Für den Ausdruck “bucklig” findet man in lateinischen Wörterbüchern die Synonyme “gibbus” und “gibbosus”. Mit Hilfe von TagPies soll diese Synonymie überprüft werden (Abbildung 5); das Ergebnis berücksichtigt die Belegstellen aller Deklinationsformen der Suchterme. Die schwarz eingefärbte Schnittmenge im Zentrum liefert besonders auf den Körper bezogene Wörter wie “triefäugig” (lippus), “Fuß” (pede) oder “gebrochen” (fracto). Der grüne Sektor für “gibbus” zeigt weitere physische Begriffe wie “Rücken” (dorso), “Kopf” (caput) und “Gehirn” (cerebri). Der rote Sektor liefert die Resultate für “gibbosus”. Sie zeigen auffällig viele Begriffe aus dem Sachfeld der christlichen “Moral”, etwa “Begierde” (cupiditatis), “geizig” (avarum), “Mäßigung” (modestia) oder “sich rühmen” (glorietur). Der TagPie zeigt demnach eine Tendenz, dass „gibbus“ eher bei physischen Eigenschaften, „gibbosus“ eher bei moralischen Eigenschaften verwendet wurde (vgl. physisch: Cels. IV 1,5; VIII 1,23; Iuv. 6,109; 10,294; moralisch: christliche Exegeten). Ein Wörterbuch wie der Thesaurus Linguae Latinae differenziert die Semantik zwar auch in “eigentlich” und “übertragen“, aber über die Häufigkeit und Tendenz der Verwendung beider Begriffe gibt es im Vergleich zu den TagPies keine Hinweise. Damit übersteigen TagPies den “rekonstruktiven” Charakter von Wörterbüchern, indem sie auf die reale Verwendung in den Textkorpora rekurrieren.

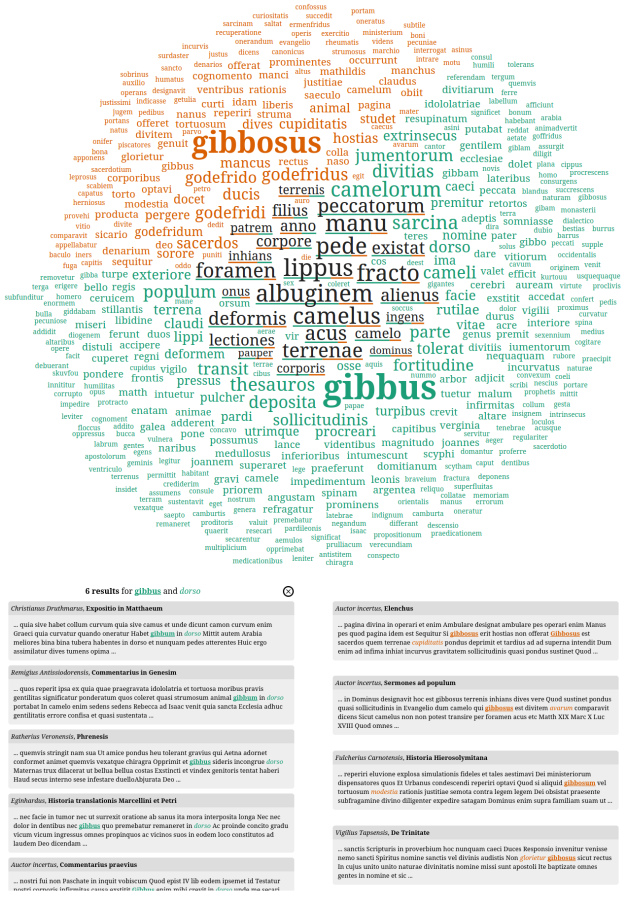


Abb. 5: Vergleichende Analyse von gibbus und gibbosus mit Belegstellen für gibbus und dorso sowie aus dem Sachfeld der christlichen “Moral”.

Vergleich von ατρεμ*, ησυχ* und ακινητ*

Für die datenbankbasierte Suche wurden die trunkierten Formen ατρεμ*, ησυχ* und ακινητ* – jeweils Synonyme für “ruhig” bzw. “unbewegt” – verwendet. Weiter sollte in der vergleichenden Analyse untersucht werden, ob ατρεμ* als medizinischer Fachterminus zu bezeichnen ist. Der resultierende TagPie (Abbildung 6) verdeutlicht aufgrund der vielen gemeinsamen Kookkurrenzen der Begriffe, gekennzeichnet durch die unterstrichenen Terme im mittleren Bereich, dass die drei Wortgruppen in der Literatur häufig synonym verwendet werden. ησυχ* liefert mit großem Abstand die meisten Belege und ist in den verschiedensten Kontexten zu finden. Für ακινητ* gibt es viele Kookkurrenzen mit κινουν (bewegen), zurückzuführen auf die aristotelische Philosophie, in der der Gegensatz zwischen Unbewegtheit und Bewegtheit eine große Rolle spielt. Da verschiedene medizinische und anatomische Begriffe ausschließlich in den Kontexten von ατρεμ* auftreten, kann man ατρεμ* als medizinischen Fachterminus bezeichnen. Beispiele hierfür sind Formen von διαφορητικος (“schweißtreibend” oder “streuend”, u. a. von Tumoren und Medikamenten) oder

schlicht ητορ “Herz”. Die entsprechenden Belegstellen finden sich dann u. a. in den medizinischen Schriften des Galen und Oribasius. Weitere Analysen der einzelnen Belegstellen, aber auch weitere Visualisierungen mithilfe der TagPies (z. B. mit Eingrenzung auf das Corpus Hippocraticum) sind geplant, um die qualitative Auswertung der Ergebnisse fortzuführen.

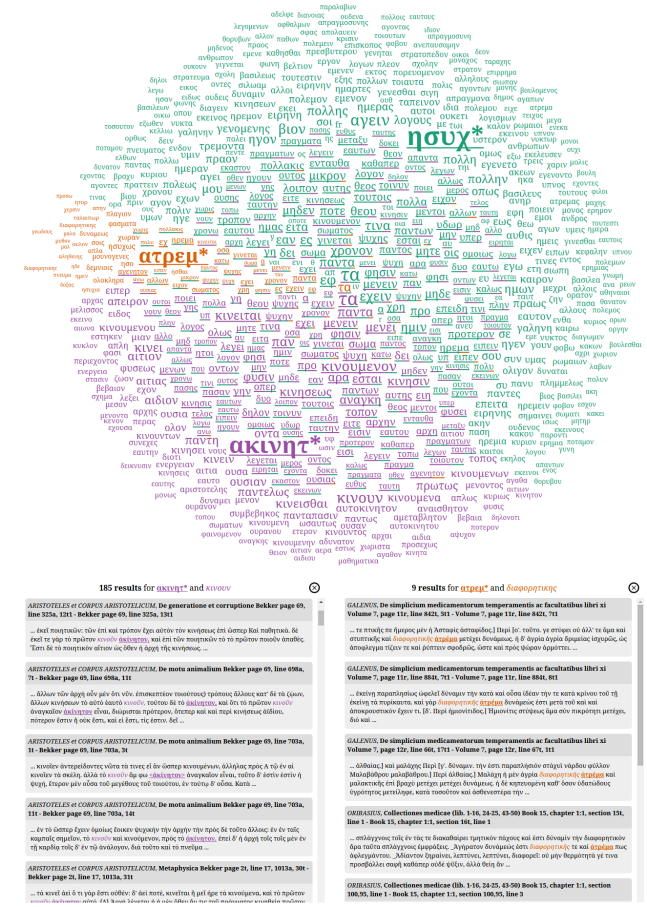


Abb. 6: Vergleichende Analyse von ατρεμ*, ησυχ* und ακινητ* mit Belegstellen für für ακινητ* und κινουν sowie ατρεμ* und διαφορητικος.

Zusammenfassung

Die Distant Reading Visualisierung TagPies ermöglicht den Vergleich verschiedener Kategorien von Schlagworten innerhalb einer Tag Cloud. Im Rahmen des Projektes eXChange hat sich die Verwendung von TagPies zur vergleichenden Analyse von Fachtermini in antiken Texten als wertvoll erwiesen. In einer Posterpräsentation würden wir typische Anwendungsszenarien der beteiligten Geisteswissenschaftler_innen demonstrieren.

Bibliographie

BStU = Bundesbeauftragter für die Unterlagen des Staatssicherheitsdienstes der ehemaligen DDR (2016): *Die DDR im Blick der Stasi*. Die geheimen Berichte an die SED-Führung. http://www.ddr-im-blick.de/ [letzter Zugriff 12. Februar 2016].

Jänicke, Stefan (2015): "TagPies", in: *vizcovery.org* http://www.tagpies.vizcovery.org/ [letzter Zugriff 12. Februar 2016].

Jänicke, Stefan/ Blumenstein, Judith / Rücker, Michaela / Zeckzer, Dirk / Scheuermann, Gerik (2015): "Visualizing the Results of Search Queries on Ancient Text Corpora with Tag Pies". To appear in *Digital Humanities Quarterly*.

Kuras, Christoph / Efer, Thomas / Adam, Christian / Heyer, Gerhard (2014): "The GDR Through the Eyes of the Stasi – Data Mining on the Secret Reports of the State Security Service of the former German Democratic Republic", in: Fred, Ana / Filipe, Joaquim (eds.): *Proceedings of the 6th International Conference on Knewledge Discovery an Information Retrieval (KDIR)*, Rom. Lisbon: SCITEPRESS 360–365.

Jänicke, Stefan / Blumenstein, Judith / Rücker, Michaela / Zeckzer, Dirk / Scheuermann, Gerik (2015): Visualizing the Results of Search Queries on Ancient Text Corpora with Tag Pies. To appear in *Digital Humanities Quarterly*, 2015.

Montemurro, Marcelo A. / Zanette, Damián H. (2013): "Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis", in: *PloS one* 8, 6: e66344.

Universität Leipzig Historisches Seminar (2013): *eXChange - Exploring Concept Change and Transfer in Antiquity* http://exchange-projekt.de/index.html [letzter Zugriff 12. Februar 2016].

Kuras, Christoph / Efer, Thomas / Adam, Christian / Heyer, Gerhard (2014): "The GDR Through the Eyes of the Stasi – Data Mining on the Secret Reports of the State Security Service of the former German Democratic Republic", in: Fred, Ana / Filipe, Joaquim (eds.): *Proceedings of the 6th International Conference on Knewledge Discovery an Information Retrieval (KDIR)*, Rom. Lisbon: SCITEPRESS 360–365.

Viégas, Fernanda B. / Wattenberg, Martin / Feinberg, Jonathan (2009): "Participatory Visualization with Wordle", in: *IEEE Transactions on Visualization and Computer Graphics* 15, 6: 1137–1144.

Montemurro, Marcelo A. / Zanette, Damián H. (2013): Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis. *PloS one*, 8, 6: e66344.

Viégas, Fernanda B. / Wattenberg, Martin / Feinberg, Jonathan (2009): "Participatory Visualization with Wordle", in: *IEEE Transactions on Visualization and Computer Graphics* 15, 6: 1137–1144.

Lizenzauswahlwerkzeuge für die digitalen Geisteswissenschaften

Kamocki, Pawel
pawel.kamocki@gmail.com
Institut für Deutsche Sprache Mannheim, Deutschland

Ketzan, Erik
ketzan@ids-mannheim.de
Institut für Deutsche Sprache Mannheim, Deutschland

Witt, Andreas
witt@ids-mannheim.de
Institut für Deutsche Sprache Mannheim, Deutschland

Tatsächlich werden allerdings, obwohl Transparenz und Reproduzierbarkeit von Forschungsergebnissen seit langem als die Grundpfeiler der wissenschaftlichen Gemeinschaft gelten, in der Praxis viele digitale Datensätze und Werkzeuge unter Lizenzen verbreitet, die unnötig restriktiv bzw. nicht zweckmäßig sind, oder bleiben gar vollständig ohne Lizenz. Dies liegt vor allem daran, dass die Wahl einer geeigneten Lizenz einem durchschnittlichen Wissenschaftler ohne hinreichende rechtliche Beratung Schwierigkeiten bereiten kann.

Als Lösung wurden einige Tools entwickelt, die den Nutzer durch den Dschungel der unzähligen, verfügbaren, „öffentlichen Lizenzen“ leiten und ihm so erlauben, die Lizenz zu wählen, die für seine Zwecke am geeignetsten ist. Diese Werkzeuge werden auf Englisch als „License Choosers“, „License Selectors“ oder „License Wizards“ bezeichnet; im Folgenden werden wir sie als Lizenzauswahlwerkzeuge bezeichnen.

Bevor diese näher beschrieben werden, soll zunächst der Begriff der „öffentlichen Lizenz“ definiert werden. Eine öffentliche Lizenz ist eine Lizenz, die nicht einem individuellen Nutzer, sondern der breiten Öffentlichkeit (also jedem potenziellen Nutzer) bestimmte Rechte einräumt. Für Software gibt es solche öffentlichen Lizenzen bereits seit den 1980er Jahren. Damals entstanden Software-Lizenzen wie die BSD-License, die MIT Lizenz oder die GNU GPL. In anderen Wissenschaftsbereichen oder im Bereich der Textdaten kamen solche öffentliche Lizenzen allerdings erst Anfang des 21. Jahrhunderts infolge der Gründung der Creative Commons (CC) Foundation auf.

Die jüngste CC-Lizenzversion, CC 4.0., die 6 verschiedene Lizenzen, einen „Waiver“ und eine „Public Domain Mark“ bereithält, eignet sich sehr gut für Datensätze, da sie nicht nur das Urheberrecht, sondern auch das Recht an Datenbanken erfasst. Auch ältere Versionen sind aber noch in Verwendung.

Bei der Wahl einer Lizenz ist darauf zu achten, dass die Lizenzen, die für Software geeignet sind, nicht gleichermaßen für Daten geeignet sind und umgekehrt.

Darüber hinaus sind nicht alle öffentlichen Lizenzen „frei“, da einige Lizenzen beispielsweise nicht den Anforderungen für das Open Access / Open Data / Open Source Label genügen.

In unserer Präsentation werden wir kurz drei Online-Tools vorstellen, die speziell für die Lizenzierung von Forschungsmaterialien entwickelt wurden.

Das Licentia Tool wurde im Jahr 2014 von Cristian Cardellino für INRIA (Französisches Institut für Forschung in Informatik und Automatisierung) entwickelt und ist genau genommen ein Verbund dreier Tools: einer Lizenzsuchmaschine, die es ermöglicht, Lizenzen zu finden, die genau die Anforderungen erfüllen, welche zuvor durch den Nutzer festgelegt wurden, eines Lizenzkompatibilitätsprüfers, der die Kompatibilität von unterschiedlich lizenzierten Daten prüft und eines Lizenz-Visualizers, der graphikbasierte Visualisierungen der Lizenzen in Open Digital Rights Language Deontology (ODRL) darstellt.

Der ELRA License Wizard, der im April 2015 von der European Language Resources Association veröffentlicht wurde, ermöglicht es dem Nutzer, bestimmte Eigenschaften festzulegen, um dann eine Auswahl an korrespondierenden Lizenzen zu treffen.

Schließlich stellen wir den License Selector vor, der im Jahr 2014 von Pawel Kamocki, Pavel Stranak und Michal Sedlak im Rahmen einer Kooperation zweier CLARIN-Zentren, dem IDS Mannheim und Karls-Universität Prag entwickelt wurde. Er benutzt einen Algorithmus (eine Serie von Ja / Nein-Fragen), der den Nutzer durch den Lizensierungsprozess begleitet. Er kann sowohl für die Lizenzwahl von Daten als auch von Software verwendet werden und bietet ein integriertes Lizenzinteroperabilitätswerkzeug. Lizenzen, die die Anforderungen für das Open Access, Open Data und Open Source Label erfüllen, sind dabei besonders gekennzeichnet. Anders als die vorstehenden Tools ist es unter Open Software- und Open Data-Lizenzbedingungen verfügbar.

Alle drei Tools haben Vor- und Nachteile. Ihr größter Nachteil ist, dass sie in verschiedenem Ausmaß eine sehr spezifische Sprache nutzen, die ein juristisches Grundverständnis insbesondere zum Urheberrecht erfordern. Zudem bedienen sie sich alle - notwendigerweise - zu einem gewissen Grad einer Verallgemeinerung, insbesondere um die Lizenzinteroperabilität zu beurteilen. Nichtsdestotrotz sind sie für die Forschungsgemeinschaft als sehr hilfreich einzuschätzen, vereinfachen sie doch die Nachnutzung von Tools und Daten in den Digital Humanities ganz erheblich.

Bibliography

Cardellino, Cristian (2014): *Licentia*. Licenses Visualizer http://licentia.inria.fr/visualize [letzter Zugriff 15. Februar 2016].

ELRA License Wizard (2015): http://wizard.elda.org [letzter Zugriff 15. Februar 2016].

Kamocki, Pawel / Stranak, Pavel / Sedlak, Michal (2014): *License Selector* http://ufal.github.io/lindat-license-selector/ [letzter Zugriff 15. Februar 2016].

Die Uwe Johnson-Werkausgabe

Kaßner, Fabian
fabian.kassner@uni-rostock.de
Universität Rostock, Deutschland

Kischel, André
andre.kischel@uni-rostock.de
Universität Rostock, Deutschland

Das Vorhaben

Die Uwe Johnson Werkausgabe ist ein Vorhaben der Berlin Brandenburgischen Akademie der Wissenschaften an der Universität Rostock. Erstmals wird mit Uwe Johnson ein Autor des 20. Jahrhunderts in einem Akademienvorhaben ediert. Die Werkausgabe gliedert sich in die Abteilungen *Werke*, *Schriften* und *Briefe* und wird sowohl als Buch wie auch digital erscheinen. Für beide Medien bilden gemäß TEI (P5) ausgezeichnete Texte die gemeinsame Datengrundlage, im Digitalen ergänzt um die digitalisierten Materialien des Uwe Johnson-Archivs. Um „lesbare“ Bücher zu ermöglichen, werden textkritische Apparate, Kommentare und Erläuterungen in einer exemplarisch sinnstiftenden Auswahl in den Druck aufgenommen, digital wird historisch-kritische Vollständigkeit angestrebt.

Zu den Eigenheiten des Autors Uwe Johnson gehört eine textsortenübergreifende Arbeitsmethode: In Briefen formuliert er Prosaversuche, aus Zeitungen übersetzt er in seine Bücher, Personen, Dokumente und Ereignisse der Zeitgeschichte treten in seinen Romanen auf.

Aus der Werkstatt

Aus dieser Arbeitsweise ergibt sich beinahe zwingend, dass im Digitalen mit den Buchdeckeln auch die Grenzen zwischen den drei die Buchausgabe gliedernden Abteilungen durchlässig werden. Anhand einiger Beispiele soll Johnsons Arbeiten gezeigt werden: welche Brücken zwischen Texten und unterschiedlichen Medien er schlägt, und welche Auswirkungen das auf einen Editionsprozess hat, der diesen Pfaden nachgehen will, um sie den Nutzern zu präsentieren. Hinzu kommen Fragen der Einbindung externer Ressourcen, audiovisuellen Materials und der Berücksichtigung von Urheberrechtsfragen.

Durch die beiden Zielmedien ergibt sich ein unterschiedlicher Bedarf in der Tiefe der TEI-Auszeichnung der jeweiligen Dokumente. Die Ansprüche an die Auszeichnung für ein Buch sind selbstverständlich andere als die, für eine digitale Edition, in der die Grenzen lediglich durch die editorischen Kriterien gezogen werden. Das Erforderliche kann aufgenommen werden, während auf zu weit Entferntes oder Weiterführendes nur hingewiesen wird. Aus diesem Grund wird für jedes Werk eine vollausgezeichnete „Masterdatei“ erstellt, aus der sich beide Versionen speisen und somit dieselbe Textgrundlage haben, wodurch eine doppelte Datenhaltung vermieden wird. Neben der Erfassung der verschiedenen Varianten finden sich in dieser Datei sowohl der text- als auch der historisch-kritische Kommentar. Das Ziel der Buchfassung, nämlich eine lesbare Ausgabe zu sein, die das zum Textverständnis Notwendige enthält, steht dem der digitalen Edition gegenüber, welche losgelöst der analogen Grenzen funktioniert. Hier kann vom Optimum ausgegangen werden und dann durch den Nutzer entsprechend seiner Fragestellung reduziert werden. Durch Zuweisung von Attributen ist es möglich per XSL Transformation das gewünschte Endprodukt zu generieren. Dazu werden einerseits einige der typischen TEI-Elemente als „Filtersignal“ genutzt, andererseits ein „Digital“- oder „Print“-Attribut, welches in dem Prozess wieder entfernt wird. Die so generierte print-XML wird per XSL-FO in PDF umgewandelt, um eine orientierende Vorlage für den Verlag zu erhalten, der diese beiden Dateien für den Satz erhält. Nach dem gleichen Prinzip wird die Digital-XML als Grundlage der digitalen Edition generiert.

Künftig wird eine weitere Auszeichnungsebene hinzukommen. In der praktischen Arbeit hat sich gezeigt, dass eine zusätzliche semantische Auszeichnung sinnvoll ist. Daher konzipieren wir zurzeit eine Möglichkeit neben der TEI- auch eine semantische Auszeichnung vorzunehmen. In welcher Ebene dies geschieht ist noch in der Überlegung. Anzunehmen ist ein erneutes Transformationsszenario.

Johnsons Arbeitsmethoden bieten einen sehr großen Anreiz zur Vernetzung und Visualisierung in der digitalen Präsentation. Intertextualität und Collage sind nur zwei von vielen Termini, die sich auf ihn anwenden und im digitalen Raum auf eine Art und Weise präsentieren lassen, wie es in einem Buch nicht möglich wäre. So können Zusammenhänge sicht- und erfahrbar gemacht werden, die sonst nur Abstrakt zu erfassen wären.

Dazu gehört auch die angestrebte Multimedialität. Johnson hielt vielerlei Lese- und sonstige Reisen ab, welche zum Teil als Video- oder Audiodokument vorliegen. Im Zusammenspiel mit seinen Manuskripten – er hat oft dokumentiert, wo er was geschrieben hat –

lässt sich die Entstehung seines Œuvres auch räumlich nachverfolgen und etwaige Koinzidenzen erforschbar machen.

Die spätere Präsentation dieser Ergebnisse wird sich der Nutzer gemäß seinen Bedürfnissen komplett selbst anpassen können. Die Textvarianten, die von Interesse sind, werden sich nach eigenem Wunsch anordnen, zu- oder abschalten lassen. Die einzelnen Anzeigeelemente werden also nicht in einem Layout „gefangen“ sein, sondern können durch den Nutzer selbst angeordnet werden. Entscheidungen, welche Varianten wie nebeneinander gelegt werden und ob ein Faksimile hinzugeschaltet werden soll liegen in der Hand des Benutzers, der seinen Schreibtisch in der Art anordnen kann, wie es seiner Arbeit dienlich ist.

Digitales Publizieren in den Geisteswissenschaften - Abschlussbericht und Handlungsempfehlungen des DFG-Projektes Fu-PusH

Kleineberg, Michael
michael.kleineberg@ub.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Kaden, Ben
ben.kaden@ub.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Projekt

Das DFG-Projekt Future Publications in den Humanities (Fu-PusH) untersuchte die Potenziale des digitalen Publizierens in den Geisteswissenschaften und erarbeitete anhand von Szenarien Handlungsempfehlungen für akademische Infrastruktureinrichtungen wie insbesondere Universitätsbibliotheken und Rechenzentren, um Publikationsprozesse zu unterstützen und dabei den funktionalen Anforderungen unterschiedlicher geisteswissenschaftlicher Fachrichtungen gerecht zu werden.

Auf dem Poster werden zum einen die Ergebnisse der Studie beschrieben und zum anderen ein speziell in diesem Projekt entwickeltes Recherche-Tool zur Auswertung qualitativer Interviews (Statement Finder) vorgestellt, das als niedrigschwelliges Open-Source-Tool der Community zur Verfügung gestellt wird. Abschließend werden eine Reihe von Handlungsempfehlungen formuliert für die an digitalen Publikationsprozessen beteiligten Akteursgruppen, namentlich für geisteswissenschaftliche Fachgemeinschaften darunter insbesondere die Digital-Humanities-Community, für Infrastruktureinrichtungen wie Bibliotheken und Rechenzentren, für Wissenschaftsverlage sowie für Förderinstitutionen und für die Wissenschaftspolitik.

Ergebnisse

Die Ergebnisse des Fu-PusH-Projektes zeigen sehr deutlich die Unterschiede im Forschungs- und Publikationsverhalten sowohl zwischen den Geisteswissenschaften und den Naturwissenschaften als auch innerhalb des disziplinären Spektrums der Geisteswissenschaften selbst. Dies betrifft insbesondere die Zurückhaltung gegenüber der Nutzung digitaler Publikationsmedien, auch angesichts ihrer anerkannten Potenziale.

Das **Publikationsverhalten** in den Geisteswissenschaften orientiert sich nach wie vor weitgehend an traditionellen Formen aus der Printkultur wie Monografien, Sammelbandbeiträge, Zeitschriftenaufsätze sowie Rezensionen. Wo digital publiziert wird, folgt man den etablierten Modellen der Verlagspublikation in einem dem Printparadigma möglichst ähnlichen Format. Hier sind auch perspektivisch nur geringe oder selektive Änderungen und Optimierungen zu erwarten. Die Einbindung von multimedialen Erweiterungen wird auf der Materialebene durch urheberrechtliche Bedingungen und auf der technischen Ebene durch den Mangel an Standards und niedrigschwelligen Lösungen eingeschränkt. Bisher lässt sich am ehesten die Form des Bloggens als dauerhafte zusätzliche Variante für die wissenschaftliche Kommunikation bestimmen.

Das Publizieren nach dem **Open-Access-Prinzip** scheint in den Geisteswissenschaften geringer ausgeprägt als in den Naturwissenschaften. Dafür lassen sich mehrere Gründe identifizieren. Zum einen fehlen an vielen Stellen bislang fachwissenschaftlich etablierte Infrastrukturen. Zum anderen genießen rein digitale Publikationen nach wie vor keinen guten Ruf, was sich beispielsweise auf die Kreditierung des Forschungsoutputs auswirkt. Schließlich wirkt im Vergleich zu naturwissenschaftlichen Publikationen der Zugangsdruck zu Neuerscheinungen an vielen Stellen durch längere Forschungszeiträume und geringere Kosten weniger stark.

Auffällig ist, dass sich stärker in Schnittstellen mit Naturwissenschaften befindliche und internationalisierte Disziplinen (z. B. Sprachwissenschaften, Archäologie) deutlich aktiver in dieser Richtung entwickeln, als die vorwiegend hermeneutisch-interpretativ arbeitenden Fächer. Eine Nutzung von frei zugänglichen Materialien erfolgt dagegen fächerübergreifend. Es existiert beim

Open Access also eine Diskrepanz zwischen Publikations- und Rezeptionsverhalten.

In einigen Bereichen vor allem unter dem Einfluss der **Digital Humanities** finden sich jedoch auch stärker digital orientierte Entwicklungen. Eine Erklärung lautet, dass viele Formen dieser Wissenschaft überhaupt erst durch digitale Technologien realisierbar werden. Dort wo größere Datenmengen flexibel verarbeitet werden müssen, etwa in der Editionswissenschaft oder der Computerlinguistik, finden sich bereits stärker etablierte Formen der digitalen Forschung und des digitalen Publizierens, die – sehr selektiv – auch auf anderen Fachbereiche inspirierend einwirken. Eine Zwischenform zwischen Publikation und Forschung, der vergleichsweise viel Potential zuerkannt wird, ist das digitale **Annotieren**. Damit zusammenhängend wird das größte Zukunftspotenzial des digitalen Publizierens im Bereich der **digitalen Editionen** gesehen, die häufig zugleich als mögliche Hybridausgaben zur differenzierten Rezeption wie auch als digitales Forschungsdatum zur weiteren Verarbeitung gesehen werden.

Die Nutzung von **Social-Media-Anwendungen** scheint sich in vielen Bereichen der Geisteswissenschaften weitgehend auf die Vernetzung durch soziale Wissenschaftsnetzwerke oder Kurznachrichtendienste (Twitter) zu beschränken. Mit Hypotheses.org etabliert sich allerdings nach und nach eine Blogplattform, die sich durchaus ein gewisses Renommee aufbaut. Das ist insofern der relevante Schritt, weil eine zentrale Hürde bei der Nutzung solcher Medien die bisher fehlende Kreditierbarkeit für wissenschaftliche Karrieren darstellt. Zudem existiert die Sorge, dass frei auf solchen Wegen zum Beispiel vor einer “ordentlichen” Publikation publizierte Ergebnisse von Anderen übernommen und verwertet werden.

Bei vielen Aspekten vernetzter und digitaler Forschung bzw. des interaktiven Publizierens zeigt sich, wie sehr wissenschaftskulturelle Aspekte der Nutzung bestimmter technologischer Formen entgegenstehen. Das betrifft insbesondere den Aspekt der Kollaboration, der Voraussetzung für den sinnvollen Einsatz **virtueller Forschungsumgebungen** ist. Hier findet sich nur eine geringe Nutzungsbereitschaft. Es ist zu vermuten, dass sowohl wissenschaftskulturelle Gepflogenheiten als auch eine vergleichsweise komplexe Nutzbarkeit die Akzeptanz und Nutzung solcher Angebote bremsen. Zweckmäßiger erscheinen hier einfache, modularisierte und miteinander verknüpfbare Lösungen.

Herausforderungen werden generell bei Fragen der technischen **Standardisierung** zur Gewährleistung von **Interoperabilität** deutlich. Dies betrifft sowohl die Werkzeuge als auch die digitalen Forschungsdaten. Zudem zeigen sich wahrgenommene Risiken, die generell von Technologien im Kontext der Digital Humanities ausgehen. Zum einen liegen bisher kaum Erfahrungswerte vor, mit denen sich eine tatsächliche Relevanzbewertung von Informationsinfrastrukturen bzw. Publikationsszenarien vornehmen lässt. Zum anderen besteht die Gefahr, dass neue technische Dispositive bestimmte Forschungs- und Erkenntnispraxen begünstigen und dafür andere weniger angemessen berücksichtigen.

Bibliographie

Universitätsbibliothek der Humboldt-Universität zu Berlin (o. J.): *Fu-PusH*. Future Publications in the Humanities https://www2.hu-berlin.de/fupush/ [letzter Zugriff 06. Januar 2016].

Graphdatenbanken für Historiker mit Perspektiven für die Historische Semantik

Kuczera, Andreas
andreas.kuczera@geschichte.uni-giessen.de
Regesta Imperii Gießen/Mainz, Deutschland

Die zunehmende Mengen an Volltexten in den Geschichtswissenschaften und vor allem auch in der Mediävistik bietet neue Chancen fu#r die Forschung, erfordern aber auch neue Methoden und Sichtweisen. Der Beitrag möchte die Verwendung von Graphdatenbanken für die Speicherung von Erschließungsinformationen vorstellen.

Momentan werden digitale Quellen und die mit ihnen verbundenen Erschließungsinformationen meist in XML oder in SQL-Datenbanken abgelegt. XML hat sich als Standard bewa#hrt und findet in vielen Editionsprojekten als Datenformat Verwendung wa#hrend Datenbanken auf Websites meist auf SQL-Datenbanken als Daten-Repositories zuru#ckgreifen. XML-Dateien sind in der Regel bis zu einem gewissen Grade noch versta#ndlich lesbar, bei SQL-Datenbanken ist die Lesbarkeit ohne Kenntnis der zu Grunde liegenden Datenstrukturen in der Regel nicht mehr gegeben. Dies liegt nicht zuletzt auch an den Architekturen der Datenbanken: um optimale Performance zu erhalten werden die Datenstrukturen normalisiert. Hier kommt es für die optimalen Nutzungsmöglichkeiten entscheidend auf die Gestaltung des Frontends der Datenbank an. Oft sind die User-Interfaces jedoch vor allem auf die Bedürfnisse jener Personen ausgerichtet, die die Datenbank selbst erstellt haben. Da diese Personen in der Regel die Datenstrukturen tief durchdrungen haben, kann es bei der Gestaltung des Frontends leicht zu einseitigen Ausrichtung auf Experten-Nutzer kommen. Solche Nutzer wissen bereits vor der Suchanfrage wie ihr Ergebnis aussieht. In den Fachwissenschaften wird eine solche Anfrage als CIN-Anfrage bezeichnet (concrete information need). Davon zu unterscheiden sind POIN-

Anfragen (problem-oriented information need), bei denen der Nutzer ohne tiefere Kenntnisse des Datenmaterials und den zu Grunde liegenden Strukturen eine Anfrage startet (Vgl. hierzu Frants / Shapiro / Voiskunskii: 1997). Die Ausrichtung auf CIN-Anfragen zeigt sich auch in den größeren Quellenportalen zur Mediävistik (Vgl. Kuczera 2014). Hier ist die Verwendung von Graphdatenbanken ein alternativer Ansatz fu#r die Speicherung von erschließendem Wissen.
In SQL-Datenbanken sind die Informationen in Tabellen abgelegt, die untereinander verknu#pft sind. Graphdatenbanken folgen hier einem vo#llig anderen Ansatz. In einem Graph gibt es Knoten und Kanten. Vergleicht man die Knoten mit einem Eintrag in einer Tabelle einer SQL-Datenbank, wa#re eine Kante eine Verknu#fung zwischen zwei Tabelleneintra#gen. Im Unterschied zu SQL- Datenbanken ko#nnen Knoten und Kanten jeweils Eigenschaften haben.

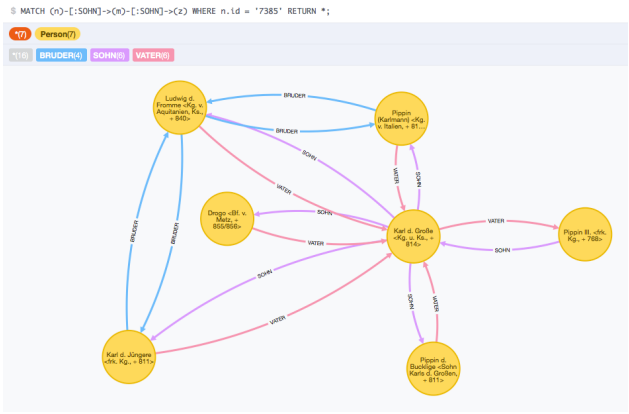


Abb. 1: Direkte Verwandschaftsverhältnisse Karls des Großen als Graph visualisiert

Daneben lassen sich die in Graphdatenbanken abgelegten Informationen sehr gut visualisieren. Gerade komplexere Datenbestände können hier sinnvoll für den Wissenschaftler erschlossen werden. Explorative Erschließungsmöglichkeiten erleichtern hierbei den Zugriff auf weitergehende Wissensdomänen des Repositoriums (Vgl. Kuczera 2015).

Das Datenmodell einer Graphdatenbank bildet quasi die semantische Repräsentation des in der Datenbank abgelegten Wissens. Ergänzt man die Eigenschaften der Knoten mit Identifikatoren wie den Angaben aus der GND oder legt man den Verknüpfungsstrukturen fachspezifische Ontologien zu Grunde können die Informationen der Graphdatenbank auch für automatisierte Abfragen über das Internet erschlossen werden.

In der Posterpräsentation werden in einem ersten Beispiel die Strukturen einer Graphdatenbank erläutert und anschließend mit der Graphenrepräsentation der Register der Regesten Kaiser Friedrichs III. und der genealogischen Datenbank Nomen-et-Gens Anwendungsbeispiele vorgestellt.

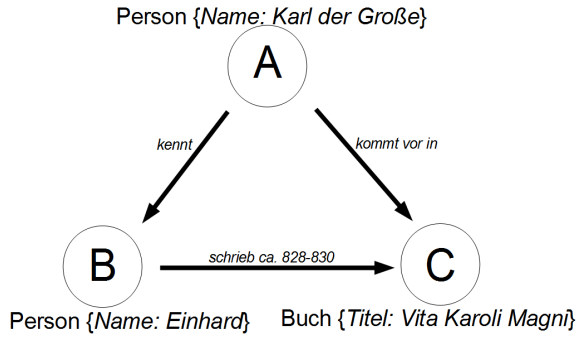


Abb. 2: Beispielgraph zu Karl dem Großen, Einhard und der Vita Karoli Magni

Bibliographie

Frants, Valery I. / Shapiro, Jacob / Voiskunskii, Vladimir G. (1997): *Automated information retrieval*. Theory and methods (= Library and information science). San Diego: Academic Press.

Kuczera, Andreas (2014): "Digitale Perspektiven mediävistischer Quellenrecherche", in: *mittelalter.hypotheses.org* http://mittelalter.hypotheses.org/3492 [letzter Zugriff 28. September 2015].

Kuczera, Andreas (2015): "Graphdatenbanken für Historiker. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi", in: *mittelalter.hypotheses.org* http://mittelalter.hypotheses.org/5995 [letzter Zugriff 28. September 2015].

CRETA (Centrum für reflektierte Textanalyse) – Fachübergreifende Methodenentwicklung in den Digital Humanities

Kuhn, Jonas
jonas.kuhn@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Alexiadou, Artemis
artemis@ifla.uni-stuttgart.de
Universität Stuttgart, Deutschland

Braun, Manuel
manuel.braun@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Ertl, Thomas
thomas.ertl@vis.uni-stuttgart.de
Universität Stuttgart, Deutschland

Holtz, Sabine
sabine.holtz@po.hi.uni-stuttgart.de
Universität Stuttgart, Deutschland

Kantner, Cathleen
cathleen.kantner@sowi.uni-stuttgart.de
Universität Stuttgart, Deutschland

Misselhorn, Catrin
catrin.misselhorn@philo.uni-stuttgart.de
Universität Stuttgart, Deutschland

Pado, Sebastian
pado@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Richter, Sandra
sandra.richter@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Stein, Achim
achim.stein@ling.uni-stuttgart.de
Universität Stuttgart, Deutschland

Zittel, Claus
claus.zittel@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Dieser Beitrag soll das Konzept des neu eingerichteten Stuttgarter DH-Zentrums CRETA [1] vorstellen, das mit literatur-, sprach-, geschichts- und politikwissenschaftlichen sowie philosophischen Fragestellungen sehr unterschiedliche textorientierte Fachdisziplinen vereint und das auf der anderen Seite Methoden und Modellierungstechniken aus dem maschinellen Lernen, der Computerlinguistik und aus der computergraphischen Visualisierung nicht nur zur Anwendung bringt, sondern begonnen hat, diese in eine gemeinsame DH-Methodik der tiefen reflektierten Textanalyse zu integrieren. Eine solche Weiterentwicklung des Methodeninventars der Digital Humanities ist ein langer Weg und braucht viele Beteiligte. Aspekte der Konzeption können wir jedoch bereits anhand von Fallstudien zu Szenarien aus laufenden Digital Humanities-Projekten konkret illustrieren, und es erscheint uns wichtig, den Ansatz breit zur Diskussion zu stellen.

Das methodische Konzept hinter CRETA geht einerseits aus von der strukturellen Gleichartigkeit vieler Teilfragestellungen über ganz unterschiedliche Teilgebiete der Digital Humanities hinweg (eingebettet in sehr unterschiedliche Gesamtzusammenhänge und methodische Rahmenbedingungen). Beispielsweise findet sich das Teilziel einer systematischen Kategorisierung von Relationen, die in einer Textquelle zwischen zwei realen oder fiktionalen Personen ausgedrückt ist, in geschichtswissenschaftlichen Fragestellungen ebenso wie in sprach-, literatur- oder sozialwissenschaftlichen Gesamtuntersuchungen. Abbildung 1 skizziert weitere Typen von wiederkehrenden Fragestellungen, die disziplinübergreifend bei der Auseinandersetzung mit Texten (und allgemeiner mit kulturellen Werken) auftauchen und bei deren Modellierung daher Synergien zu erwarten sind. Eine komputationelle Modellierung des Teilfrage-Typs kann so für ganz unterschiedliche Rahmenuntersuchung die Erschließung größerer Korpora per Aggregation über Aspekte des Textinhalts bzw. der –form erschließen.

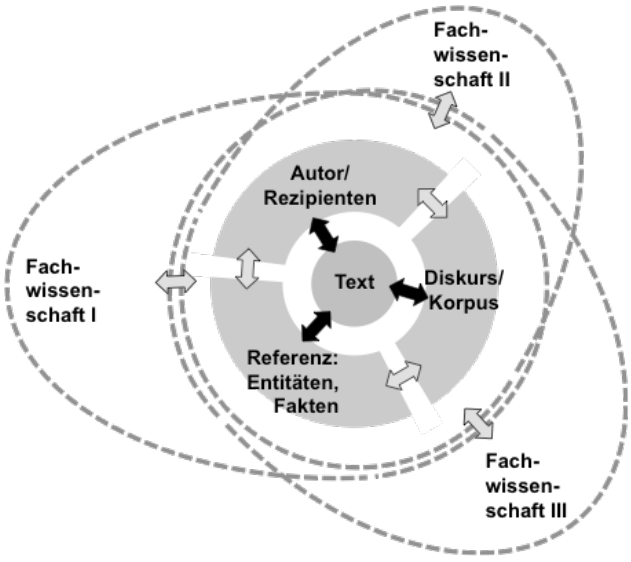


Abb. 1

Zugleich anerkennt das methodische Konzept die Unterschiedlichkeit sowohl der jeweiligen inneren Ausprägung der Fragestellung (so gehen im genannten Beispiel Texteigenschaften und Relationstypen weit auseinander) als auch der interpretatorischen Anforderungen, die sich aus dem jeweiligen Modellierungs- und Fragekontext ergeben. Das technische Ziel, eine einzige optimale Werkzeuglösung für jede der fachübergreifend identifizierten Teilfragen zu entwickeln bzw. aus dem *Text Mining* oder *Data Mining* zu übernehmen, greift also zu kurz. Für die Mehrzahl der Einsatzgebiete wäre die Lösung suboptimal, und bei der Anwendung wäre schwer zu unterscheiden, in

welchen Aspekten sie den methodischen Ansprüchen der einbettenden Untersuchungen genüge tut und in welchen nicht. Also sollten die gleichartigen Teilfragen zwar gemeinsam gedacht werden, als Instanzen derselben Modellklasse. Sie können (und müssen teilweise) jedoch für den jeweiligen Kontext angepasst und optimiert werden.

Der zentrale CRETA-Gedanke zur Erschließung von disziplinübergreifenden Synergien ist folgender: Für eine praktisch umsetzbare und dennoch methodisch adäquate Integration in die jeweilige Gesamtfragestellung kann es vorteilhaft sein, Modellinstanzen anfänglich auch über Kontexte hinweg zu übertragen, deren Randbedingungen nicht in vollem Maße übereinstimmen, die eingebetteten Teilmodelle aber sehr bewusst als vorläufig anzusehen – als Gegenstand eines **fortlaufenden Verbesserungsprozesses** (ganz im Sinne des Modellierungsbegriffs, den McCarty (2005) als Kernelement der Digital Humanities identifiziert). Die ohnehin angezeigte methodenkritische Hinterfragung des eingeschlagenen Weges (die jedoch gerade beim Einsatz von komputationellen Werkzeugen häufig nicht oder nicht in ausreichender Tiefe erfolgt) rückt damit zentral ins Blickfeld, und es ist nicht nur eine Frage des Nutzungskomforts, dass ein Instrumentarium zur Verfügung gestellt werden, das eine reflektierte Diagnostik der ineinandergreifenden komputationellen und klassischen Analyseschritte ermöglicht.

Folgerichtig wird bei aufgedeckten Unzulänglichkeiten die **Anpassung** der Modellierungslösung für eine Teilfrage an die kontextuellen Anforderungen zu einer Aufgabe, die nicht in der technisch-digitalen Peripherie einer geisteswissenschaftlichen Studie zu bearbeiten ist, sondern ihr kommt durch das Gesamtgeflecht aus Teilschritten für die übergeordnete Untersuchungsfrage zentrale Bedeutung zu.

Die notwendigen Anpassungen der vorläufigen Modellinstanzen lassen sich mit Techniken aus der Informatik (insbes. maschinellen Lernverfahren) prinzipiell ohne weiteres umsetzen – dabei muss jedoch die Zielrichtung der Optimierung vorgegeben sein (beim maschinellen Lernen in der Regel unterschiedliche Eingabe- / Ausgabe-„Trainingsdatensätze“, anhand derer die Parameter für eine gegebene Modellklasse eingestellt werden). Und hier beginnt die eigentliche Herausforderung für eine echte fachübergreifende Methodenintegration: selbst wenn man – rein hypothetisch – für eine geisteswissenschaftliche *Gesamt*fragestellung eine ausreichende Menge von Eingabe- / Ausgabe-„Daten“ bereitstellen könnte (also eine Annotation der interpretatorischen Zielkategorien für repräsentative Texte / Textabschnitte), die eine Modell-Optimierung ermöglichen würde, müsste dieser (vermutlich extrem ressourcenintensive) Prozess für jede Studie neu vorgenommen werden, da die gleiche Gesamtfragestellung in den Geisteswissenschaften wohl niemals zweimal gestellt wird. Methodische Erkenntnisse zu empirischen Faktoren bei der Modellierung aus einer Studie ließen sich nur schwer auf eine andere übertragen. (Abgesehen davon dürfte mit der Bereitstellung einer vollständig adäquaten Zielannotation häufig auch die Gesamtfragestellung gelöst sein, so dass der Bedarf an einer komputationellen Modellierung hinfällig wird.)

Naheliegender Weise wird man vielmehr versuchen, Modelle für relativ eng umrissene Teilfragestellungen empirisch zu optimieren, die dann in ein Geflecht von Analyseschritten einfließen. Der Annotationsaufwand für die Erzeugung von Referenzdaten hält sich damit in vertretbaren Grenzen und Erkenntnisse zu studienübergreifend gleichartigen Teilaspekten lassen sich so systematisch übertragen.

Der Identifikation von sinnvollen Teilfragestellungen, die über unterschiedliche Projekt- und Fachkontexte hinweg tragen – einer „Modularisierung“ – kommt also auch aus praktischen Erwägungen heraus eine zentrale Bedeutung zu. Was aus informatischer Sicht wie eine Binsenweisheit klingt, ist jedoch in der Modellierungspraxis extrem anspruchsvoll, ist bei vielen übergeordneten Fragen eine Untergliederung in effektive Teilschritte doch alles andere als klar. Eine Vorstrukturierung auf dem Reißbrett ist nur in Einzelfällen möglich (wie im Fall der Sprachwissenschaft mit ihrer etablierten Ebenenstruktur der Sprachbetrachtung möglich ist, die auch die computerlinguistische Modulstruktur prägt, selbst wenn bewusst klassische Teilschritte kombiniert werden).

Für alle offenen Fragen der Modularisierung bietet die komputationelle Modellierung und die Verwendung von digitalen Arbeitsumgebungen Potenziale, die noch lange nicht ausgeschöpft sind: alternative Modularisierungen können exploriert und gegeneinander abgewogen werden. Der CRETA-Ansatz legt diese Exploration in die interdisziplinären Verantwortung: statt auf dem Reißbrett die plausibelste Untergliederung einer Projekt-Problematik festzuhalten, Softwarelösungen anhand dieser Spezifikation umzusetzen und nach zwei Jahren Entwicklung auf die inhaltliche Fragestellung anzuwenden, findet ein Dialog zwischen komputationellen Modellierungsexpertinnen und –experten und Fachwissenschaftlerinnen und –wissenschaftlern unterschiedlicher Disziplinen statt.

Überlegungen aus der fachspezifischen Kultur der Fragestellung müssen herangezogen werden, um eine geeignete Einbindung eines technisch übertragbaren Teilmodells in den Erkenntnisprozess und seine methodenkritische Reflexion zu gewährleisten. Gleichzeitig fleißen aus den informatischen Disziplinen Überlegungen zur formalen Adäquatheit möglicher Modellklassen, Erfahrungswerte aus der zu erwartenden Qualität, sowie Möglichkeiten einer Visualisierung und explorativen Ergebnispräsentation ein, um die wechselseitige Optimierung von Modellierungskomponenten zu unterstützen.

Konkret stellt sich das Vorgehen bei der Modellierung folgendermaßen dar: Im multidisziplinären Dialog im

Rahmen von Werkstattklausuren werden für geistes- und sozialwissenschaftliche Fragestellungen mit Bezug zu ausgewählten digitalen Ressourcensammlungen

- für vergleichbare Teilaufgaben eines bestimmten Typs die charakteristischen Parameter so definiert, dass sowohl die argumentative Funktion der Teilaufgabe innerhalb des fachwissenschaftlichen Vorgehens als auch das Spektrum der formal-komputationellen Implementierungen im Wesentlichen aus diesen Parametern heraus abgeleitet werden kann,
- für jeden gängigen Typ von Teilaufgaben ein Instrumentarium von Methoden und Werkzeugen zur Verfügung gestellt für die Evaluation der Qualität bei der Aufgabenbearbeitung, für das Auffinden möglicher Fehler, Aggregation und Meta-Analyse von Ergebnissen (jeweils in Kombination von analytischen Werkzeugen und interaktiver Ergebnisvisualisierung),
- der interaktive Prozess einer Korrektur und Anpassung von Komponenten sowie die Kombination von Ergebnissen mit Hilfe stark visuell orientierter Interfaces unterstützt.

Die angesprochenen methodischen Desiderate eines transparenten Zugangs zu den Analyse-Teilergebnissen und der Adaptierbarkeit von analytischen Teilmodellen haben wir exemplarisch anhand mehrerer Erweiterungsszenarien des Relationsextraktionsmodells aus Blessing und Kuhn (2014) umgesetzt: Über die ursprüngliche Zielrelation (Emigrationsbewegungen, die in Kurzbiographien textuell beschrieben werden) können andere Relationen interaktiv trainiert werden. Eine Erweiterung des Korpusbestands um Texte aus weiteren Quellen wurde vorgenommen, einschließlich eines Wechsels der Sprache (Übertragung des Teilmodells als Erweiterung einer deutschen Analysekette auf eine französische). Eine analoge Adaptionsplattform wurde für Zeitungstexte erstellt, die in politikwissenschaftlichen Studien zum öffentlichen Diskurs analysiert werden.

Fallstudien zum Einsatz der resultierenden Analysekette zeigen, dass eine kritische Betrachtung der übertragenen Teilmodelle vor allem durch den Wechsel des Blickwinkels auf aggregierte Daten mit einer Verlinkung von Einzelinstanzen unterstützt werden: Textuelle Einzelinstanzen eines Relationstyps (z. B. Emigration einer Person X aus dem Land A in ein Land B) werden aggregiert und das Aggregationsergebnis kann beispielsweise geographisch visualisiert werden.

Das interaktive Springen zwischen unterschiedlichen Dimensionen der Aggregation bzw. zwischen aggregierter Sicht und Einzelinstanzen erlaubt es, Datenpunkte gezielt unter die Lupe zu nehmen, die von allgemeinen Tendenzen in bestimmter Weise abweichen. Fu#r solche Beobachtungen ist zu klären, ob es sich (a) um einen aus bekannten Zusammenhängen erklärbaren, (b) einen neuartigen, validen Effekt oder (c) um einen technisch erklärbaren Scheineffekt handelt, der durch eine methodische Verbesserung eliminiert werden könnte. Ein Beispiel ist die fehlerhafte Klassifikation von UN-Resolution 1261 und 1973 in Zeitungsartikeln als Datumsangaben. Bei der Visualisierung der Extraktionsergebnisse auf einem Zeitstrahl fällt ein unerwartetes Muster beim Jahr 1261 auf (während Scheineffekte zum Jahr 1973 möglicherweise zunächst unerkannt bleiben). Die Fallstudien unterstützen die These, dass interaktive Nachforschungen und ein adaptierbares Instrumentarium gerade bei nicht perfekten Analysekomponenten die kritische Distanz zum Modellinventar unterstützen.

Notes

1. Das Zentrum wird seit Januar 2016 vom BMBF gefördert. Dieser Beitrag fasst wichtige Aspekte des Antragskonzepts zusammen; die Liste der Autorinnen und Autoren entspricht den Antragstellern für die BMBF-Förderung. Durch ihren Wechsel an die Humboldt-Universität zu Berlin ist Artemis Alexiadou nicht mehr direkt in die Umsetzung involviert.

Bibliographie

Blessing, André / Kuhn, Jonas (2014): "Textual Emigration Analysis (TEA)", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland.

McCarty, Willard (2005): *Humanities Computing*. London: Palgrave.

Darf wissenschaftliches Design in DH-Projekten emotional ansprechen?

Lambertz, Michael
lambertz@uni-trier.de
Universität Trier (Trier Center for Digital Humanities), Deutschland

Wie müssen virtuelle, integrative und interaktive Forschungs-, Kommunikations- und Präsentationsumgebungen aussehen, damit sie der Umsetzung des fächerübergreifenden Forschungsparadigmas nützen, die Schaffung und Darstellung von Wissen ermöglichen und eine aktive Beteiligung der Öffentlichkeit an diesen Prozessen erlauben, lautet eine der zentralen Fragen nicht nur der DHd 2016, sondern generell der Digital Humanities (DH). Zu selten wird dabei bedacht, dass Visualisierung

mehr ist als nur die reine grafische Darstellung und Gestaltung geisteswissenschaftlicher Inhalte. Die Rolle, die das Design eines User Interface (UI) spielt, ist in den DH noch zu wenig im Forschungsfokus. Die Entscheidung, welches Design und welche UI-Komponenten letztlich gewählt und umgesetzt werden, ist leider selten so wissenschaftlich fundiert wie der Inhalt, sondern stattdessen oft das Ergebnis eines Kompromisses der beteiligten Personen und deren persönlichem Geschmack, deren Vorlieben und Gewohnheiten. Daraus resultierende Mängel und Probleme reichen von Unsicherheit und Bedienungsproblemen bis zu Abneigung gegenüber dem Design und damit dem Projekt, und zeigen sich wenn überhaupt erst später bei der Nutzung der Benutzeroberfläche. Ein Grund für diese mangelnde wissenschaftliche Fokussierung liegt sicherlich häufig in der finanziellen Unterausstattung dieses Bereichs in DH-Projekten, stehen doch den digitalen Geisteswissenschaften nicht die Gelder[1] für Design, geschweige denn für die Erforschung der „visuellen Kommunikation" zur Verfügung, wie sie etwa im (Neuro-)Marketing für Usability-Tests, Eye-Tracking- oder bildgebende Verfahren vorhanden sind. Und doch könnte mehr Wissen über Design- und Usability-Fragen auch hier gebraucht werden.[2]

Mehr noch als die funktionellen Gesichtspunkte von (UI-)Design sind es die emotionalen Aspekte von Design, die bislang nicht nur in der DH-Forschung noch unterrepräsentiert sind. Es gibt noch immer wenig Wissen darüber, wie im Gehirn Gefühle vernetzt sind und wirken, und über das Zusammenspiel von Emotionen und Gedanken können Jahrtausende alte Philosophien und Weisheiten oft mehr sagen als die Forschung auf dem Gebiet der Neurobiologie[3] . Dennoch ist ziemlich offensichtlich, dass es ein Zusammenspiel gibt[4] , und dass diesbezüglich das Verhältnis von Emotionen und wissenschaftlichem Design einer näheren Betrachtung unterzogen werden muss. Grafikdesign wirkt immer auch unterhalb der Oberfläche von Sachlichkeit, Rationalität und Kontrolle. Designelemente wie Bilder (z. B. weinendes Kind, Bergpanorama), Typographie (z. B. Romantik, Bauhausstil) und Farben (z. B. schwarz/rot/weiß der NS-Zeit) können je nach „Zielgruppe" (Targeting) mehr oder weniger starke Emotionen wie z. B. Mitleid, Glück, Sicherheit, Freude, Freiheit, Stolz, Unmut, Scham, ja sogar Wut hervorrufen.

Dies scheint konträr zum Rationalitäts- und Objektivitätsanspruch der Wissenschaft zu sein. Vermeintlich wissenschaftlich neutral gestaltete Designs nutzen Schriftarten und Farben, die in anderen Zielgruppen eventuell sogar Irritationen hervorrufen können. Je nach Mode und Zeit variieren diese Designelemente zwar, aber im Allgemeinen strahlt das typisch sachliche, wissenschaftliche Design eine gewisse Kühle und Geschmeidigkeit aus. Eine emotionale Wirkung ist also auch hier vorhanden.

Zentrale Fragen, die sich gerade die DH mit ihren noch nicht ausgeschöpften gestalterischen Möglichkeiten stellen müssen, sind daher: *Inwieweit darf wissenschaftliches Design ästhetisch sein und gezielt emotional ansprechen? Wie korreliert es mit der vermeintlichen wissenschaftlichen Objektivität? Kann und wenn ja wie kann emotionale Wirkung beobachtet werden? Kann / darf sie gelenkt werden? Welche Kriterien sind wesentlich für Designentscheidungen in den DH?* Auf einer Metaebene sollte dabei auch exploriert und reflektiert werden, wie bewusst sich DH-Projekte für oder gegen ein Design- und UI-Konzept entscheiden.

Diesen Fragen möchte sich die Posterpräsentation auch auf der Ebene des Visuellen und des Designs stellen und anhand des Heinrich Heine-Portals exemplifizieren. Für die am Trier Center for Digital Humanities aufgebaute und zwischenzeitlich „in die Jahre gekommene" digitale Edition werden ein neues Designkonzept und Visualisierungskomponenten entworfen und realisiert, anhand deren Aspekte der emotionalen Ansprache von Präsentationen sowie Fragen, die sich daraus für den ästhetischen Anteil wissenschaftlicher Darstellungsformen diskutiert werden sollen.

Notes

1. Neben finanziellen und projektlaufzeitgebundenen Gründen führen die zusätzlichen Schwierigkeiten, die sich aus der parallelen Entwicklung von (UI-)Design zu der fachwissenschaftlichen Entwicklung der Inhalte und der Modellierung von Daten bzw. deren Architektur ergeben, auch oft zu pragmatischen und wenig adäquaten Lösungen.
2. Dennoch lassen sich Forschungsergebnisse hinsichtlich der Kaufentscheidungen von Zielgruppen in der Marktforschung und im Neuromarketing nicht eins zu eins auf die Wissensdarstellung übertragen, da die Ziele hier größtenteils andere sind.
3. Erst seit den 1990er Jahren gibt es bildgebende Verfahren in den Neurowissenschaften, die nachweisen konnten, dass Gefühle im Gehirn repräsentiert werden. Trotz einiger Fortschritte in der Forschung ist bezüglich Komplexität und Dynamik von Gefühlen wenig bekannt (vgl. die Kritik von Kagan 2007).
4. Das ist vermutlich der Grund, warum große Unternehmen so viel Geld in Neuromarketing investieren.

Bibliographie

Frevert, Ute (2009): "Was haben Gefühle in der Geschichte zu suchen?", in: *Geschichte und Gesellschaft* 35: 183-208.

Kagan, Jerome (2007): *What is Emotion?* History, Measures, and Meanings. Binghamton: Yale University Press.

McCandless, David (2009): *Information is Beautiful*. London: Collins.

Meirelles, Isabel (2013): *Design for Information*. An introduction to the histories, theories, and best practices behind effective information visualizations. Beverly Massachusetts: Rockport.

Pferdt, Frederik G. (2012): *Designbasierte Didaktik (DbD)*. Lernumgebungen mit social media innovativ gestalten. Paderborn: Eusl-Verlag.

Ruecker, Stan / Radzikowska, Milena / Sinclair, Stéfan (2011): *Visual interface design for digital cultural heritage*. A guide to rich-prospect browsing. Farnham: Ashgate.

Gernika – Visualisierung der Interkonnektivität medialer Öffentlichkeiten in Europa

Loebel, Jens-Martin
loebel@bitgilde.de
Angewandte Medienwissenschaft (Digitale Medien), Universität Bayreuth / bitGilde IT Solutions UG

Holly, Eva Maria
eva.holly@uni-duesseldorf.de
Lehrstuhl für Neuere Geschichte, Geschichtswissenschaften, Heinrich-Heine-Universität Düsseldorf

Einleitung

Gernika-Lumo ist eine Stadt im spanischen Baskenland und zeichnet sich seit jeher durch seine zentrale kulturhistorische Rolle für die Basken aus. Bis in die Gegenwart trägt Gernika internationale Bedeutung, nicht zuletzt aufgrund Picassos Antikriegsbild „Guernica“, das er als Reaktion auf die Zerstörung Gernikas am 26. April 1937 durch den Luftangriff der deutschen Legion Condor und italienischer Kampffliegerverbände während des spanischen Bürgerkrieges malte. Die Bombardierung unbefestigter Städte aus der Luft sowie die gezielte Terrorisierung der Zivilbevölkerung stellte zu diesem Zeitpunkt ein Novum dar und entfachte eine heftige öffentliche Debatte in den europäischen Tageszeitungen. Davon ausgehend entstand über die Zeit hinweg bis in die Gegenwart ein komplexes intermediales Geflecht aus wechselseitigen Bezugnahmen und vielseitigen Vernetzungen. Durch den Einsatz verschiedener Medien wurden Kommunikationsräume aufgespannt, die nationale und transnationale Öffentlichkeiten hervorbrachten.

Den Kern des Forschungsvorhabens bilden die Untersuchung und Visualisierung dieser Kommunikationsstrukturen ausgehend von der Berichterstattung ausgewählter Tageszeitungen der Länder Deutschland, Frankreich, Großbritannien und der Schweiz, um daraus dezidierte Aussagen über die Struktur von Öffentlichkeiten ableiten zu können.

Zwar existieren Arbeiten über Öffentlichkeiten im Kontext des spanischen Bürgerkriegs (Brinkmann 2002) und Presseanalysen zu Gernika (Southworth 1977), allerdings wurden hier andere, zumeist inhaltliche Fragestellungen wie beispielsweise nach den Verantwortlichen, verfolgt. Eine umfassende und präzise Analyse der Beziehungsgeflechte erscheint jedoch erst durch die Zuhilfenahme digitaler Forschungsplattformen möglich, die u. a. eine tiefergehende semantische Analyse der Beziehungen und Anbindung an externe Linked-Open-Data-Ressourcen (LOD) ermöglichen.

Beschreibung der Forschungsumgebung

Die Forschungsgrundlage besteht aus einer gescannten Sammlung der Presseberichterstattung überregionaler Tageszeitungen ausgewählter Länder über die Zerstörung Gernikas, die in die virtuelle Forschungsumgebung *HyperImage* importiert werden. Zeitgleich wird eine erste Codierung der Zeitungsartikel mittels SPSS vorgenommen. Zusammen mit dem derzeit in Entwicklung befindlichen Nachfolgesystem *Yenda* zur semantischen Erschließung wird das Korpus katalogisiert und semantische Bezüge zu Unterthemen (Topics, wie z. B. der Schuldfrage) hergestellt. In einem nächsten Schritt kann das komplexe Geflecht der unterschiedlichen semantischen Bezugnahmen zwischen den Zeitungsartikeln dargestellt werden (z. B. Berliner Tageblatt bezieht sich auf einen Artikel der Times über Gernika, der sich wiederum auf die Neue Zürcher Zeitung bezieht). Außerdem wird angezeigt zu welchen Topics Bezüge vorliegen.

Konkret erlaubt die Anwendung den Abruf aller Artikel zu Gernika (intern und in LOD-Beständen) nach einem bestimmten Topic, die Darstellung der unterschiedlichen Bezugnahmen zu diesem Topic und die zeitliche Gewichtung wann Bezugnahmen am intensivsten (a) generell; (b) zu bestimmten Topics und (c) zwischen welchen Ländern vorliegen. Zudem können übergeordnete Themen (z. B. die Rolle Europas, vgl. Kaelble 2002), die in allen Topics vertreten sein könnten, als horizontal verlaufende Thematik visualisiert werden.

Yenda erlaubt so die Verfolgung von Diskurskontinuitäten sowie die Modellierung und Visualisierung eines gemeinsamen Bezugsnetzwerkes über die semantische Verknüpfung von Topics. Querbezüge zu anderen Diskursen und Medien werden erfasst und in einer gemeinsamen Umgebung dargestellt.

Die Vernetzung erfolgt zum Einen automatisiert über Schlagworte und RDF-Terme und zum Anderen über das manuelle Einpflegen weiterer Referenzen im Zusammenhang mit der SPSS-Codierung der Quelldaten.

HyperImage und Yenda sind Open-Source-Forschungsumgebungen. HyperImage ist als Werkzeug zur Unterstützung des Bilddiskurses in den Digitalen Geisteswissenschaften seit vielen Jahren etabliert und wird in Forschung und Lehre an Forschungseinrichtungen in Deutschland und Europa eingesetzt. Zwischenergebnisse wie endgültige Fassungen können jederzeit als hypermediale online- oder offline-Publikation erstellt werden und sind über standardisierte APIs (wie IIIF) abfrag- und nachnutzbar (Loebel et al. 2014).

Yenda bietet als neues Werkzeug die Möglichkeit zur semantischen Annotation und Analyse von Mixed-Media-Daten auf Basis von RDF sowie des Open Annotation Data Models (OADM) und ist die konsequente Weiterführung des HyperImage-Gedankens.

Ergebnisse, Zeitrahmen, Online-Veröffentlichung

Für die historische Forschung ergeben sich durch die digitalen Werkzeuge neue Möglichkeiten, komplexe und weitrechende Vernetzungen in vielen Bereichen aufzuzeigen und diese konkret zu visualisieren. Insbesondere mit Blick auf synchrone und diachrone Vergleiche erlaubt dies die Bildung neuer Theorien und Hypothesen.

So lässt sich beispielsweise ein Vergleich mit Ereignissen in anderen Epochen, wie z. B. Gernika in der Zeit vor 1945 und europäischer Integration mit der Diskurslage um die Jugoslawienkriege der 1990er Jahre aufstellen, der eine tiefergehende Analyse von gleichsam verwendeten Diskursthemen erlaubt.

Kontinuitäten in diachronen Vergleichen der Berichterstattung können erkannt und die Veränderung kommunikativer Verdichtungen in Bezugnahmen über die Zeit visualisiert werden. Das semantische Netz (RDF-Graph) lässt sich mit Informationen in weiteren Medien und LOD-Ressourcen erweitern. Spannend ist dies auch z. B. für Fragen im Rahmen der Europäischen Integration (Meyer 2010) für Disziplinen wie Geschichts-, Kommunikations-, Politik- und Sprachwissenschaften.

Das Poster wird erste Zwischenergebnisse der Forschungsarbeit sowie das Werkzeug Yenda en Detail vorstellen. Die Veröffentlichung der Ergebnisse ist für 2017 geplant. Die Forschungsumgebung Yenda wird ab voraussichtlich Frühjahr 2016 unter http://yenda.tools/ als öffentliche Betaversion zur Verfügung stehen.

Bibliographie

Brinkmann, Sören (2002): "Bilder eines Krieges: Europa und der Bürgerkrieg in Spanien", in: Requate, Jörg / Schulze Wessel, Martin (eds.): *Europäische Öffentlichkeit*. Transnationale Kommunikation seit dem 18. Jahrhundert. Frankfurt am Main: Campus Verlag.

HyperImage (o. J.): http://hyperimage.ws [letzter Zugriff 15. Februar 2016].

Kaelble, Hartmut (2002): "Das europäische Selbstverständnis und die europäische Öffentlichkeit im 19. und 20. Jahrhundert", in: Kaelble, Hartmut / Kirsch, Martin / Schmidt-Gernig, Alexander (eds.): *Transnationale Öffentlichkeiten und Identitäten im 20. Jahrhundert*. Frankfurt am Main: Campus Verlag 85-110.

Loebel, Jens-Martin / Kuper, Heinz-Günter / Arnold, Matthias / Decker, Eric (2014): "Hachiman Digital Handscrolls – Semantische Anreicherung mit HyperImage und Yenda", in: *EVA Konferenz Berlin 2014*. Elektronische Medien & Kunst, Kultur und Historie, Berlin: Staatliche Museen Preußischer Kulturbesitz 262-267.

Meyer, Jan-Henrik (2010): *The European Public Sphere*. Media and Transnational Communication in European Integration 1969–1991. Stuttgart: Franz Steiner Verlag.

Southworth, Herbert R. (1977): *Guernica! Guernica!* Berkeley / Los Angeles / London: University of California Press.

Datenressourcen der Arbeitsstelle des Deutschen Wörterbuchs (Neubearbeitung)

Mederake, Nathalie
nmedera@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Blanck, Wiebke
Wiebke.Blanck@gmx.net
Akademie der Wissenschaften zu Göttingen, Deutschland

Zwischen den Digital Humanities und der Lexikografie besteht eine interessante Wechselwirkung: Zum einen gehört die Erschließung textueller Informationsquellen zur alltäglichen Wörterbucharbeit und ist dort fester Bestandteil des lexikografischen Redaktionsprozesses. Verschiedene digitale Instrumente und Korpora unterstützen diese Arbeit, indem sie besagte Quellen auffindbar machen, beispielsweise in Bibliotheken oder als Digitalisate im Internet. Zum anderen wird über digitale Wörterbücher der Zugang zu umfangreichen (digitalen und nicht-digitalen) Textkorpora

und elektronischen Datensammlungen eröffnet, die für Wissenschaftler verschiedenster Disziplinen von Bedeutung sind. Digital vorliegende Wörterbücher eröffnen damit eine wichtige Schnittstelle zwischen den traditionellen Geisteswissenschaften und den Digital Humanities. Diese Schnittstelle sowie die lexikografisch aufbereitete Information als solche wird angesichts der Tatsache, dass Wörterbücher der Zukunft Informationssysteme sein werden, die aufgrund sehr großer Sprachdatenbanken existieren, nichts an ihrer Bedeutung einbüßen.

In der Neubearbeitung des Deutschen Wörterbuchs (= ²DWB), einem (teilweise digitalisierten) traditionellen geisteswissenschaftlichen Unternehmen, sind im Laufe der Zeit zwei Nebenprojekte entstanden, die sich an eben dieser Schnittstelle befinden und lexikografische bzw. lexikologisch relevante Informationen erschließen: das Quellenverzeichnis zum Deutschen Wörterbuch und die Kartei Literatur zur Wortforschung. Die digitale Aufarbeitung und Bereitstellung dieser beiden Projekte wurde in den vergangenen Jahren in Zusammenarbeit mit der Staats- und Universitätsbibliothek Göttingen (SUB) entwickelt und steht in beiden Fällen kurz vor dem Abschluss.

Das Quellenverzeichnis der Neubearbeitung des Deutschen Wörterbuchs ist ein unerlässliches Hilfsmittel für jeden, der mit diesem Wörterbuch arbeitet. Ziel der digitalen Erschließung des Quellenverzeichnisses ist es, die vorhandenen elektronischen Daten in die noch zu erarbeitende digitale Version der Neubearbeitung des Deutschen Wörterbuchs zu integrieren, um Nutzern dieses Wörterbuchs einen möglichst barrierefreien Zugriff auf die Quellenlage zu gestatten. Überdies bieten sich über die Quelleneinträge Verlinkungsmöglichkeiten der Wörterbuchartikel an und damit ein größerer Umfang der Wörterbuchnutzung.

Die Kartei Literatur zur Wortforschung, die sogenannte LW-Kartei, war ursprünglich ein internes Hilfsmittel im ²DWB, die den Bearbeitern die Artikelarbeit erleichtern sollte, indem sie auf Forschungsliteratur zu Einzelwörtern verwies. Diese Arbeit wurde in den 1970er-Jahren begonnen. Im Laufe der Jahre ist ein Zettelkatalog entstanden, der ca. 14.000 Einträge umfasst. Er enthält wissenschaftliche Literatur zu Stichwörtern von A-Z und wurde in zwei Sortierungen angelegt (sowohl nach Verfassern als auch nach Einzelwörtern). Seit 2011 wird die Kartei Literatur zur Wortforschung kontinuierlich auf den aktuellen Forschungsstand gehoben. Dazu werden entsprechende Informationen aus der Forschungsliteratur von 1990 bis heute exzerpiert; d. h. aus dem Zeitraum, den der Zettelkatalog nicht mehr erfasst. Aufgrund der Fülle an Veröffentlichungen ist dies ein umfangreiches Vorhaben, das in der verbleibenden Laufzeit des ²DWB (Ende 2016) nicht vollständig zu bewältigen sein wird. Sollte die LW-Kartei jedoch innerhalb eines anderen wissenschaftlichen Projekts weitergeführt werden, ist es wichtig, bereits jetzt an die Grundlagen für eine weiterführende Bearbeitung zu denken, d. h. den Aufbau und die Bearbeitung (d. i. Exzerption) eines Korpus' zur Einzelwortforschung zu diskutieren. Bis Ende 2016 werden alle bis dahin entstandenen Exzerpte in der Datenbank verfügbar gemacht.

Durch die Projektkooperation mit der SUB Göttingen ist es möglich geworden, zwei geprüfte lexikografische Datensammlungen, die parallel zur ständig wachsenden und elektronisch verfügbaren Wörterbuchlandschaft entstehen, angemessen und zugleich verständlich aufzuarbeiten und zu präsentieren. Sie bilden eine sinnvolle Ergänzung sowohl zum Wörterbuchprodukt als auch zur Arbeit mit Wörterbuchergebnissen. Das erarbeitete digitale Format ist zweifelsohne ein unerlässlicher Beitrag zur weiteren Erschließung textueller und lexikologischer Quellen; es soll anderen Wörterbuchprojekten Anknüpfungsmöglichkeiten über die ermittelten Literaturressourcen eröffnen und das Potenzial lexikografischer Arbeit über den unmittelbaren Wörterbuchkontext hinaus vermitteln.

Die beiden Sammlungen, die als Nebenprodukte der lexikografischen Arbeit entstanden sind, beschreiben ferner Arbeiten, die für weitere Erschließungsmethoden der digitalen und historischen Lexikografie von grundlegender Bedeutung sind. Eine Integration dieser Ressourcen in bestehende Wörterbuchnetze ist zudem zwingend erforderlich. Auch wenn für das ²DWB der Kern der Wörterbucharbeit bis zum Ende der Projektlaufzeit eher traditionellen lexikografischen Maßstäben verpflichtet ist, sind die mit dieser Arbeit in Verbindung stehenden Entwicklungen bezeichnend für die sich verändernden wissenschaftlichen Mittel und formulieren neue Anknüpfungspunkte lexikografischer Fragestellungen, die zu diskutieren sind.

Nutzerorientierte Softwareentwicklung revised – Die Perspektive der Editorinnen und Editoren in digitalen Musik und Medieneditionen

Meise, Bianca
bianca.meise@upb.de
Universität Paderborn, Deutschland

Schloots, Franziska
franzi.margarete@gmail.com
Universität Paderborn, Deutschland

Meister, Dorothee
dm@ex.uni-paderborn.de
Universität Paderborn, Deutschland

Müller-Lietzkow, Jörg
jml@mail.upb.de
Universität Paderborn, Deutschland

Erkenntnisse der Digital Humanities, so die Hoffnung vieler Wissenschaftler, führen das Wissen aus den Elfenbeintürmen der Universitäten heraus (vgl. etwa Lauer 2013). Viele Bildungspotenziale, wie die Demokratisierung von Wissensbeständen im Sinne von Zugang, Partizipation, Reflexion und Emanzipation, lassen sich den Digital Humanities problemlos zuschreiben. Auch wenn die Digitalisierung, Auszeichnung und Verarbeitung digitaler Artefakte bislang im Zentrum der Digital Humanities stehen, gilt es gleichzeitig, die Perspektive der Nutzerinnen und Nutzer zu erschließen. Letztendlich entstehen Wissen und Bildung nicht singulär, sondern in sozialen Kontexten, indem Menschen auf Basis ihres bislang erarbeiteten Wissenstandes Ressourcen nutzen, sich mit ihnen auseinandersetzen und diese transformieren. Wie aber werden Technologien innerhalb der Digital Humanities gegenwärtig genutzt, welche wissenschaftlichen Potenziale eröffnen sich und welche Restriktionen bestehen? Damit gilt es die Wissenschaftlerin und den Wissenschaftler in den Blick zu nehmen (vgl. auch Edwards 2012) und vom forensic zum formal layer (vgl. Kirschenbaum 2008) zu wechseln. Aber auch Kirschenbaums formal layer bringt nicht ganz zum Ausdruck, was Drucker (2013) mit der performativen Ebene von Materialität beschreibt: Handeln, der Umgang der Nutzer mit kulturellen, aber auch immateriellen Artefakten, prägt die Wahrnehmung, Beurteilung und kulturelle Bedeutung dieser Artefakte mit. Digitale Editionen, Softwareentwickler und Editorinnen und Editoren befinden sich neuerdings in einem komplexen reziproken Verhältnis und entwickeln derzeit durch ihre Arbeit andere Repräsentationen und damit ein neues Verständnis von Editionen.

In Bezug auf digitale Musik- und Medieneditionen gilt es festzuhalten, dass es DEN Anwender gar nicht gibt. Aufgrund der Komplexität der Editionsanwendungen sind ganz verschiedene Aspekte wissenschaftlicher Tätigkeiten und Anwendungsmöglichkeiten damit verbunden. Diese gilt es differenziert zu betrachten, sollen mögliche Potenziale letztlich nicht nur illusorischer Natur sein oder ungenutzt bleiben. Im Rahmen des Projekts „Zentrum Musik-Edition- Medien. Musik und nicht-textuelle Objekte im Kontext digitaler Editionen“ befasst sich ein Teilprojekt damit, die Perspektive der Editorinnen und Editoren mittels medien- und sozialwissenschaftlicher Befragungen zu erheben. Ziel ist es zum einen, die veränderte wissenschaftliche Editionspraxis zu untersuchen und zum anderen, die empirischen Untersuchungsergebnisse in die Optimierung der Editionssoftware einfließen zu lassen. Die Editorinnen und Editoren arbeiten an der Schnittstelle vom computer und cultural layer (Manovich 2001), d. h. sie arbeiten mit Metadaten und Auszeichnungssprachen und müssen somit die Logiken des Prozessierens des Computers verstehen. Gleichzeitig arbeiten sie mit den Transformationen an der Oberfläche, lassen sich Teile oder Überblicke bestimmter Werkaspekte anzeigen, um editorische Entscheidungen zu treffen und bilden damit einen ganz versierten Wissenschaftstypus ab. Darüber hinaus gibt es weitere Anwenderinnen und Anwender wie etwa Tonmeister, Dozenten und Studenten, deren Perspektiven noch erhoben werden sollen.

Zunächst wurde von Juli bis August 2015 sowohl eine qualitative als auch eine quantitative Erhebung bei Editorinnen und Editoren durchgeführt. Bei der qualitativen Studie wurden acht Interviews von eineinhalb bis dreieinhalb Stunden Dauer geführt. Bei der quantitativen Befragung wurden bislang 60 Editorinnen und Editoren befragt. Ziel der quantitativen Erhebung ist es, möglichst viele Editorinnen und Editoren unabhängig von der Art der bearbeiteten Editionen zu ihrer wissenschaftlichen Arbeit, Kommunikation und Softwarenutzung zu befragen. Mit der qualitativen Befragung hingegen wurden nur Editorinnen und Editoren interviewt, die mit der Editionssoftware Edirom arbeiten. Hier standen die persönlichen Erfahrungen, Erwartungen, Arbeitsweisen und Orientierungen der Wissenschaftler im Zentrum.

Die quantitative Erhebung zeigt, dass durch die überwiegend projektbezogene Forschung neue Anforderungen an Werkzeuge gestellt werden, welche das vernetzte, kooperative und kollaborative Arbeiten unterstützen. Dabei deutet sich an, dass sich die über viele Jahrzehnte bewährten Arbeitsabläufe in den Geisteswissenschaften im Zuge der Digitalisierung grundlegend verändert haben. So zeigt sich bereits ein stetiger Prozess, in welchem sich die neuen Werkzeuge in die Arbeitsabläufe der Editorinnen und Editoren einfügen. Die Befragung bestätigt zudem Warwicks (2012) These, dass Geisteswissenschaftler eben keine Technikverweigerer sind, sondern einfach spezifische Anforderungen und Arbeitsabläufe formulieren, was zu den in den DH diskutierten und zu bearbeitenden Herausforderungen für die Entwickler von Werkzeugen und Software führt.

In den qualitativen Interviews wurde deutlich, dass mit der Editionssoftware ganz neue Arbeitskontexte und Repräsentationsmöglichkeiten entstehen, die die bisherige Editionspraxis unterstützen, erweitern und den Forscherinnen und Forschern andere Perspektiven ermöglichen. Darüber hinaus ergaben sich daraus spannende Einblicke in bisherige Karriere- und Ausbildungswege der Editorinnen und Editoren, die nicht nur durch musikwissenschaftliche Expertise vorweisen müssen. Hinsichtlich der nutzerorientierten

Softwareentwicklung bieten die Interviews sehr gute Hinweise auf sich etablierende Handlungsroutinen, die in der Software abgebildet werden müssen. Ebenso gibt es Verweise auf erhebliche Arbeitserleichterung durch die Softwarenutzung und notwendige Weiterentwicklungen, um bislang nicht verfügbare Repräsentationsmöglichkeiten einzubeziehen und innovative Fragestellungen bearbeiten zu können. Aus dem qualitativen Material lässt sich bereits jetzt ablesen, dass durch die digitalen Werkzeuge und Repräsentationsmöglichkeiten Editionen als solche als Gegenstand akademischer Auseinandersetzungen interessant werden: Welche Möglichkeiten der (Re-) Präsentation gibt es, welche Informationen gehören dazu, welche Analysemöglichkeiten entwickeln sich durch die digitale Sammlungen, wann ist ein Endpunkt der wissenschaftlichen Erkenntnis erreicht, oder wird lediglich ein Kontinuum der Wissensproduktion dokumentiert? In diesem Sinne zeigt sich, dass die empirische Nutzerforschung zum einen sowohl die Softwareentwicklung bereichert, als auch neue Impulse für aktuelle theoretische Diskurse in den Digital Humanities anbietet.

Bibliographie

Drucker, Johanna (2013): "Performative Materiality and Theoretical Approaches to Interface", in: *Digital Humanities Quartely* 7,1 http://www.Digitalhumanities.org/dhq/vol/7/1/000143/000143.html [letzter Zugriff 03. September 2015].

Edwards, Charlie (2012): "The Digital Humanities and Its Users", in: Gold, Matthew K. (ed.): *Debates in the Humanities* http://dhdebates.gc.cuny.edu/debates/text/31 [letzter Zugriff 03. September 2015].

Kirschenbaum, Matthew (2008): *Mechanisms*. New Media and the Forensic Imagination. Cambridge: MIT University Press.

Lauer, Gerhard (2013): "Die digitale Vermessung der Kultur. Geisteswissenschaften als Digital Humanities", in: Geiselberger, Heinrich / Moorstedt, Tobias (eds.): *Big Data*. Das neue Versprechen der Allwissenheit. Berlin: Suhrkamp.

Manovich, Lev (2001): *Language of New Media*. Cambridge: MIT Press.

Warwick, Claire (2012): "Studying users in digital humanities, in: Warwick, Claire / Terras, Melissa / Nyhan, Julianne (eds.): *Digital Humanities in Practice*. London: Facet Publishing.

WissKI – Wissenschaftliche Kommunikations-Infrastruktur

Merz, Dorian
dorian.merz@fau.de
Friedrich-Alexander Universität Erlangen-Nürnberg, Deutschland

Fichtner, Mark
m.fichtner@wiss-ki.eu
Germanisches Nationalmuseum Nürnberg, Deutschland

Das DFG-finanzierte Projekt “WissKI” hat in den Jahren 2009 bis 2011 zur Entwicklung einer digitalen Forschungsumgebung für die Anwendung im Bereich der Digital Humanities geführt. Hauptaspekt der Datenerfassung und -haltung in WissKI sind die semantischen Zusammenhänge zwischen einzelnen Fakten und Datensätzen. Dies wird durch umfassende Unterstützung aktueller Semantic Web Technologien erreicht. Die Einordnung und Speicherung der erhobenen Daten erfolgt auf Grundlage einer Domänenontologie, deren Konzepte und Relationen - zu sogenannten Pfaden verbunden - als Vorlage für die Masken und Felder im System dienen. Auf Basis dieser Technologie werden solitär erscheinende Daten zu einem gemeinsamen, semantischen Netzwerk verbunden und damit die unmittelbare Sichtbarkeit weiterer, tiefergehender Zusammenhänge ermöglicht. Hierdurch bietet sich ein Mehrwert, der in der Vergangenheit in flachen Hierarchien wie Datenbanktabellen gar nicht oder nur mit sehr hohem Aufwand erkennbar gemacht werden konnte. Das Web-basierte Systemdesign und der dadurch ermöglichte Zugriff über das Internet, die Anbindung von externen kuratierten Datenquellen (sog. Authority Files) und die Möglichkeit zur Bereitstellung ausgewählter Daten über gängige Online-Schnittstellen (Web-Frontend, SPARQL-Endpoint, ...) betonen den Semantic-Web-Gedanken hinter der Infrastruktur.

Die Speicherung der Daten erfolgt in einem Triple-Store, der die eingegebenen Fakten in einer Subjekt-Prädikat-Objekt-Satzform ablegt. Die Aneinanderreihung der hier verwendeten Prädikate zu Pfaden erfolgt im Kern des Systems, dem sogenannten Pathbuilder, mit dem die semantische Bedeutung der einzelnen Masken-Einträge in Bezug auf das beschriebene Objekt (auch Person, Ort o. Ä.) anhand der Ontologie festgelegt wird. Die Eingabe der Daten erfolgt über eine, mit den gängigen Datenbankoberflächen vergleichbare, Editier-Oberfläche. Sie ist aus Feldern aufgebaut, die wiederum je einem bestimmten Feldtyp zugeordnet sind. Feldtypen bestimmen die Ein- und Ausgabemodalitäten der Daten.

So stehen zum Beispiel ein- oder mehrzeilige Textfelder. verschiedenartige Auswahldialoge und Möglichkeiten zur Bildanzeige zur Verfügung. In einigen davon kann WissKI den Anwender durch Auto-Vervollständigung unterstützen, indem zu angefangenen Eingaben ähnlich lautende Einträge aus der lokalen Datenhaltung oder den Authority Files vorgeschlagen werden.

Weitere umfassende Möglichkeiten zur Datenaufbereitung sind das halbautomatische Auszeichnen und Erkennen bzw. Einordnen von Personen, Orten, Zeiten und anderen Entitäten aus Texten (sog. Named Entity Recognition), die Verwaltung der Revisionsgeschichte von Texten, das Werkzeug zur Bildannotation und die Verknüpfung mit der Literaturverwaltungssoftware Zotero. Dabei verzichtet die Software nicht auf die aus dem Bereich der Content Management Systeme bekannten Funktionalitäten wie z. B. die Generierung von Websites, Foren, Wikis oder auch die detaillierte Verwaltung der Nutzer und ihrer Zugriffsrechte.

Inzwischen ist die Software in verschiedenen Forschungsprojekten an unterschiedlichen, namhaften Institutionen im kunst- und kulturhistorischen, sowie biologischen und technischen Bereich erfolgreich im Einsatz. Als Domänenontologie im Museums- und Sammlungsbetrieb kommen individuelle Erweiterungen des “Conceptual Reference Model” des Comité international pour la documentation zum Einsatz (CIDOC-CRM: ISO 21127), dessen Umsetzung in der Web-Ontology-Language OWL ebenfalls vom Projekt besorgt wurde und über die Website http://erlangen-crm.org frei zur Verfügung steht.

Das Poster stellt nun die konsequente Weiterführung des Projektes (DFG Proj.-Nr. GR1471/9) und das damit einhergehende Update der Software vor. Neben der grundlegenden Aktualisierung der zugrundeliegenen Frameworks und Technologien (Drupal 7, php 5.5, SPARQL 1.1) sind dabei einige Erweiterungen der Oberflächenfunktionalität sowie deutliche Erleichterungen in der Bedienbarkeit vorgesehen. So können nun beliebige Feldtypen aus der “Field API” des verwendeten Content Management Systems Drupal in die Ein- und Ausgabeansichten integriert werden. Dies umfasst neben den altbewährten Textfeldern und -bereichen und Bildern (incl. Zoomviewer für sehr hochauflösende Bilder) auch interaktive Landkarten, 3D-Animationen, Zeitstrahlen und alle denkbaren Medientypen, sowohl zur direkten Ansicht als auch zum Download. Zusätzlich zu diesen bereits eingebundenen Formaten ermöglicht die offene Architektur von WissKI² auch die Einbindung anderer, gängiger Feldtypmodule, die für Drupal zur Verfügung stehen.

Zu den erwähnten Erleichterungen zählt ebenso ein Update des System-Kerns, dem Pathbuilder, mit dem die Pfadschablonen durch die Domänenontologie auf einer graphischen Oberfläche ausgewählt bzw. erzeugt werden können. Daneben soll eine umfassende Bibliothek mit Musterontologien, -masken und -pfaden bereitgestellt werden, die die Einstiegshürde für Erstbenutzer minimal halten wird.

Ein weiteres neu integriertes Werkzeug zur Verbesserung der Datenqualität werden automatische Inferenzmaschinen, sog. Reasoner, sein, die in der Lage sind, widersprüchliche Daten zu erkennen und somit etwaige Eingabe- oder Modellierungsfehler aufzudecken.

: aichinger

Mueller, Mathias
mathiasmueller@aon.at
ÖAW, Österreich

Dittrich, Andreas
dittricha@gmail.com
ÖAW, Österreich

Waltl, Gilbert
Gilbert_Waltl@gmx.at
ÖAW, Österreich

Csillag, Marlene
marlene@software-multimedia.at
ÖAW, Österreich

Godler, Katharina
katharina.godler@oeaw.ac.at
ÖAW, Österreich

Ivanovic, Christine
christine.ivanovic@univie.ac.at
ÖAW, Österreich

Ziel: Erfassung und Analyse literarischer Topographien

Der Fokus der Untersuchung liegt auf der literarischen Repräsentation von Raum. Bisherige Untersuchungen ihres Werks haben erwiesen, dass Aichingers Bezugnahmen auf Orte und Ereignisse in Wien zentrale Bedeutung zukommt (Fässler 2013). Dabei fällt auf, dass Aichinger Raumbezüge in verschiedenen Phasen ihres Werks auf ganz unterschiedliche Weise elaboriert: Der Wienbezug ihres ersten Romans, *Die größere Hoffnung* (1948), ist für den Leser / die Leserin unzweifelhaft erkennbar, obwohl Aichinger konsequent auf die Nennung identifizierbarer Ortsnamen verzichtet. Im mittleren Werk werden Ortsbezüge zunehmend abstrakt; in ihren spätesten Texten hingegen häufen sich exakte Ortsangaben im Stadtraum Wien.

Ziel des Projekts ist es, grundlegende Strukturen in Aichingers Referenzierung auf Orte zu ermitteln und deren Zusammenhang zur historischen Erfahrung herauszuarbeiten, deren Darstellung im Zentrum ihres Werks steht. Zu diesem Zweck sollen alle Angaben zu Ort, Zeit und Person in ihren Texten so codiert werden, dass sie einer maschinellen Abfrage zugänglich und damit sowohl systematisch als auch vollständig evaluiert werden können.

Arbeitsschritte

Digitale Texterfassung

Textgrundlage ist die achtbändige Ausgabe der Werke Ilse Aichingers (S. Fischer Verlag 1991) sowie die danach erschienenen Einzelbände. Diese Bände wurden gescannt und mittels OCR (Optical Character Recognition) erfasst und dadurch maschinenlesbar gemacht. Als Vergleichskorpora sollen zusätzlich die davon abweichenden Textfassungen der Erstausgabe des Romans sowie der zwischen 2000 und 2004 in Tageszeitungen publizierten Texte erfasst werden.

Digitale Texterschließung

Im zweiten Arbeitsschritt wird eine TEI-konforme Datei erstellt, in der die Texte mithilfe von Standards wie RDF (Resource Description Framework), XML (Extensible Markup Language) und PoS (Part-of-Speech-Tagging) codiert und dadurch der maschinellen Abfrage durch Abfragesprachen wie SPARQL (SPARQL Protocol And RDF Query Language) zugänglich gemacht werden. Im Hinblick auf den primären Fokus der Untersuchung, die Erfassung und Analyse der literarischen Topographien Aichingers, werden vorrangig Personennamen sowie Orts- und Zeitangaben kodiert. Außerdem ist eine Analyse anhand semantischer Felder geplant, wofür eine Vernetzung mit unterschiedlichen Datenbanken (z. B. Dornseiff) vorgesehen ist. Von vornherein soll so gearbeitet werden, dass die Möglichkeit weitere bzw. speziellere Codierungen zu ergänzen offen bleibt.

Erhebung und Einpflege zusätzlicher Daten / Metadaten

Ergänzend zur digitalen Erfassung und Erschließung der Texte werden weitere Metadaten eingebracht. Dies können textgenetisch relevante Daten sein wie Entstehungs- und Publikationsdaten, oder Sacherläuterungen, wie sie in Apparaten wissenschaftlicher Editionen oder Kommentaren üblich sind, sowie Hinweise auf Varianten, Querverweise, Illustrationen etc. Ein Teil dieser Daten ist durch Recherchen in Wiener Archiven oder am Aichinger-Vorlass im Deutschen Literaturarchiv (DLA) in Marbach zu erheben. Die Auszeichnung durch RDF ermöglicht aber auch die Verlinkung mit online Datenbanken und damit den Anschluss an das semantic web (Hitzler et al. 2008; Ivanovic / Frank 2015).

Darstellung der Ergebnisse

Das Textkorpus :aichinger soll die Basis bilden für die Durchführung von Abfragen und Analysen, die eine präzise, systematische und vollständige Evaluierung der Raumbezugnahmen im Gesamtwerk der Autorin ermöglichen soll. Erfassbar werden dadurch beispielsweise Personenkonstellationen in Verbindung mit Orten sowie Frequenzen der Nennung bestimmter Orte resp. Wege in Korrelation zur beschriebenen Zeit wie zur Zeit der Textabfassung. Diese Ergebnisse verlangen unterschiedliche Darstellungsformen. So sind Diagramme möglich, oder Wordclouds, die wiederum Häufungen oder Übereinstimmungen bzw. Korrelationen darstellen können. Auf der Basis der RDF Codierung lassen sich z. B. maschinell Karten generieren, in denen die erwähnten Orte oder Wege der in den Texten genannten Figuren u. a. aufscheinen. Die kartographische Darstellung ermöglicht es darüber hinaus Leerstellen ihres Werkes (nie genannte Orte oder Zonen) oder verdeckte Strukturen (die Wientopographie, die der für Aichinger maßgebliche Film Der Dritte Mann darstellt) sichtbar zu machen. Insbesondere anhand solcher Karten wird das Poster das Konzept und die Analysemöglichkeiten unseres Projektes darstellen.

Prototyp

Das Projekt dient der Sichtbarmachung und besseren Analyse der raumrelevanten Strukturen im Werk Aichingers und deren Relevanz für die erinnerungskulturell motivierte Schreibweise, der die Autorin verpflichtet ist. Das Projekt hat insofern paradigmatischen Charakter, als die an diesem Beispiel entwickelten Methoden den Status eines Prototyps haben und auch bei der Analyse anderer Textkorpora Anwendung finden sollen.

Bibliographie

Aichinger, Ilse (1991): *Werke in acht Bänden*. Herausgegeben von Richard Reichensperger (Die größere Hoffnung / Der Gefesselte / Eliza Eliza / Schlechte Wörter / Kleist, Moos, Fasane / Auckland / Zu keiner Stunde / Verschenkter Rat). Frankfurt / Main: S. Fischer.

Fässler, Simone (2011): *Von Wien her, auf Wien hin*. Ilse Aichingers „Geographie der eigenen Existenz.“ Wien: Böhlau.

Hitzler, Pascal / Krötzsch, Markus / Rudolph, Sebastian (2008): *Semantic Web*. Berlin / Heidelberg: Springer.

Ivanovic, Christine / Frank, Andrew (2015): „Auf der Suche nach dem erfüllten Raum: Digitale Korpusanalyse in der Literaturwissenschaft am Beispiel Ilse Aichinger“, in: *Tagung der Digital Humanities im deutschsprachigen Raum*, Graz.

Historische Begriffe der Erziehungswissenschaft - Erzeugung einer Ontologie

Müller, Lars
l.mueller@dipf.de
DIPF, Deutschland

Historische Begriffe der entstehenden Erziehungswissenschaft sollen als maschinenlesbare Terminologie für digitale historische Bildungsforschung bereitgestellt werden. Bibliografische Titelaufnahmen von Lemmata aus 24 historischen erziehungswissenschaftlichen Nachschlagewerken (1774 – 1942) werden hierfür in eine Ontologie transformiert (Abbildung 1). Die Lexikonbeiträge dokumentieren Herausbildung und Wandel erziehungswissenschaftlicher Fachsprache und Gegenstände. Christian Ritzi (2003: 114) zeigte anhand der „Prügelstrafe“, wie diese "in pädagogischen Nachschlagewerken als pädagogisches Problem im Sinne einer 'Kodifizierung des Wissens' in das System des ausgebreiteten pädagogischen Wissens integriert" wurde. Das Beispiel illustriert das Potential, das für die Forschung in einem integrierten, semantisch modellierten, historischen Vokabular liegt. Im Fachportal Scripta Paedagogica Online (SPO) (vgl. BBF) sind die digitalisierten und detailliert mit Metadaten beschriebenen Lexika (Ritzi 2003: 103ff.) bereits frei zugänglich.

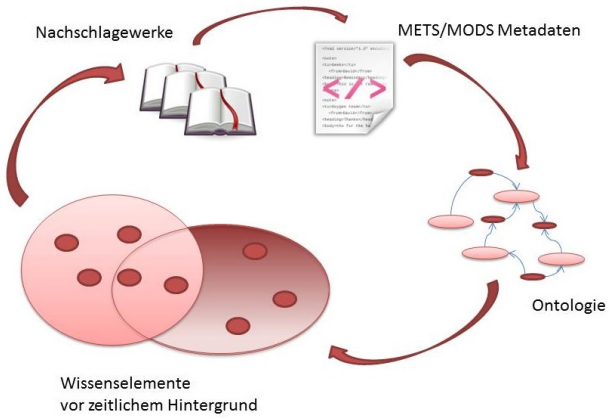


Abb. 1: Rekonstruktion einer Wissensdomäne aus bibliografischen Daten

Ell et al. (2013) haben gezeigt, wie Bibliografische Daten zur Bildungsgeschichte in die virtuelle Forschungsumgebung Semantic CorA (vgl. German Institute for International Educational Research et al.) zur semantischen Repräsentation und Analyse integriert werden können. Die digitalen Dienste der Bibliothek für Bildungshistorische Forschung (BBF) können künftig über die Bereitstellung von Daten hinausgehen und gegenüber den bestehenden Funktionen um eine Wissens-Ebene erweitert werden (vgl. Oramas et al. 2014; Feng et al. 2005), um die Bildungsgeschichte bei ihrer Hinwendung zu digitalen Methoden optimal zu unterstützen.

Im hier dargestellten Ansatz werden bibliografische Titeldaten als Begriffe bzw. Entitäten konzeptualisiert, neu modelliert und semantisch angereichert. So können sie in bildungshistorischen Analysen ausgewertet oder als Datensets für Textanalysen verwendet werden. Diese zu Forschungsdaten transformierten Katalogdaten erweitern somit die Informationsinfrastruktur für die historische Bildungsforschung auf der Wissens-Ebene und stellen exemplarisch einen neuen bibliothekarischen Infrastrukturservice für Digital Humanities dar, der in der Erzeugung und Bereitstellung semantischer Vokabulare besteht.

Als Ausgangsmaterial liegen in SPO frei zugängliche METS / MODS-Metadatensätze digitalisierter Nachschlagewerke und Lemmata im RDF / XML-Format vor. Mittels XSLT können sie auf die für das Vorhaben relevanten Elemente reduziert und umgeformt werden. Die als Wissenselemente betrachteten Titeldaten werden dabei aus der Wissensdomäne Buch- und Bibliothekswesen in die Wissensdomäne historische Bildungsforschung überführt. Die Ontologie wird als Linked Open Data (LOD) in RDF aufgebaut.

Jedes Lemma wird im ersten Verarbeitungsschritt zum Deskriptor eines Konzepts, das durch den zugehörigen Lexikonbeitrag definiert wird. Das Vokabular wird in SKOS (vgl. W3C 2012) modelliert, indem die Instanzen umklassifiziert werden. In den Metadaten

befindliche Identifier für Katalogeinträge und digitale Dokumente werden mitgenommen und erhalten die Beziehung zu den Quellen. Im folgenden Schritt werden die erzeugten Vokabulare so überarbeitet, dass explizite Begriffsbeziehungen innerhalb jedes einzelnen Nachschlagewerks abgebildet werden. Zunächst wird jedes Nachschlagewerk einzeln transformiert. Anschließend werden die so entstandenen historischen Vokabulare zur Pädagogik durch Terminologiemapping (vgl. Keil 2012) aufeinander bezogen.

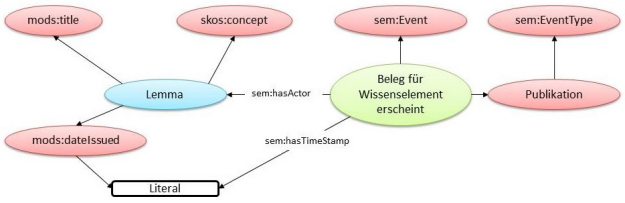


Abb. 2: Schematische Darstellung des Transformationsmodells

Eine besondere Herausforderung stellen Erhalt und Darstellung des zeitlichen Wandels von Sprachgebrauch und Kategorien dar. Für die Abbildung des Zeitbezugs wird jedes Nachschlagewerk als Snapshot-Ontologie modelliert (vgl. Ide / Woolner 2007: 142; Kauppinen / Hyvönen 2007). Jeder Begriff wird mit dem Erscheinungsdatum des ihn beinhaltenden Nachschlagewerks assoziiert, welches als „Publikationsereignis“ im Simple Event Model (SEM) (van Hage et al. 2011) modelliert wird (Abbildung 2). Da die aus den Lexika extrahierten Lemmata viele Überschneidungen aufweisen werden, sollen sie abschließend auch in einer einzigen Ontologie zusammengeführt werden (vgl. Kalfoglou / Schorlemmer 2003: 4).

Im Ergebnis entsteht eine Ontologie zu historischen Begriffen der Erziehungswissenschaft, die zur Nachnutzung in einem geeigneten Forschungsdatenrepositorium publiziert wird. Damit die Zuverlässigkeit der Daten gewährleistet und überprüft werden kann, werden auch Datensätze der Teilergebnisse zusammen mit den Metadaten des Transformationsprozesses dokumentiert. Damit bleibt jeder Verarbeitungsschritt replizierbar (Abbildung 3). Aus der neuen Datenbasis lassen sich Versionen ableiten, die bspw. als Gazetteer für automatische Entitätenerkennung, Entity Linking, Netzwerkanalysen oder visuelle Datenexploration eingesetzt werden können.

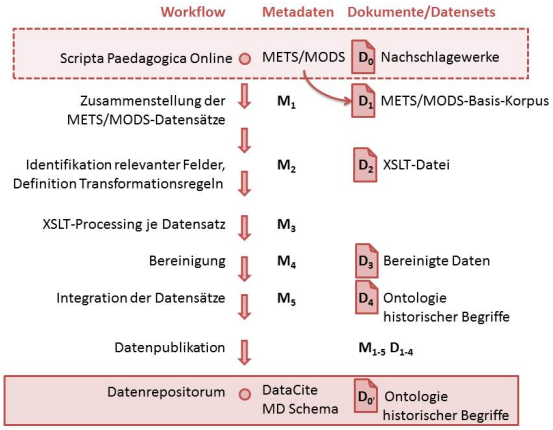


Abb. 3: Workflow zur Transformation und Erzeugung von Daten- und Metadatensets

Bibliographie

Bibliothek für Bildungsgeschichtliche Forschung (BBF) (o.J.): *SPO*. Scripta Pedagogica Online http://goobiweb.bbf.dipf.de/viewer [letzter Zugriff 12. Februar 2016].

Ell, Basil / Schindler, Christoph / Rittberger, Marc (2013): "Semantically Enhanced Interactions between Heterogeneous Data Life-Cycles", in: Garoufallou, Emmanouel / Greenberg, Jane (eds.): *Metadata and Semantics Research*. Cham: Springer International Publishing 277–288.

Feng, Ling / Jeusfeld, Manfred A. / Hoppenbrouwers, Jeroen (2005): "Beyond information searching and browsing: acquiring knowledge from digital libraries", in: *Information Processing & Management* 41, 1: 97–120 http://conceptbase.sourceforge.net/mjf/itrs008.pdf [letzter Zugriff 12. Februar 2016].

German Institute for International Educational Research / Karlsruher Institute for Technology / Georg-August-Universität Göttingen (o. J.): *Semantic CorA* [letzter Zugriff 12. Februar 2016].

Ide, Nancy / Woolner, David (2007): "Historical Ontologies", in: Ahmad, Khurshid / Brewster, Christopher / Stevenson, Mark (eds.): *Words and Intelligence II*. Dordrecht: Springer Netherlands 137–152.

Kalfoglou, Yannis / Schorlemmer, Marco (2003): "Ontology mapping: the state of the art", in: *Knowledge Engineering Review*18,1: 1–31 http://drops.dagstuhl.de/volltexte/2005/40/pdf/04391.KalfoglouYannis.Paper.40.pdf [letzter Zugriff 12. Februar 2016].

Kauppinen, Tomi / Hyvönen, Eero (2007): "Modeling and Reasoning About Changes in Ontology Time Series", in: Sharman, Raj / Kishore, Rajiv / Ramesh, Ram (eds.): *Ontologies*. A handbook of Principles,

Concepts ans Applications in Information Systems. Boston, MA: Springer 319–338.

Keil, Stefan (2012): "Terminologie Mapping: Grundlagen und aktuelle Normungsvorhaben", in: *Information - Wissenschaft & Praxis* 63, 1: 45-55 http://eprints.rclis.org/16716/1/Stefan%20Keil%20-TM%20Grundlagen%20und%20Normungsvorhaben-Repository.pdf [letzter Zugriff 12. Februar 2016].

Oramas, Sergio / Sordo, Mohamed / Serra, Xavier (2014): "Automatic creation of knowledge graphs from digital musical document libraries", in: *9th Conference on interdisciplinary musicology - CIM14*. Proceedings http://mtg.upf.edu/system/files/publications/CIM14_MAIN.pdf [letzter Zugriff 12. Februar 2016].

Ritzi, Christian (2003): "Scripta Paedagogica Online : Digitales Textarchiv zur Bildungsgeschichte des deutschsprachigen Raums", in: Thaller, Manfred (ed.): *Digitale Bausteine für die geisteswissenschaftliche Forschung*. Göttingen: Dührkohp & Radicke 103–135.

van Hage, Willem Robert / Malaisé, Véronique / Segers, Roxane / Hollink, Laura / Schreiber, Guus (2011): "Design and use of the Simple Event Model (SEM)", in: *Web Semantics: Science, Services and Agents on the World Wide Web* 9, 2: 128–136.

W3C = World Wide Web Consortium (2012): *SKOS*. Simple Knowledge Organization System [letzter Zugriff 12. Februar 2016].

neonion - Kollaboratives Annotieren zur Erschließung von textuellen Quellen

Müller-Birn, Claudia
clmb@inf.fu-berlin.de
Freie Universität Berlin, Deutschland

Breitenfeld, Andre
andre.breitenfeld@fu-berlin.de
Freie Universität Berlin, Deutschland

Im Rahmen einer Posterpräsentation stellen wir die kollaborative Annotationssoftware neonion vor, dessen Entwicklung inspiriert wurde von der Vision des Memex. Vannevar Bush führt in seinem Artikel dazu aus, dass "[a] record if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted." (Bush 1945: 37). Ein solcher "record" kann beispielsweise ein historisches Dokument sein. Am Anfang des Forschungsprozesses, ist noch wenig darüber bekannt, aber das Wissen um dieses Dokument wächst kontinuierlich durch die Forschungsarbeit der Wissenschaftler_innen. Mit Hilfe von *neonion* sollen Forschende Dokumente gemeinschaftlich mit Hilfe von Annotationen erschließen können (Müller-Birn et al. 2015). Die Software wird mit dem Ziel entwickelt, als Forschungsumgebung zu fungieren, um kollaborative Wissensgenierung im Rahmen von Annotationsprozessen in der Forschungsarbeit zu analysieren.

Die grundsätzlichen funktionalen Anforderungen an die Software neonion wurde basierend auf den Erkenntnissen einer Interviewstudie durchgeführt. Ingesamt wurden sechs Interviews durchgeführt. Diese Interviews waren semi-strukturiert und wurden am Arbeitsplatz durchgeführt, um auch Informationen über das Arbeitsumfeld zu erlangen und einen direkten Einblick in die genutzten Softwarewerkzeuge und Abläufe zu erhalten. Die Interviews dauerten eine bis anderhalb Stunden und wurden anschließend transkribiert. Alle Teilnehmer_innen waren Mitglieder der gleichen Forschungsreinrichtung (für weitere Informationen s. Müller-Birn et al. 2015). Das Ziel dieser Studie war es, den Kontext der Forschungsarbeit in den Geisteswissenschaften besser zu verstehen. Solche Interviews sind zentral bei der nutzerzentrierten Softwareentwicklung, einem Ansatz, der bei uns konsequent verfolgt wird. Die Ergebnisse der Interviews wurden nach vier Gesichtspunkten ausgewertet: Form, Funktion, Wert und Status (in Anlehnung an Marshall 1998). Wir nutzen diese Ergebnisse im Folgenden, um die grundsätzlichen Funktionen von neonion vorzustellen.

Der Bereich Form setzt sich mit der Struktur von Annotationen auseinander. Die Mehrzahl der Befragten gab an, vor allem basierend auf Textdokumenten zu forschen. Daher wurde entschieden, neonion zunächst für Textdokumente zu verwenden. Die Softwarearchitektur wurde mit Webframeworks umgesetzt. Hierzu findet Django im Bereich des Back-Ends und vorrangig Bootstrap und AngularJS im Front-End Anwendung. Die erzeugten Annotationen werden zur dauerhaften Aufbewahrung zum einen in den AnnotationStore, der auf Grundlage von ElasticSearch arbeitet, gespeichert, zum andern in den Sesame Triple Store eingespeist.

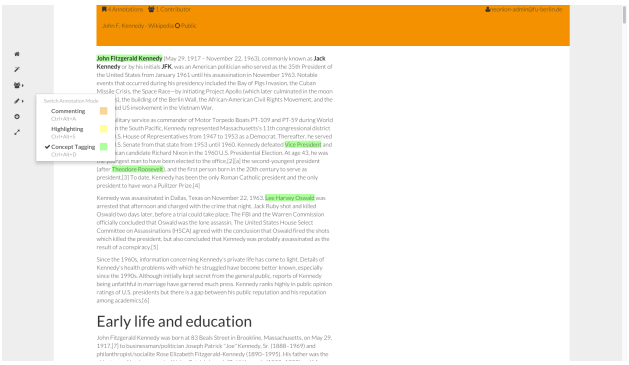


Abb. 1: Auswahl des Annotationsmodus im Annotator

Im Bereich Funktion wurde der Frage der Verwendung nachgegangen. So wurde ersichtlich, dass unterschiedliche Annotationsmodi (s. Abbildung 1) notwendig sind. In neonion werden daher drei Arten der Annotation

unterschieden: die Markierung für Zitate, der Kommentar für Paraphrase und semantische Tags für das semantischen Erschließen von Dokumenten (z. B. basierend auf einer zugrundeliegenden Ontologie). Auch wenn diese drei Arten von Annotationen aus den Interviews entstanden sind, können diese Annotationsmodi je nach Anwendungszweck sehr variabel eingesetzt werden. Ein Einsatz von neonion im Bereich der Linguistik wäre beispielsweise möglich, aber ein praktischer Anwendungsfall fehlt bisher.

Zur Implementierung der Annotationskomponente kommt die quelloffene JavaScript Bibliothek Annotator.js zum Einsatz. Die Bibliothek der OKFN wurde zusätzlich durch eigene Plug-Ins, insbesondere zur Realisierung einer semantischen Annotation, erweitert.

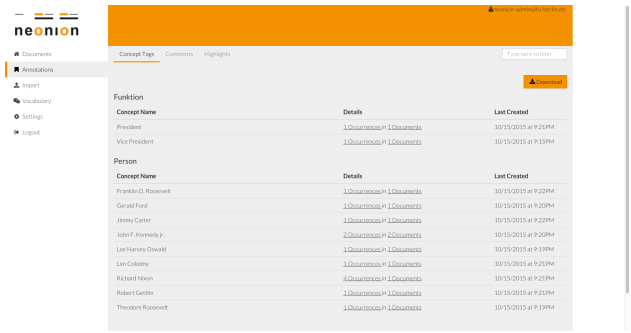


Abb. 2: Annotationsübersicht gefiltert nach semantischen Tags

Der längerfristige Wert der Annotation wurde im dritten Bereich untersucht. Hier wurde von allen Interviewteilnehmer_innen angegeben, dass eine Weiterverwendung der Annotationen in anderen Kontexten nicht möglich ist, da die verwendete Software den Export der Annotationen verhinderte. Diese Mangel sollte in neonion behoben werden. Alle Annotation werden einerseits innerhalb eines standardisierten Datenmodells – dem Open Annotation Data Models (OADM) - gespeichert und anderseits haben Nutzende die Möglichkeit, alle ihre Annotationen nach unterschiedlichen Gesichtspunkten zu filtern und in ein Textdokument zur Weiterverarbeitung, z. B. in ein Textverarbeitungsprogramm zu exportieren (s. Abbildung 2). Mit Hinblick auf die Kollaboration besteht die Möglichkeit Annotationen innerhalb von Gruppen mit anderen Nutzer_innen zu teilen bzw. gemeinschaftlich Dokumente zu annotieren. Ebenfalls können die verwendeten strukturierten Vokabulare gemeinschaftlich erstellt werden. Es ist geplant, hier entsprechend benötigte Diskussionsfunktionen einzubauen.

Darüber hinaus wurde das bestehende Open Annotation-Datenmodell um die Möglichkeit erweitert, semantische Tags über eine typisierte Verbindung in Relationen zueinander zu setzen. Die semantischen Tags stellen in diesem Zusammenhang Instanzen von vordefinierten Konzepten mit eigener URI (Unified Resource Identifier) dar und ermöglichen durch die Beziehung der Instanzen zueinander eine mehrstufige Analyse von Annotationen.

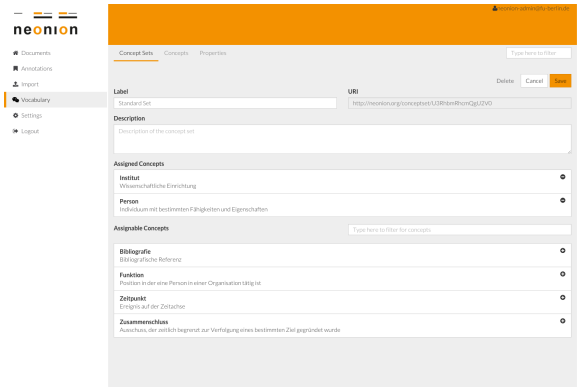


Abb. 3: Defintion eines Begriffsystems als Concept Sets

Im vierten Bereich wurde der Frage nachgegangen, wie Annotationen inhaltlich geteilt werden. Unsere Interviewpartner führten aus, dass vor allem im Bereich der strukturierten Annotationen (semantische Tags basieren auf einem vordefinierten Begriffssystem) es sehr umständlich und zeitaufwändig ist, ein gemeinschaftliches Begriffssystem zu erstellen. In neonion können solche Begriffssysteme, die zu einer Ontologie weiterentwickelt werden können, einfach als sogenannte Concept Sets (s. Abbildung 3) hinterlegt werden. Diese Concept Sets können dann auch wieder anderen Personen in neuen Forschungskontexten zur Verfügung gestellt werden. Aus technischer Sicht bietet das Backend von neonion verschiedene Dienste an, um Annotationen und Concept Sets mit unterschiedlichen Systemen über eine spezifizierte REST API oder SPARQL Endpoint auszutauschen.

Bibliographie

Bush, Vannevar (1945). „The atlantic monthly". in: *As we may think* 176, 1: 101-108.

Marshall, Catherine C. (1998) „Toward an ecology of hypertext annotation", in: *Proceedings of the ninth ACM conference on Hypertext and hypermedia*: links, objects, time and space —structure in hypermedia systems 40-49.

Müller-Birn, Claudia / Klüwer, Tina / Breitenfeld, Andre / Schlegel, Alexa / Benedix, Lukas (2015) „neonion: Combining Human and Machine Intelligence", in: *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing* 223-226.

Eine musikwissenschaftliche Edition in virtueller Umgebung: Die Einbindung der Anton Webern-Gesamtausgabe in SALSAH

Münnich, Stefan
stefan.muennich@unibas.ch
Universität Basel, Schweiz

Zusammenfassung

Die virtuelle Forschungsplattform SALSAH (*System for Annotation and Linkage of Sources in Arts and Humanities*) ermöglicht sowohl die Erzeugung, Bearbeitung und Verknüpfung von Daten und Inhalten als auch deren Präsentation in ein und derselben Umgebung. Ihre hierarchisierbare Datenmodellierung begünstigt die Herausbildung eines 'semantic web' aus digitalen Objekten mit ihren jeweiligen Annotationen und Verknüpfungen. Die an der Universität Basel ansässige Anton Webern-Gesamtausgabe nutzt diesen Ansatz für ihre historisch-kritische Edition des musikalischen Werks Anton Weberns gleich mehrfach: als Quellenarchiv, als Dokumentationsdatenbank sowie als editionspraktisches Arbeitswerkzeug. Die Posterpräsentation soll die strukturelle Konzeption für die Einbindung einer (musik-)wissenschaftlichen Edition in eine solche virtuelle Forschungsumgebung veranschaulichen.

Die virtuelle Forschungsumgebung SALSAH

SALSAH wurde vom Digital Humanities Lab der Universität Basel ursprünglich im Zusammenhang mit einem kunsthistorischen Projekt (Kunsthistorisches Seminar der Universität Basel / Digital Humanities Lab der Universität Basel) entwickelt, hat aber seither das Anwendungsspektrum auf andere geisteswissenschaftliche Disziplinen, so auch die Musikwissenschaft, erweitert. Als rein auf Internettechnologien basierende Forschungsumgebung unterstützt SALSAH eine standortunabhängige, interdisziplinäre Kollaboration von Forschenden bei der Arbeit mit und an digitalen Quellenbeständen. Über ein via Webbrowser zugängliches grafisches Benutzerinterface erfolgt dabei eine Anreicherung des Quellenmaterials mit Informationen (Annotationen), wie z. B. Quellenbeschreibungen, Transkriptionen, Literaturverweise u. ä., sowie deren Verknüpfung untereinander; auch externe (über das Internet erreich- und referenzierbare) Quellenbestände können hier eingebunden und auf gleiche Weise annotiert und verknüpft werden. Ein mehrgliedriges Berechtigungskonzept erlaubt zudem die flexible Verwaltung von Benutzerzugangsrechten für jede einzelne Annotation.

Erstellung einer musikwissenschaftlichen (Teil-)Edition in SALSAH

Die Webern-Gesamtausgabe verwendet diese Funktionalitäten von SALSAH in mehrfacher Weise: Zum einen als internes digitales Quellenarchiv, das nach der Sequenzierung von über 3600 hochauflösenden Einzeldigitalisaten das virtuelle Quellenmaterial in der Originalreihenfolge der realen Konvolute für die Editoren jederzeit zugänglich macht. Darüber hinaus fungiert SALSAH hier als Dokumentationsinstrument, in dem Informationen zu biographischen und chronologischen Hintergründen, zu Ergänzungsmaterialien wie Briefen oder Tagebüchern sowie zusätzliche Werkinformationen und Quellenlisten aufgenommen und für die weitere Forschung öffentlich und kostenfrei zugänglich bereitgestellt werden. Neben der Funktion als Datenbank und Archiv hat SALSAH aber vor allem eine wichtige Rolle als Arbeitsinstrument für die direkte editorische Arbeit der Webern-Gesamtausgabe, die hybrid konzipiert ist mit einem Print- (gedruckte Notenbände) und einem Online-Anteil (digital publizierte, transkribierte Notentexte und Textmaterialien, vor allem die Kritischen Berichte). Letzterer benötigt ein ausgefeiltes Strukturmodell, das die unterschiedlichen Teilabschnitte der Edition als digitale (Unter-)Objekte anlegt, hierarchisierbar verknüpft und in semantische Beziehungen zueinander setzt. So sind zum Beispiel der edierte Notentext und der Kritische Bericht Teilobjekte des Hauptobjekts Edition, während Einleitung / Entstehungsgeschichte, Quellenübersicht, -beschreibung und -bewertung sowie Textkritische Anmerkungen wiederum Teilobjekte des Objekts Kritischer Bericht darstellen (analog zum Aufbau der Druckbände). Besondere Rücksicht gilt dabei den unterschiedlichen textlichen Erscheinungsformen (Fließtext, Listen, tabellarische Darstellungen) der Teilobjekte, die sich in deren strukturellen Eigenschaften widerspiegeln müssen. Letztendlich sollen sämtliche Teilabschnitte des Kritischen Berichts als auch der edierte Notentext direkt innerhalb der virtuellen Forschungsumgebung erzeugt, bearbeitet und publiziert werden können. Ebenso muss eine in ihrer Reihenfolge der Teilobjekte festlegbare Konvertierung in verschiedene Ausgabeformate (z. B. für eine daraus zu erstellende Druckfassung des Kritischen Berichts) möglich sein.

Die Einbindung der Gesamtausgabe in eine solche virtuelle Forschungsumgebung dient somit nicht nur dem Selbstzweck oder einer bloßen Präsentation von (retro-)digitalisiertem Quellenmaterial, sie bringt genuin digital erzeugtes, frei verknüpfbares Forschungswissen hervor, und bildet somit einen vollgültigen Teilbereich des Editionsprojekts.

Bibliographie

Kunsthistorisches Seminar der Universität Basel / Digital Humanities Lab der Universität Basel (o. J.): *Die Bilderfolgen der Basler Frühdrucke: Spätmittelalterliche Didaxe als Bild-Text-Lektüre* http://www.salsah.org/incunabula [letzter Zugriff 15. Februar 2016].

Rosenthaler, Lukas / Schweizer, Tobias (2012): "SALSAH – eine webbasierte Forschungsplattform für die Geisteswissenschaften", in: *SAGW Bulletin* Januar 32-33.

Schingnitz, Barbara / Schweizer, Tobias: "SALSAH in der Nutzung durch die Anton Webern-Gesamtausgabe", in: Ahrend, Thomas / Schmidt, Matthias (eds.) (in Vorbereitung): *Webern-Studien*. 3: Webern-Philologien.

Schweizer, Tobias / Rosenthaler, Lukas (2011): "SALSAH – eine virtuelle Forschungsumgebung für die Geisteswissenschaften", in: *EVA Konferenz 2011 Berlin*. Elektronische Medien & Kunst, Kultur, Historie, die 18. Berliner Veranstaltung der Internationalen EVA-Serie Electronic Imaging & the Visual Arts 147-153.

Stefan George Digital

Neuber, Frederike
frederike.neuber@uni-graz.at
ZIM, Universität Graz, Österreich

Untersuchungsgegenstand und Zielsetzung

In der neueren deutschen Literatur steht Stefan George (1868-1933) wie kein anderer für die außergewöhnliche Beschäftigung eines Autors mit bzw. für die Verwendung von Typografie. Ab 1904 werden seine Werke in *Stefan-George-Schrift* (St-G) gedruckt. Der Formenkanon der serifenlosen Type basiert auf Georges Buchschrift[1] sowie auf der *Akzidenz-Grotesk* der Schriftgießerei Berthold und zeigt zudem Einflüsse historischer Schriften wie der Unziale und der karolingischen Minuskel. Bis zu der vom Dichter autorisierten *Gesamt-Ausgabe der Werke* (1927-1934) entwickelt sich das Typenrepertoire der Schriftart, sodass die St-G-Schrift nicht in einer, sondern in mehreren Fassungen vorliegt.

Eine serifenlose Schrift inmitten der in Deutschland tobenden Antiqua-Fraktur Debatte zu verwenden, ihr Design an der eigenen Handschrift zu orientieren und gleichzeitig auf historische Vorbilder zu referieren – lediglich ausschnitthaft verdeutlichen diese Aspekte die große Relevanz von Typografie für Georges Werk. Umso verwunderlicher ist es, dass bisherige Editionen[2] keine tiefentypografische Analyse der Drucke vornehmen. Die Abhandlungen der einschlägigen Forschung zur Gestaltung und Genese der Type sowie zu ihrer Verwendung und Wirkung sind dementsprechend dürftig. Daher ist das Ziel des Projekts *Stefan George Digital* (StGD) die erstmalige Edition der Drucküberlieferung der Georgeschen Lyrik, wobei der Schwerpunkt auf der Erschließung typografischer Formen mittels eines semantischen Modells liegt.[3]

Vorgehen und Methodik

Das Editionskorpus StGDs besteht aus 29 Druckausgaben der insgesamt 11 lyrischen Werke[4] Georges, in denen die Anwendung und die Entwicklung der typografischen Gestaltung sichtbar werden. Die digitalen Volltexte werden größtenteils aus bestehenden Repositorien (z. B. Deutsches Textarchiv und TextGrid Repository) semi-automatisch in ein projektspezifisches XML/TEI Schema konvertiert. Da der Schwerpunkt der Edition auf der buchwissenschaftlichen Erschließung des Materials liegt, werden die Daten entsprechend mit Metadaten (z. B. FRBR, METS) angereichert. Schließlich werden digitale Faksimiles, vereinzelt aus bestehenden Repositorien (z. B. ULB Düsseldorf), größtenteils jedoch erstmalig hochauflösend digitalisiert, auf einen IIIF basierten Imageserver mit angepassten Viewern (Open Seadragon, Mirador) in die Edition integriert. Die digitalen Bilder werden sowohl parallel zum edierten Text als auch separat typografisch annotiert. Während die Ebenen der Meso- (Schrift in der Fläche), Makro- (Organisation von Schrift) und Paratypografie (Materialität und Technik) im Rahmen der digitalen Edition weitestgehend mit bestehenden Datenmodellen erfasst werden können, wird das Modell zur Erschließung der Mikrotypografie (Formausstattungsmerkmale) erst im Rahmen des Projekts entwickelt (zu den typografischen Ebenen vgl. Stöckl 2004).

Die Modellierung der typografischen Detailformen erfolgt in Form einer Ontologie, welche ihre eindeutige Identifikation, formalisierte Beschreibung und Zitation ermöglicht. Damit wird eine netzwerkartige Erschließung und Verknüpfung unterschiedlicher Aspekte und Charakteristika von Schrift unternommen, welche sowohl für Mensch als auch für Maschine interpretierbar sind. Die Technologien des Semantic Web zur Wissensrepräsentation in Thesauri (SKOS), Klassenmodelle (RDFs) und Ontologien (OWL) können

dafür ebenso verwendet werden wie Methoden der Daten- und Softwaremodellierung (UML).

Sowohl die Digitale Edition als auch die Ontologie zur Beschreibung von Typografie werden unter CC BY-SA Lizenz auf der GAMS , der technischen Infrastruktur des Grazer Zentrums für Informationsmodellierung , bereitgestellt.

Kontextualisierung in den Digitalen Geisteswissenschaften

Das Projekt ist vorrangig für die digitale Editorik relevant, die seit geraumer Zeit verstärkt auch die Materialität von Dokumenten zu erschließen versucht. Statt den Weg der Abbildung von Schrift mittels Faksimiles oder ihrer Rekonstruktion im Rahmen der Transkription (z. B. Schriftfaksimile) zu gehen, wählt das Projekt die formale Modellierung und macht die Informationen so analysierbar. In diesem Zusammenhang trägt StGD auch zur Bildung einer [5] noch kaum bestehenden Digitalen Buchwissenschaft und Typenkunde bei, für welche der Ansatz der semantischen Modellierung von Typenformen ebenfalls neu ist.[6]

Schließlich kann das StGD auch als exemplarisch für den Einsatz und Umgang mit projektspezifischen Datenmodellen gelten. Um die Ontologie für die breitere Forschungscommunity nutzbar zu gestalten, wird die Übertragbarkeit des Modells auf verschiedene Arten von Typen, wie beispielsweise bewegliche Lettern und frühneuzeitliche Typen, getestet. Außerdem ist der Versuch eines mappings des Modells auf Handschriften in Zusammenarbeit mit DigiPal vorgesehen.

Zentrale Aspekte auf dem Poster

Neben einer Gesamtpräsentation des Projekts StGD, wird das Poster vorrangig drei aktuelle Herausforderungen illustrieren:

- Modellierung mikrotypografischer Formen basierend auf einer stabilen Terminologie, festgelegten Beschreibungskategorien (z. B. Form und Stil) sowie zentraler Unterscheidungsmerkmale (z. B. Serifen und Strichdicke).
- Vernetzung der vier typografischen Ebenen Mikro-, Meso, Makro- und Paratypografie, welche mit unterschiedlichen Datenmodellen an unterschiedlichen Orten der Edition erfasst werden.
- Visualisierung typografischer Genese sowie gestalterischer Brüche und Kontinuitätslinien über das lyrische Gesamtwerk Georges hinweg.

Notes

1. Ab circa 1896 stilisierte George seine Kursivhandschrift verstärkt zu einer Buchschrift, welche in der Folgezeit auch von Mitgliedern des George-Kreises verwendet wurde.
2. Neben den Editionen einzelner Werke v.a.: *Sämtliche Werke in 18. Bänden*. Herausgegeben von der Stefan George-Stiftung; bearbeitet von Peter Landmann und Ute Oelmann. Stuttgart 1981-2013.
3. Das Projekt Stefan George Digital entsteht im Rahmen von DiXiT (Digital Scholarly Editions Initial Training Network), als Teil der Marie Curie Actions im 7. Rahmenprogramm der Europäischen Kommission.
4. Die Bandzählung und -aufteilung folgt der Veröffentlichung der Werke in der Gesamt-Ausgabe (1927-1934). „Der Teppich des Lebens“ ist beispielsweise vier Mal (1900/1,1901/2,1904/3, 1932/Gesamt-Ausgabe), „Der Stern des Bundes“ nur zwei Mal (1914/1, 1928/Gesamt-Ausgabe) im Korpus repräsentiert. Im Laufe des Projekts werden außerdem auch einzelne Gedichte, welche im Vorfeld in der von George gegründeten Zeitschrift *Blätter für die Kunst* (1892-1919) veröffentlicht wurden, in das Editionskorpus aufgenommen.
5. Vgl. die „stilisierte Ursprungsschrift“ (Tool Dreifachlupe der Universität Würzburg).
6. Das Typenrepertorium des Gesamtkatalogs der Wiegendrucke beschreibt die formalen Aspekte von Typen beispielsweise in Prosaform (vgl. Inkunabelreferat der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz).

Bibliographie

Eide, Øyvind (2015): "Ontologies, Data Modeling, and TEI", in: *Journal of the Text Encoding Initiative* 8. http://jtei.revues.org/1191 [letzter Zugriff 13. Oktober 2015].

Inkunabelreferat der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (o.J.): *TW*. Typenrepertorium der Wiegendrucke http://tw.staatsbibliothek-berlin.de/ [letzter Zugriff 15. Februar 2016].

Lucius von, Wolf D. (2012): "Buchgestaltung und Typographie bei Stefan George", in: Aurnhammer, Achim / Braungart, Wolfgang / Breuer, Stefan / Oelmann, Ute (eds.): *Stefan George und sein Kreis*. Ein Handbuch, 1. Berlin / Boston: De Gruyter 467-491.

Stöckl, Hartmut (2004): "Typographie: Körper und Gewand des Textes. Linguistische Überlegungen zu typographischer Gestaltung", in: *ZfAL Zeitschrift für Angewandte Linguistik* 41: 5-48.

Stokes, Peter A. (2011): *Describing Handwriting*. Part I: http://www.digipal.eu/blog/describing-handwriting-part-i/ [letzter Zugriff 13. Oktober 2015].

Universität Würzburg (o. J.): *Dreifachlupe* http://vb.uni-wuerzburg.de/ub/lskd/dreifachlupe.html [letzter Zugriff 15. Februar 2016].

Wehde, Susanne (2000): *Typographische Kultur*. Eine zeichenhistorische und kulturgeschichtliche Studie zur Typographie und ihrer Entwicklung. Tübingen: Niemeyer.

Der Lehrpraxis im Transfer-Facharbeitskreis "Digitale Geisteswissenschaften in Sachsen"

Pfeil, Patrick

ppfeil@uni-leipzig.de
Universität Leipzig, Deutschland, Alte Geschichte

Mehner, Caroline

caroline.mehner@uni-leipzig.de
Universität Leipzig, Deutschland, Hochschuldidaktisches Zentrum Sachsen, Lehrpraxis im Transfer

Auf dem Poster soll die Arbeit des Facharbeitskreises „ Digitale Geisteswissenschaften in Sachsen “ vorgestellt werden. Dieser ist ein loser Zusammenschluss der im Bereich der Digital Humanities lehrenden und forschenden Wissenschaftler_innen an den sächsischen Hochschulen. Zusätzlich sind auch Einrichtungen und Hochschulen aus Thüringen und Sachsen-Anhalt beteiligt. Der Facharbeitskreis wird im Rahmen des Projektes Lehrpraxis im Transfer (LiT), welches durch den Qualitätspakt Lehre finanziert ist, gefördert und ist somit mit dem Hochschuldidaktischen Zentrum Sachsen (HDS) assoziiert. Die Koordination hat der Lehrstuhl für Alte Geschichte der Universität Leipzig unter der Leitung von Charlotte Schubert übernommen. Hervorgegangen ist er aus dem ebenfalls vom HDS geförderten Lehr-Lern-Projekt „ Neue Medien in den Geisteswissenschaften in Lehre und Forschung “, welches von April 2014 bis März 2015 am Lehrstuhl für Alte Geschichte an der Universität Leipzig angesiedelt war.

Ziel des Projektes ist die Bündelung der Initiativen und Projekte im mitteldeutschen Raum, die sich mit den Digital Humanities in der Forschung sowie in der Lehre beschäftigen. Durch regelmäßige Treffen sollen eine Vernetzung erreicht, Projektideen diskutiert, Synergieeffekte erzeugt und inhaltliche Punkte diskutiert werden. Der Facharbeitskreis existiert seit 2014 und hat zurzeit ca. 30 aktive Wissenschaftler_innen als Mitglieder. Diese kommen von der Universität Leipzig, der TU Dresden, der TU Chemnitz, der TU Bergakademie Freiberg, der Friedrich-Schiller-Universität Jena, HTWK Leipzig, Martin-Luther-Universität Halle-Wittenberg, Universität Erfurt sowie vom Forschungszentrum Gotha, der Klassik-Stiftung Weimar und dem Forschungsverbund Marbach-Weimar-Wolfenbüttel.

Mit der Arbeit des Facharbeitskreises konnte der Vernetzungsgedanke auch an verschiedene Hochschulen weitergetragen werden. So unterstützte er beispielsweise die Gründung des DHnet Jena an der Friedrich-Schiller-Universität Jena, welches zusammen mit dem ICE (Interdisciplinary Center of E-Humanities in History and Social Sciences) des Max-Weber-Kollegs Erfurt und dem „ Netzwerk für digitale Geisteswissenschaften Erfurt “ das „DH-Treffen Thüringen“ als Plattform für einen regelmäßigen Austausch der Initiativen und Projekte der Digital Humanities in Thüringen initiierte.

Inhaltlich werden von einzelnen Mitgliedern des Facharbeitskreises und deren Arbeitsgruppen in den Treffen eigene Themen zunächst vorgestellt und dann gemeinsam bearbeitet. Im Jahr 2014 und der ersten Hälfte 2015 stand die Diskussion um die Rolle der Digital Humanities in der Hochschullandschaft im Vordergrund. Hierbei wurde eine umfassende Diskussion um die konträren Meinungen, also Digital Humanities als Teil der Fachdisziplinen oder als eigene Wissenschaftsdisziplin, geführt. Für die zweite Hälfte 2015 ist die Entwicklung eines Positionspapiers zum Thema Infrastruktur vorgesehen, das verschiedene Streitpunkte in diesem Bereich aufgreifen, analysieren und kommentieren soll. Nach Fertigstellung des Papiers ist geplant, dieses im Kreise der Digital Humanities als Diskussionsansatz oder auch Handreichung zu veröffentlichen.

Das Poster soll einerseits zeigen, wie eine regionale Vernetzung initiiert werden kann und welche Möglichkeiten eine solche bietet. Es sollen der Gründungsprozess und die Dissemination der Initiative dargestellt werden.

Andererseits soll ein Einblick in die aktuelle inhaltliche Arbeit des Facharbeitskreises gegeben werden. Deshalb wird das bis zur DHd-Tagung entwickelte Positionspapier zum Thema Infrastruktur in den Digital Humanities auf dem Poster präsentiert. Es sollen der Diskussionsprozess dargestellt und verschiedene umstrittene Punkte im Detail behandelt werden.

Damit ist das Poster beispielgebend für regionale Vernetzung auch über die Grenzen einzelner Bundesländer hinaus und es zeigt weiter, dass in Organisationen wie dem LiT-Facharbeitskreis „Digital Humanities in Sachsen“ neben der wichtigen Vernetzungs- und Koordinationstätigkeit auch inhaltliche Arbeit im Bereich Digital Humanities stattfindet.

Bibliographie

Digital Humanities an der Friedrich-Schiller-Universität Jena (o. J.): *DHnet Jena* http://dhnet.uni-jena.de/index.php?id=124 [letzter Zugriff 15. Februar 2016].

Universität Erfurt (o. J.): *Netzwerk für digitale Geisteswissenschaften an der Universität Erfurt* [letzter Zugriff 15. Februar 2016].

Little Data on Big Map (Operative Verbildlichung von lokal existierten Daten der linguistischen Feldforschung)

Pourtskhvanidze, Zakharia
pourtskhvanidze@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Der Poster beschreibt den gegenwärtigen Stand des Projekts WALDI (World Atlas of Little Data Infrastructure), das am Institut für Empirische Sprachwissenschaft gemeinsam mit dem StudiumDigitale seit Mai 2015 realisiert wird.

Durch empirische Arbeit in der Sprachforschung entstehen verhältnismäßig kleine Daten, die i. d. R. von einzelnen Forschern erhoben werden und für die aktuelle Zwecke lokal (d. h. nicht-netzbasiert) abgespeichert. Die Verwendung von solchen „Little Data“ ist durch eine konkrete Fragestellung und das Forschungsinteresse beschränkt. Aus diesem Grund geraten sie nach dem Auslaufen einer Forschungsphase schnell in Vergessenheit.

Das primäre Ziel des Projektes ist die Konstruktion eines webbasierten Tools, das den linguistischen Feldforschern erlaubt lokal abgespeicherte Sprachdaten auf einem digitalen Atlas abzubilden und die Existenz dieser Daten für die Nutzer des Atlasses bekanntzugeben.

Eine webbasierte Anwendung erlaubt die freie Nutzeranmeldung auf einer digitalen Weltkarte (basierend auf Google Maps). Die Nutzer_innen editieren die Karte, indem sie die elizitierten Daten der dokumentierten Sprache (kleine Sprachkorpora) geographisch fixieren (pinnen) und den fixierten Pin mit einem TEI-ähnlichen Header versehen. Das Pinnen geschieht anhand von GPS-Angaben, die durch einfache Mauszeigerbewegungen generiert werden (man kann für viele Sprachen z. B. im Amazonas oder kaukasischen Hochland keine Orte in Google Maps lokalisieren). Der Header beinhaltet alle relevanten Sprachdateninformationen, die in folgende 5 Reiter-Kolumnen unterteilt sind: 1. Korpus Information (Autoren, Institutionen, Release, Größe etc.); 2. Sprache (Status, Schrifttum, Typologie etc.); 3. Die Art und der Aufarbeitungsstand der Daten (Transkription, Annotation etc.); 4. Erfasste Texte (Art, Epoche, Genre etc.); 5. Zugänglichkeit und Kontakt.

Zur technischen Realisierung des Tools kommen die Elemente aus Google Maps, JavaScript, MySQL zum Einsatz.

Bibliographie

Bubenhofer, Noah (2014): *Visual Linguistics - Sprache sehen*. http://www.visual-linguistics.net/ [letzter Zugriff 14. Februar 2016].

Dryer, Matthew S. / Haspelmath, Martin (2013): *World Atlas of Language Structures Online* (WALS). Leipzig: Max Planck Institute for Evolutionary Anthropology http://wals.info/ [letzter Zugriff 14. Februar 2016].

Krämer, Sybille (2009): „Operative Bildlichkeit. Von der ‘Grammatologie’ zu einer ‘Diagrammatologie’? Reflexionen über erkennendes Sehen“, in: Hessler, Martina / Mersch, Dieter (eds.): *Logik des Bildlichen*. Zur Kritik der ikonischen Vernunft. Bielefeld: transcript 94-123.

"Excerpta Constantiniana: vom Palimpsest zur Edition einer mittelalterlichen Enzyklopädie"

Rafiyenko, Dariya
dariya.rafiyenko@uni-leipzig.de
Universität Leipzig, Deutschland

In diesem Poster wird die digitale Edition der *Excerpta Constantiniana* (im Weiteren *Excerpta*), einer byzantinischen Geschichtsenzyklopädie, die im 10. Jahrhundert in Konstantinopel in Altgriechisch verfasst wurde, vorgestellt.

Zugrunde liegt ein disziplinspezifisches Forschungsprojekt im Bereich Klassische und Byzantinische Philologie, dessen Ziel darin besteht, die Edition einer wichtigen Quelle der byzantinischen Geschichtsschreibung vorzulegen. Das Ziel dieses Posters ist fachübergreifend und besteht darin, die Rolle des Herausgebers sowie das Konzept der Präsentation einer historischen Quelle in digitaler Umgebung zu definieren.

Bei den *Excerpta* handelt es sich um ein groß angelegtes, mehrere Bände umfassendes Werk. Es besteht aus mehreren Tausend einzelnen Auszügen (Exzerpten), die inhaltlich aus etwa drei Dutzend antiken und byzantinischen Geschichtswerken stammen. Die erhaltenen Reste umfassen etwa 560 000 Wörter (es wird vermutet, dass fast das Zehnfache verlorengegangen ist). Die beiden erhaltenen Originalhandschriften der *Excerpta* (je ein Band) zeichnen sich durch ein bemerkenswertes Layout aus: zum Zweck der Navigation durch den Inhalt

wurden an deren Rändern mehrere hundert Notizen und Piktogramme angebracht (s. Abbildung 2).

Die Edition des Gesamtwerks befindet sich in der Vorbereitungsphase; exemplarisch wurde bereits ein Abschnitt aus den *Excerpta* ediert, und zwar 24 Seiten aus der Originalhandschrift *Vaticanus graecus* 73 (ca. 9 000 Wörter). Bei der Handschrift handelt es sich um einen Palimpsest: der Text der *Excerpta* wurde etwa im 14. Jahrhundert ausradiert und mit einem anderen Text überschrieben, sodass der frühere Text heute nur mühsam lesbar ist (s. Abbildung 1).

Die Standardlösung wäre es, das Faksimile der Handschrift zu publizieren und eine historisch-kritische Edition des Texts anzufertigen. Die Publikation von dermaßen beschädigten Seiten erwies sich jedoch als wenig ergiebig. Auch die traditionelle Art der Textgestaltung im Rahmen einer historisch-kritischen Edition war kaum für die Wiedergabe der *Excerpta* geeignet. So bedurfte es beispielsweise einer originalgetreuen Reproduktion der Notizen und Piktogramme an den Rändern, die einen wichtigen Beitrag für das Verständnis des Textes leisten.

Die Art der Gestaltung der *Excerpta* war ausschlaggebend bei der Entwicklung des Konzepts dieser Digitaledition. Es wurde eine *pluralistische* Herangehensweise an den Text zugrunde gelegt, die gleichzeitig mehrere *Ansichten* (Wiedergabemöglichkeiten) desselben Textes ermöglicht. Dabei wurden drei Grundansichten ausgewählt: (1) die digitale Rekonstruktion der Handschrift (*topographische Edition*), (2) die diplomatische Abschrift (*diplomatische Edition*) sowie (3) die normalisierte, historisch-kritische Version des Textes (*digitale historisch-kritische Edition*).

Topographische Edition: Die topographische Edition umfasst die digitale Rekonstruktion des Originals und die zweidimensionale, detaillierte Darstellung der Oberfläche. Das bedeutet, dass der ausradierte Text der Handschrift, stark vergrößert, auf Touchscreen mit Stylus nachgemalt wird (s. Abbildungen 1 und 2). Diese Methodik wurde hier m.W. zum ersten Mal angewendet und wird daher im Poster vorgestellt. Sie kombiniert menschliches Expertenwissen mit den aktuellen technischen Möglichkeiten.

Diplomatische Edition: Unter einer diplomatischen Edition wird eine möglichst originalgetreue Abschrift einer Handschrift verstanden. In dieser Ansicht wird die Gestaltung der Originalhandschrift, d. h. vor allem das Layout und die Navigationselemente des ursprünglichen Texts, visualisiert (s. Abbildung 2). Nach Möglichkeit wird die ursprüngliche Orthographie wiedergeben. Hier ist auch die Option vorgesehen, innerhalb derselben Ansicht auf normalisierte Orthographie umzuschalten (dies ermöglicht z. B. die Wahl zwischen mittelalterlicher und moderner Zeichensetzung, zwischen der Schreibweise mit Abbreviaturen oder mit deren Auflösung u. ä.)

Digitale historisch-kritische Edition: Diese Ansicht hat das Layout einer modernen Edition, die Orthographie wird weitgehend normalisiert. Außerdem wird unter dieser Ansicht als Option die Möglichkeit der Hervorhebung unterschiedlicher Textinhalte angeboten, wie z. B. von Zitaten, Orten, Personennamen, Völkerbezeichnungen usw. In diese Ansicht gehören ferner der kritische Apparat sowie Indices mit Namen, Orten usw.

Technisch wird das folgenderweise implementiert. Die topographische Edition wird in Form von Bildern hergestellt. Die Transkription der Handschrift für die diplomatische und historisch-kritische Edition wird in TEI-XML angefertigt. Erscheinungen, die mit XML-Tags kodiert werden, werden in größere *Blöcke* aufgeteilt. Die wichtigsten Blöcke sind:

- *Physischer Zustand der Handschrift und physische Struktur des Texts*: physische Schäden der Handschrift, Lesbarkeit des Textes, Seiten- und Zeileneinteilung an den Stellen, wo sie nicht mit der logischen Struktur des Texts in Verbindung stehen (s. u.);
- *Logische Struktur des Texts sowie alle Elemente des Layouts, welche die Navigation im Text unterstützen*: Einheiten wie Bände, Kapitel, Exzerpte; Elemente der Gestaltung, die auf diese Einteilung verweisen (z. B. größere Leerräume im Text); Piktogramme und Randnotizen;
- *Orthographie der Handschrift*: originale Zeichensetzung, Akzentuierung, Abbreviaturen und Ligaturen;
- *Normalisierte Orthographie*: moderne Zeichensetzung, Worttrennung (fehlt in der Handschrift), Groß- und Kleinschreibung nach modernen Normen.
- *Inhalte im Text*: Zitate, Namen, Orte, Völker u. a.

Die Webdarstellung wird auf der Basis von XSLT erstellt. Für jede Ansicht wird einzeln modelliert, wie die einzelnen Blöcke der Tags transformiert werden sollen. So ist beispielsweise für das Layout der diplomatischen Edition der physische Zustand der Handschrift und die physische Struktur des Texts maßgebend, während für das Layout der historisch-kritischen Edition die logische Struktur des Texts entscheidend ist. Die endgültige Darstellung wird über Cascading Stylesheets (CSS) gestaltet. Es ist geplant, die Edition bis Mitte des Jahres 2016 online zu stellen.

Abb. 1: Der Prozess der graphischen Rekonstruktion des Palimpsests

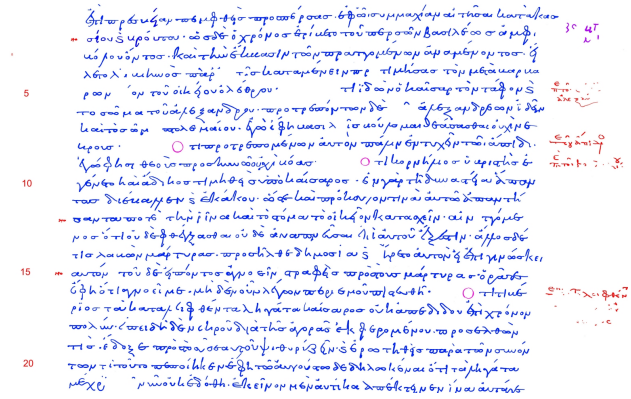


Abb. 2: Die topographische Edition

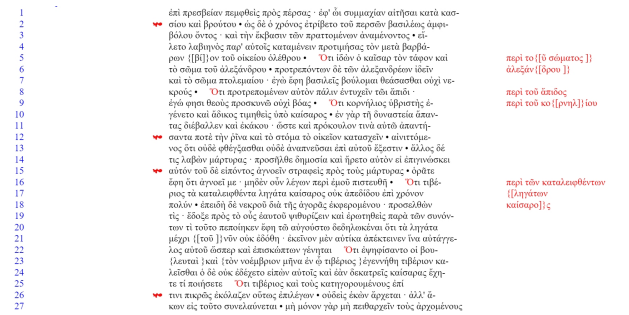


Abb. 3: Die diplomatische Edition

Fragment N. 1 (R)

Λαβιηνος ἐπὶ πρεσβείαν πεμφθεὶς πρὸς Πέρσας, ἐφ' ᾧ συμμαχίαν αἰτῆσαι κατὰ Καισσίου καὶ Βρούτου· ὡς δὲ ὁ χρόνος ἐτρίβετο τοῦ Περσῶν βασιλέως ἀμφιβόλου ὄντος καὶ τὴν ἔκβασιν τῶν πραττομένων ἀναμένοντος, εἵλετο Λαβιῆνος παρ' αὐτοῖς καταμένειν, προτιμήσας τὸν μετὰ βαρβάρων βίον τοῦ οἰκείου ὀλέθρου.

Fragment N. 2 (R)

Ὅτι ἰδὼν ὁ Καῖσαρ τὸν τάφον καὶ τὸ σῶμα τοῦ Ἀλεξάνδρου, προτρεπόντων δὲ τῶν Ἀλεξανδρέων ἰδεῖν καὶ τὸ σῶμα Πτολεμαίου, „ἐγὼ" ἔφη „βασιλεῖς βούλομαι θεάσασθαι, οὐχὶ νεκρούς."

Fragment N. 3 (R)

Ὅτι προτρεπομένων αὐτὸν πάλιν ἐντυχεῖν τῇ Ἀπίδι, „ἐγὼ" φησι „θεοὺς προσκυνῶ, οὐχὶ βοῦς."

Fragment N. 4 (R)

Ὅτι Κορνήλιος ὕβριστὴς ἐγένετο καὶ ἄδικος τιμηθεὶς ὑπὸ Καίσαρος· ἐν γὰρ τῇ δυναστείᾳ ἅπαντας διέβαλλεν καὶ ἐκάκου, ὥστε καὶ Πρόκουλόν τινα αὐτῷ ἀπαντήσαντά ποτε τὴν ῥῖνα καὶ τὸ στόμα τὸ οἰκεῖον κατασχεῖν αἰνιττόμενον, ὅτι οὐδὲ φθέγξασθαι οὐδὲ ἀναπνεῦσαι ἐπὶ αὐτοῦ ἔξεστιν. ἄλλος δέ τις λαβὼν μάρτυρας προσῆλθε δημοσίᾳ καὶ ἤρετο αὐτὸν εἰ ἐπιγινώσκει αὐτόν. τοῦ δὲ εἰπόντος ἀγνοεῖν, στραφεὶς πρὸς τοὺς μάρτυρας „ὁρᾶτε" ἔφη „ὅτι ἀγνοεῖ με· μηδὲν οὖν λέγων περὶ ἐμοῦ πιστευθῇ."

Abb. 4: Die historisch-kritische Edition

DARIAH-DKPro-Wrapper

Reimer, Nils
reimers@ukp.informatik.tu-darmstadt.de
TU Darmstadt, Deutschland

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
Universität Würzburg, Deutschland

Pernes, Stefan
stefan.pernes@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reger, Isabella
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Dieses Poster soll den DARIAH-DKPro-Wrapper vorstellen, der aus einer Kooperation zwischen dem Lehrstuhl für Computerphilologie der Universität Würzburg und dem Ubiquituous Knowledge Processing Lab der TU Darmstadt im Rahmen von DARIAH-DE entstanden ist.

DKPro integriert zahlreiche (unabhängig entstandene) Softwarekomponenten zum Natural Language Processing (NLP) und ermöglicht so dem Nutzer die Anwendung typischer NLP-Aufgaben wie Tokenisierung, Part-of-Speech-Tagging, Named Entity Recognition oder Dependency Parsing mit State-of-the-Art Werkzeugen. Es basiert auf dem Framework UIMA. Für Nutzer, die nicht aus dem Umfeld der Informatik oder Computerlinguistik kommen, ist die Schwelle zur Verwendung allerdings recht hoch: das komplexe Framework muss in Java angesprochen werden.

Um diese Hürde zu senken und einer größeren Zahl auch von weniger technisch versierten Nutzern die Verwendung zu ermöglichen, wurde der DARIAH-DKPro-Wrapper entwickelt. Dieser ermöglicht es, eine Pipeline mit mehreren Komponenten über die Kommandozeile auszuführen und damit auch längere Textdokumente und Textsammlungen zu verarbeiten. Zudem können eine ganze Reihe von Einstellungen bequem und individuell über Konfigurationsdateien vorgenommen werden: über die Auswahl der Sprache bis hin zur Aktivierung und Deaktivierung einzelner Komponenten und der Auswahl bestimmter Komponenten

oder Modelle. Auf diese Weise kann jeder Nutzer vorgefertigte Pipelines verwenden oder eine auf seine Bedürfnisse zugeschnittene Pipeline individuell zusammenstellen. Der Wrapper ist stets aktuell über GitHub (https://github.com/DARIAH-DE/DARIAH-DKPro-Wrapper) verfügbar, ebenso wie die dazugehörige Dokumentation des DARIAH-DKPro-Wrapper v0.4.3 (2016).

Um die anschließende Weiterverarbeitung derart prozessierter Dokumente ebenfalls zu vereinfachen, wurde ein entsprechendes Ausgabeformat entwickelt. Dieses lehnt sich an das CoNLL2009-Format an und stellt die Ergebnisse der Pipeline in tabellarischer Form dar. Dabei befindet sich in jeder Zeile ein Token, während die dazugehörigen Informationen wie Lemma, POS-Tag und ähnliches je in einer Spalte stehen. Dadurch werden alle durch Komponenten der Pipeline ermittelten Informationen in einer Datei zusammengefasst. Dieses Format hat den Vorteil, dass es für menschliche Nutzer übersichtlich und gut lesbar ist. Zudem ist es als Tabstopp-getrennte Datei auch für gängige Skriptsprachen wie Python oder R, sowie Tabellenkalkulationsprogramme wie Excel leicht zugänglich.

Um die Verwendung des Wrappers und die Weiterarbeit mit dem Ausgabeformat zusätzlich zur Dokumentation anschaulich zu beschreiben, wurden außerdem eine Reihe von Tutorials zu Beispielanwendungen aus Bereichen der digitalen Literaturwissenschaft, wie zum Beispiel der Stilometrie oder dem Topic Modeling, verfasst. Die Dokumentation sowie die Tutorials sind ebenfalls auf GitHub zu finden.

Das Poster wird all diese Punkte in übersichtlicher Form zusammenführen und potentiellen Nutzern präsentieren. Dabei werden die Funktionsweise der Pipeline, die Arbeit mit den Konfigurationsdateien, der Aufbau und die Verwendung des Ausgabeformats sowie Anwendungsbeispiele im Mittelpunkt stehen.

Bibliographie

Dokumentation: DARIAH-DKPro-Wrapper v0.4.3 (2016): *User guide DARIAH-DKPro-Wrapper v0.4.3* DARIAH2 - Cluster 5, Use Case 1 Team. Universität Würzburg, TU Darmstadt - DARIAH-DE https://rawgit.com/DARIAH-DE/DARIAH-DKPro-Wrapper/master/doc/user-guide.html [letzter Zugriff 08. Januar 2016].

CoNLL-2009 Format (2008-*): *CoNLL-2009 Shared Task*. Syntactic and Semantic Dependencies in Multiple Languages. Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic, Faculty of Mathematics and Physics https://ufal.mff.cuni.cz/conll2009-st/task-description.html [letzter Zugriff 08. Januar 2016].

Schichten über Schichten - Die Zukunft der Handschriftenforschung

Schaßan, Torsten
schassan@hab.de
Herzog August Bibliothek Wolfenbüttel, Deutschland

Die Handschriftenforschung hat sich in der Vergangenheit auf die Erschließung durch Katalogisierung auf der einen Seite und Textedition auf der anderen gestützt. Sowohl Katalogisierung als auch Textedition stützen sich nunmehr immer stärker auf digitale Daten. Insbesondere sind immer mehr Handschriften als Imagedigitalisate verfügbar. Zentrale Nachweisportale wie Manuscripta Mediaevalia, e-codices oder Biblissima bieten Einstiegspunkte für alle Arten von handschriftenbezogenen Informationen. Mit der Entität „Handschrift“ im Mittelpunkt können Digitalisate, Katalogisate, Nachweise zu Wasserzeichen, Einbandstempeln oder Provenienzen verbunden werden. Solche Verknüpfung erfordern die Anreicherung mit Normdaten, um so zur Grundlage für Linked Open Data und das Semantic Web zu werden.

Während die Notwendigkeit bzw. Nützlichkeit, Normdaten für Personen, Institutionen, Orte und auch teilweise Bildinhalte anzubieten und zu nutzen schon relativ weit verbreitet ist und diese auch von der Forschung eingesetzt werden, denken die Handschriftencommunities weiter: Auch die Objekte selbst, Handschriften und Texte sollen in normierter Form zugänglich gemacht werden. Entsprechende Normdatensysteme zur eineindeutigen Identifizierung von Objekten des kulturellen Erbes befinden sich in der Entwicklung (Kailus 2013).

Die Entwicklung zur Erschließung historischer Materialien macht hier allerdings nicht halt: Die Überlegung, dass historische Dokumente Schichten von vielerlei Informationen aufweisen und auch in Zukunft immer neue Schichten anhäufen werden, ist der Ausgangspunkt für die Entwicklung eines neuen Datenformats, Shared Canvas:

- Die Materialoberfläche kann durch Schichten von Bildern repräsentiert werden, doch auch sie selbst kann aus Schichten bestehen, die durch Abbildungsverfahren wie der Multispektralfotographie sichtbar gemacht werden können.

- Auch der Beschreibprozess kann als Schichtung historischer Prozesse verstanden werden. Das Befüllen der Seite geschieht in einer chronologischen Reihenfolge, die in einer Edition sichtbar gemacht werden kann. Dieser Ansatz hat in der theoretischen Strömung der genetischen Edition seinen Niederschlag gefunden. (Brüning et al. 2013; Burnard et.al. 2011)

- Einzelne Oberflächen können in beliebige Kontexte gesetzt oder auch in beliebige Reihenfolgen gebracht werden: Virtuelle Sammlungen entstehen durch forschungsgeleitete Zusammenstellung; auseinander gerissene Handschriften, sogenannte 'Codices discissi', können virtuell zusammengesetzt, falsch gebundene Handschriften virtuell in die richtige Reihenfolge gebracht werden.

- Schließlich wird die Annotation der (Bild)Oberfläche selbst, durch Forscher und Handschriftenkundler, weiter Schichten des Wissens über die Objekte versammeln und zugänglich machen, die ebenfalls visualisiert sein wollen.

Um all das zu erlauben, ist in internationaler Zusammenarbeit die Idee des Shard Canvas (Sanderson & Albritton 2013) entstanden welches eine theoretische Antwort auf die vielfältigen Bedürfnisse der Handschriftenforschung geben soll. Basierend auf Semantic Web-Technologien wie RDF und JSON sollen die geschichteten Informationen abgerufen, in Beziehung gesetzt und visualisiert werden können. Mit IIIF (International Image Interoperability Framework, sprich: Triple-I, F) liegt ein implementierendes Format vor.

Im Poster werden die Entwicklungen im Feld der Handschriftenforschung, ihrer theoretischen Grundlagen und die daraus resultierenden Möglichkeiten im Semantic Web aufgezeigt.

Bibliographie

Biblissima (2013): *Bibliotheca bibliothecarum novissima* Patrimoine écrit du Moyen Âge et de la Renaissance, Bibliothèque numérique, VIIIe au XVIIIe siècle. http://www.biblissima-condorcet.fr [letzter Zugriff 31. Dezember 2015].

Brüning, Gerrit / Henzel, Katrin / Pravida, Dietmar (2013): „Multiple Encoding in Genetic Editions: The Case of 'Faust'", in: *Journal of the Text Encoding Initiative* 4 http://jtei.revues.org/697 [letzter Zugriff 31. Dezember 2015].

Burnard, Lou / Jannidis, Fotis / Pierazzo, Elena / Rehbein, Malte (2011): *An Encoding Model for Genetic Editions* http://www.tei-c.org/Activities/Council/Working/tcw19.html [letzter Zugriff 31. Dezember 2015]

Burrows, Toby (2011): „Applying Semantic Web Technologies to Medieval Manuscript Research", in: Fischer, Franz / Fritze, Christiane / Vogeler, Georg (eds.): *Kodikologie und Paläographie im digitalen Zeitalter 2*. Norderstedt: BoD 117-131 http://kups.ub.uni-koeln.de/4346/ [letzter Zugriff 31. Dezember 2015].

EBDB (2001-*): *Einbanddatenbank* Württembergische Landesbibliothek Stuttgart, Herzog August Bibliothek Wolfenbüttel, Bayerische Staatsbibliothek München, Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Universitäts- und Landesbibliothek Darmstadt, Universitätsbibliothek Rostock http://www.hist-einband.de [letzter Zugriff 31. Dezember 2015].

e-codices (2003-*): *e-codices: Virtuelle Handschriftenbibliothek der Schweiz*. Universität Freiburg, Schweiz http://e-codices.unifr.ch [letzter Zugriff 31. Dezember 2015].

IIIF (2011-*): *International Image Interoperability Framework* http://iiif.io [letzter Zugriff 31. Dezember 2015].

Kailus, Angela (2013): *RDA in der Dokumentation von Kunst und Architektur*. Workshop am 10. September 2013. Deutsche Nationalbibliothek in Frankfurt am Main http://www.dnb.de/SharedDocs/Downloads/DE/DNB/standardisierung/rdaKultur2013Kailus.pdf [letzter Zugriff 31. Dezember 2015].

Manuscripta Mediaevalia (2000-*): Staatsbibliothek zu Berlin- Preußischer Kulturbesitz, Deutsches Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg, Bayerische Staatsbibliothek München http://www.manuscripta-mediaevalia.de [letzter Zugriff 31. Dezember 2015].

Sanderson, Robert / Albritton, Benjamin (2013): *Shared Canvas Data Model* http://iiif.io/model/shared-canvas/1.0/ [letzter Zugriff 31. Dezember 2015].

Sanderson, Robert / Albritton, Benjamin / Schwemmer, Rafael / Van de Sompel, Herbert (2011): "SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination", in: *arXiv:1104.2925* http://arxiv.org/abs/1104.2925 [letzter Zugriff 31. Dezember 2015].

Stinson, Timothy (2009): „Codicological Descriptions in the Digital Age", in: Rehbein, Malte / Sahle, Patrick / Schaßan, Torsten (eds.): *Kodikologie und Paläographie im digitalen Zeitalter*. Norderstedt: BoD 35-51 http://kups.ub.uni-koeln.de/2959/ [letzter Zugriff 31. Dezember 2015].

Terras, Melissa (2011): „Artefacts and Errors: Acknowledging Issues of Representation in the Digital: Imaging of Ancient Texts", in: Fischer, Franz / Fritze, Christiane / Vogeler, Georg (eds.): *Kodikologie und Paläographie im digitalen Zeitalter 2*. Norderstedt: BoD 43-61 http://kups.ub.uni-koeln.de/4342/ [letzter Zugriff 31. Dezember 2015].

WZIS(2010-*): *Wasserzeichen-Informationssystem*. Bayerische Staatsbibliothek München, Deutsche Nationalbibliothek Leipzig, Landesarchiv Baden-Württemberg, Österreichische Akademie der Wissenschaften, Staatsbibliothek zu Berlin, Württembergische Landesbibliothek Stuttgart, Universitätsbibliothek Leipzig http://www.wasserzeichen-online.de [letzter Zugriff 31. Dezember 2015].

Datenbank für Gesprochenes Deutsch (DGD)

Schmidt, Thomas
thomas.schmidt@ids-mannheim.de
IDS Mannheim, Deutschland

Eine Korpusplattform für die Arbeit mit mündlichen Daten

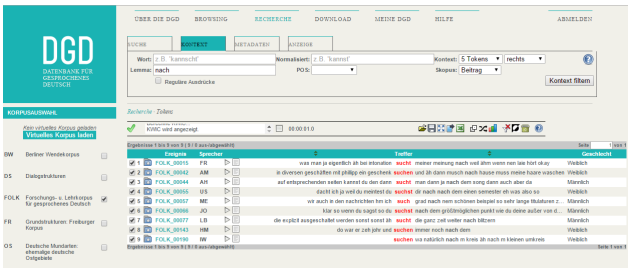
Die Datenbank für Gesprochenes Deutsch (DGD) (vgl. Institut für Deutsche Sprache; Schmidt 2014a) ist die zentrale Plattform für den Zugriff auf Daten des Archivs für Gesprochenes Deutsch (AGD). Über die DGD werden 23 mündliche Korpora des Deutschen im Gesamtumfang von mehr als 3000 Stunden Audio und 8 Millionen transkribierter Wort-Tokens angeboten.

Der Bestand umfasst erstens mehrere große Variationskorpora des Deutschen, insbesondere das Korpus „Deutsche Mundarten“ (Zwirner-Korpus) mit mehreren Satelliten-Korpora nach dem gleichen Design („Deutsche Mundarten: Ehemalige deutsche Ostgebiete“, „Deutsche Mundarten: DDR“ und mehrere kleinere, regional begrenzte Sammlungen von Dialektaufnahmen) sowie das Korpus „Deutsche Umgangssprachen“ (Pfeffer-Korpus). Diese Dokumentation binnendeutscher Mundarten wird komplementiert durch Korpora auslandsdeutscher Varietäten, z. B. das Korpus „Australiendeutsch“ und drei Korpora zum Emigrantendeutsch in Israel.

Zweitens bietet die DGD Zugriff auf verschiedene Gesprächskorpora, u. a. das Berliner Wendekorpus, die Korpora „Grundstrukturen“ (Freiburger Korpus) und „Dialogstrukturen“, sowie das Korpus „Elizitierte Konfliktgespräche“. Mit dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, Schmidt 2014b) wird im AGD ein großes, breit stratifiziertes Gesprächskorpus des Deutschen aufgebaut, das technisch und methodisch auf dem Stand aktueller bester Praktiken ist und der wissenschaftlichen Gemeinschaft ebenfalls über die DGD zur Verfügung gestellt wird.

Die Bestände sind nach einem einheitlichen XML basierten Metadatenschema dokumentiert und durch Transkriptions- und Annotationsdaten, die ebenfalls auf einem gemeinsamen XML-Datenmodell basieren, für die wissenschaftliche Analyse erschlossen.

Die DGD erlaubt zum einen ein exploratives Browsen auf diesen Daten. Korpus-, Sprecher- und Ereignisdokumentationen können eingesehen und die zugehörigen Audiodateien online abgespielt werden. Mit dem Audio alignierte Transkripte werden dem Benutzer in einer HTML5-basierten Darstellung präsentiert, die das Anspringen beliebiger Stellen im Transkript und das synchronisierte Abspielen der entsprechenden Segmente der Aufnahme ermöglicht. Diese Form des Zugangs dient sowohl dem Kennenlernen der Datenbestände als auch dem Einstieg in deren qualitative Analyse.



Für die gezielte Auswertung der Daten in quantitativer Hinsicht bietet die DGD zum anderen mehrere Recherchefunktionen. Über eine strukturierte Metadatensuche können nach flexibel spezifizierbaren Kriterien (z. B. Gespräche mit Sprechern aus dem norddeutschen Raum, älter als 40 Jahre) Teilmengen des Gesamtbestands ausgewählt und als virtuelle Korpora gespeichert werden. Die strukturierte Tokensuche erlaubt korpuslinguistische Anfragen über mehrere Annotationsebenen (Transkription, orthographische Normalisierung, Lemmatisierung, POS-Annotation), deren Ergebnisse in vielfältiger Hinsicht kontextualisiert (d. h. mit Metadaten korreliert, auf Transkript- und Aufnahmekontext rückbezogen) werden können.

Für die weitere Bearbeitung von Ausgangsdaten oder Analyseergebnissen bietet die DGD schließlich verschiedene Möglichkeiten zum Download von Datensätzen oder geeigneten Ausschnitten.

Bei allen Funktionen zum Browsen und Durchsuchen der Daten legt die DGD Wert darauf, korpusgesteuerte Analysemethoden zu ermöglichen, in denen Hypothesen aus den Daten selbst generiert und in einer interaktiven Auseinandersetzung mit selbigen schrittweise verfeinert werden können.

Die DGD ist seit Ende 2012 online und hat mittlerweile mehr als 4000 registrierte Nutzer_innen aus Forschung und Lehre. Datenbestände und Funktionalität werden kontinuierlich erweitert.

Bibliographie

Institut für Deutsche Sprache (IDS) (o. J.): *DGD.* Datenbank für Gesprochenes Deutsch http://dgd.ids-mannheim.de [letzter Zugriff 12. Februar 2016].

Schmidt, Thomas (2014a): "The Database for Spoken German – DGD2", in: *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland 1451-1457 http://www.lrec-conf.org/proceedings/lrec2014/pdf/171_Paper.pdf [letzter Zugriff 12. Februar 2016].

Schmidt, Thomas (2014b): "The Research and Teaching Corpus of Spoken German – FOLK", in: *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland 383-387 http://www.lrec-conf.org/proceedings/lrec2014/pdf/290_Paper.pdf [letzter Zugriff 12. Februar 2016].

CFDB: eine paläographische Datenbank neu- und spätbabylonischer Keilschriftzeichen

Schopper, Daniel
daniel.schopper@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Pirgruber, Reinhard
reinhard.pirngruber@univie.ac.at
Universität Wien, Österreich

Jursa, Michael
michael.jursa@univie.ac.at
Universität Wien, Österreich

Die hier unter dem Arbeitstitel **CFDB** vorgestellte Datenbank stellt den derzeit einzigartigen Versuch dar, Fragen zur paläographischen Entwicklung der babylonischen Keilschrift im ersten vorchristlichen Jahrtausend mit den Mitteln der Digital Humanities zu beantworten. Sie wird im Zuge des unter der Leitung von Michael Jursa an der Universität Wien durchgeführten Projekts *Diplomatik und Paläographie neu- und spätbabylonischer archivalischer Dokumente* (FWF P 26104) am *Austrian Centre for Digital Humanities* der Österreichischen Akademie der Wissenschaften entwickelt. Diese Datenbank stellt ein dynamisches und flexibles Untersuchungsinstrument dar, das im Hinblick auf die neu- und spätbabylonische Epigraphik eine beträchtliche Lücke in der Forschung zu schließen beabsichtigt.

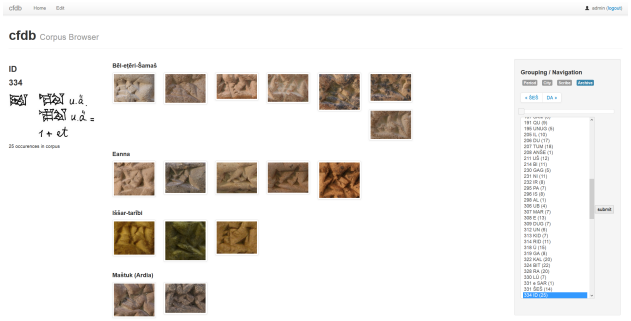
Verglichen mit alphabetischen Zeichensystemen weist die Keilschrift auf der Ebene des Schriftduktus eine hohe Zahl an potentiell objektivierbaren Merkmalen auf: Eigenschaften wie Schreibwinkel, Drucktiefe, Reihenfolge, Anordnung und Clustering von Keilen, Zeichengröße und andere Kriterien eignen sich ausgezeichnet für paläographische Untersuchungen. Vorrangiges Interesse dieser Seite des Forschungsprojekts ist die Identifizierung von gehäuft auftretenden standardisierten Zeichenformen (auf der Ebenen von Einzelzeichen und Ligaturen) bzw. die Frage, ob sich Abweichungen davon an die Größen Datum, Entstehungsort, Archiv oder Schreiber rückbinden lassen.

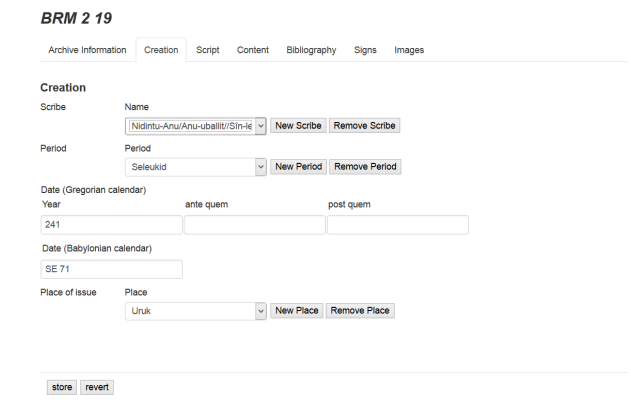
CFDB ist analog zu konventionellen assyrologischen Zeichenliste strukturiert: Jedem Graphem dieser Liste wird ein Korpus attestierter Formen (Allographen) in Form von annotierten Bildsegmenten beigestellt. Die Allographen werden anhand ihrer Datierung und Periodisierung sowie ihres Herkunftsortes, des Archives, der Textsorte und, sofern möglich, des Schreibers klassifiziert. Somit erlaubt CFDB die Untersuchung dieses Zeichenkorpus einerseits hinsichtlich des Verhältnisses verschiedener Allographen zueinander in einer diachronen Perspektive sowie andererseits in Bezug auf die oben genannten unterschiedlichen Aspekte des Schreibduktus, und zwar über die Ebene von Einzelzeichen oder -tafeln hinaus. Auf einer Metaebene bietet die Datenbank die Möglichkeit, in der Literatur häufig anzutreffende, allerdings lediglich impressionistisch begründete Differenzierungen zwischen „formalen“, „kalligraphischen“ oder „kursiven“ Schriftduktus zu objektivieren und zur Diskussion zu stellen.

Die Implementierung der Applikation, die sich derzeit im Beta-Stadium befindet und deren Quellcode im Laufe der Entwicklung der Forschungscommunity Open Source verfügbar gemacht werden wird, basiert auf der XML-Datenbank exist-db . Eine integrierte, browserbasierte Arbeitsumgebung erlaubt die Eingabe und Manipulation der Metadaten zu einzelnen Tafeln, die Bearbeitung der digitalen Standardzeichenliste, Upload und Verwaltung von einem oder mehreren Faksimiles einzelner Tafeln sowie die manuelle Bildsegmentierung nach Einzelzeichen und deren Annotation. Die Verwendung von XForms für die Dateneingabe einerseits und von REST-Endpoints für die Kommunikation zwischen Annotierungsoberfläche und Server anderseits ermöglichen es, Teile der Anwendung auch in verändertem technischen Kontext (bspw. vor einer relationalen Datenbank) wiederverwenden zu können. Die Applikation sieht sich damit auch als kleiner Beitrag zum Aufbau einer nachhaltigen, da modularen Forschungsinfrastruktur.

Das Datenmodell von CFDB beruht auf den aktuellen *Guidelines* der *Text Encoding Initiative* und verwendet Bestandteile der Module transcr (Kapitel 11: *Representation of Primary Sources)* für das Markup der Bild-Text-Relation sowie des gaiji-Moduls (Kapitel 5: *Characters, Glyphs, and Writing Modes)*. Die digitale Version der Standardzeichenliste, die im Zuge des Projekts erstellten Bildsegmente, Transkriptionen sowie die zugehörigen Metadaten und Annotationen werden nach Projektende sowohl eingebettet in cfdb als auch als Datenset in TEI-XML der Forschungsöffentlichkeit zur Verfügung gestellt. Die Anbindung an relevante Initiativen zur Digitalisierung und zum Korpusaufbau von Keilschrifttexten (Cuneiform Digital Library Initiative CDLI, Neo-Babylonian Cuneiform Corpus NaBuCCo) ist geplant.

CFDB ist als ein dynamisches Werkzeug für Untersuchungen zur Keilschrift konzeptualisiert, das zur einfachen Referenz in Forschung und Lehre dient und mit Blick auf die Nachhaltigkeit von Forschungsdaten entwickelt wurde. Gleichzeitig kann die Applikation leicht für analog strukturierte Forschungsprojekte in anderen Bereichen der Bild-Text-Korrelation adaptiert werden.





Bibliographie

Biggs, Robert D. (1973): "On regional cuneiform handwritings in Third Millennium Mesopotamia", in: *Orientalia* 42: 39-46.

Daniels, Peter (1995): "Cuneiform calligraphy", in: Mattila, Raija (ed.): *Nineveh, 612 BC*. The glory and fall of the Assyrian Empire. Helsinki: Helsinki University Press 81-90.

Devecchi, Elena (ed.) (2012): *Palaeography and Scribal Practices in Syro-Palestine and Anatolia in the Late Bronze Age*. Papers read at a symposium in Leiden, 17–18 December 2009 (= PIHANS 119). Leiden: Nederlands Instituut voor het Nabije Oosten

Jursa, Michael (in Vorbereitung): "Late Babylonian Epigraphy: a Case Study", in: *Current Research in Cuneiform Palaeography*. Proceedings of the Workshop organized at the 60th Rencontre Assyriologique Internationale, Warsaw 2014

Powell, Marvin A. (1981): "Three problems in the history of cuneiform writing: Origins, Direction of Script, Literacy", in: *Visible Language* 15, 4: 419-440.

Sallaberger, Walther (2001): "Die Entwicklung der Keilschrift in Ebla", in: Meyer, Jan-Waalke / Novák, Mirko / Pruß, Alexander (eds.): *Beiträge zur Vorderasiatischen Archäologie Winfried Orthmann gewidmet.* Frankfurt am Main: Johann Wolfgang Goethe-Universität 436-445.

Ziele und Aktivitäten der Arbeitsgruppe Digitale Romanistik

Schöch, Christof
christof.schoech@uni-wuerzburg.de
Universität Würzburg, Deutschland

von Ehrlich, Isabel
Isabel.vonEhrlich@lrz.uni-muenchen.de
LMU München, Deutschland

Ehrlicher, Hanno
hanno.ehrlicher@phil.uni-augsburg.de
Universität Augsburg, Deutschland

Gerstenberg, Annette
gerstenberg@zedat.fu-berlin.de
FU Berlin, Deutschland

Kraft, Tobias
kraft@bbaw.de
BBAW Berlin, Deutschland

Rißler-Pipka, Nanette
nanette.rissler@gmail.com
Universität Siegen, Deutschland

Völker, Harald
harald_voelker@gmx.net
Universität Würzburg, Schweiz

Mühlschlegel, Ulrike
Muehlschlegel@iai.spk-berlin.de
Ibero-Amerikanisches Institut, Deutschland

Einleitung

Mit dem hier beschriebenen Poster möchte die Anfang 2014 gegründete Arbeitsgruppe Digitale Romanistik, die beim Deutschen Romanistenverband (DRV) angesiedelt ist, ihre Ziele und Aktivitäten vorstellen. Zentrales Anliegen der Arbeitsgruppe ist es, die Romanistik mittelfristig und nachhaltig in den digitalen Geisteswissenschaften zu verankern. Daher sieht die AG die Poster-Präsentation bei der DHd-Tagung einerseits als Gelegenheit, die Vertreter_innen der Digitalen Geisteswissenschaften über die Existenz und Aktivitäten der Arbeitsgruppe zu informieren. Diese Aktivitäten haben zuletzt insbesondere die Themen Langzeitarchivierung von Forschungsdaten und Verbreitung von digitalen Methoden betroffen. Andererseits verfolgt die AG mit dem Poster bei der DHd-Tagung auch das Ziel, mit vergleichbaren Initiativen (andere Arbeitsgruppen, Infrastrukturen, Verbänden, Disseminations-Initiativen) zum Thema der Rolle digitaler Daten, Methoden und Tools in einzelnen Disziplinen ins Gespräch zu kommen.

Arbeitsgruppe Digitale Romanistik

In der Romanistik wird zunehmend wahrgenommen, wie der geisteswissenschaftliche Alltag seit einigen Jahren tiefgreifenden und nachhaltigen Veränderungsprozessen unterliegt, die auf der gesellschaftlich und technologisch bedingten Bedeutungszunahme und zunehmenden Selbstverständlichkeit der Verwendung von digitalen Medien, elektronisch verfügbaren Informationen und computergestützten Werkzeugen beruhen. Die weitreichende und immer stärkere Vernetzung der Forschenden, die schnellere Kommunikation von Forschungsergebnissen und die zunehmende Digitalisierung der Forschungsgegenstände sind in diesem Kontext nur drei zentrale Aspekte eines weitreichenden Prozesses, der bekanntlich häufig unter dem Stichwort der „Digitalen Geisteswissenschaften“ verhandelt wird. Der Vorstand des Deutschen Romanistenverbands hat die Einrichtung der Arbeitsgruppe „Digitale Romanistik“ gut geheißen, weil es sich hier gleichermaßen um ein wissenschaftliches und wissenschaftspolitisches Arbeitsfeld handelt. Die deutschsprachige Romanistik sollte über die Möglichkeit verfügen, sich in die zunehmend wichtigen Prozesse von Standard- und Normsetzungen in diesem Bereich einbringen zu können.

Ziele der Arbeitsgruppe

Das übergeordnete Ziel der Arbeitsgruppe „Digitale Romanistik“ ist es, die Konsequenzen der Digitalisierung in ihren Herausforderungen und Chancen für unterschiedliche Fachgebiete und Teilaspekte zu reflektieren. Dies bedeutet, die spezifische Perspektive der romanistischen Sprach-, Literatur-, Kultur- und Medienwissenschaften sowie der Fachdidaktik auf die Digitalisierung sichtbar zu machen und die Bedürfnisse der Romanistik an digitale Datenbestände, Infrastrukturen, Ausbildungsmöglichkeiten, Förderstrategien und vieles mehr zu formulieren. Die Arbeitsgruppe möchte auf diese Weise den DRV und die Romanistik als Fach dabei unterstützen, zu den anstehenden Fragen eigene Positionen weiter zu entwickeln, Empfehlungen für Zukunftsstrategien zu formulieren, sich aktiv an nationalen und europäischen Prozessen zu beteiligen sowie das bedeutende Gewicht der Romanistik in den Geisteswissenschaften auch im Bereich der digitalen Geisteswissenschaften gegenüber Forschungsförderern, Universitätsleitungen und der breiteren Öffentlichkeit deutlich zu machen. Außerdem wollen wir Ansprechpartner für Kolleg_innen sein, die mit konkreten Fragen zum Thema "Digitale Romanistik" an die Arbeitsgruppe heran treten möchten, weiterführende Informationen benötigen oder eine strategische Beratung suchen.

Erster Schwerpunkt: Langzeitarchivierung von digitalen Forschungsdaten für die Romanistik

Das erste Schwerpunktthema der Arbeitsgruppe war die Langzeitarchivierung von digitalen Forschungsdaten für die Romanistik unter den veränderten Rahmenbedingungen in den letzten beiden Jahrzehnten. Die AG hat im Winter 2014 eine Umfrage zu den aktuellen Diskussionen und Bedürfnissen der Fachwissenschaftler_innen in der Romanistik durchgeführt. Ziel war es, die romanistischen Bedürfnisse auf diesem Wege zu ermitteln, um sie in die aktuellen Strukturdebatten innerhalb der DFG, zwischen den Fachverbänden und an den Universitäten einbringen zu können. Aus den Ergebnissen der Umfrage leiten sich aus Sicht der AG Digitale Romanistik mehrere Schlussfolgerungen ab:

+ Texeditionen, Korpora und andere digital vorliegende Forschungsdaten werden intensiv und auf vielfältige Weise genutzt.
+ Es besteht Handlungsbedarf, da tragfähige Konzepte der Langzeitarchivierung fehlen.
+ Es besteht Informationsbedarf, um die zukünftigen Nutzer_nnen in den Entwicklungsprozess einzubeziehen.
+ Es sollte ein Weg zwischen "Insellösungen" (Zergliederung des Angebots) und einer klaren fachbezogenen Identität gefunden werden.

Aktueller Schwerpunkt: Verbreitung digitaler Methoden in der Romanistik

Der derzeitige Schwerpunkt der Arbeit der Arbeitsgruppe bezieht sich darauf, die Verbreitung digitaler Methoden in der Romanistik zu unterstützen. Dies geschieht insbesondere durch drei Aktivitäten: Erstens die Sammlung laufender romanistischer Forschungsprojekte und aktueller romanistischer Publikationen mit Bezug zu digitalen Methoden. Zweitens durch die Vermittlung und Durchführung von kleineren Methodenworkshops, die in ausgewählte Verfahren der digitalen Geisteswissenschaften spezifisch für ein romanistisches Publikum einführen. Und drittens durch die Publikation von Überblicksbeiträgen, die über die Forschungsaktivitäten in der digitalen Romanistik informieren und so den Austausch und die Netzwerkbildung befördern.

Wichtig scheint es aus Perspektive der Arbeitsgruppe Digitale Romanistik darüber hinaus, dass sich auch in den Fächern Kommunikationsstrukturen etablieren, über die dem digitalen Paradigma in den Fächern mehr Geltung verschafft wird und auf deren Grundlage sich die am Thema interessierten Wissenschaftlerinnen und Wissenschaftler vorbereitende Gespräche zur Anbahnung von Kooperationen führen können. Es geht also neben der Kommunikation mit den anderen Fächern auch um die Schaffung von Strukturen innerhalb der Fächer.

Weiterführende Informationen

Webseite der "Arbeitsgruppe Digitale Romanistik" des DRV Deutschen Romanistenverbandes unter:

Arbeitsgruppe Digitale Romanistik (2013-*): "Isabel von Ehrlich, Hanno Ehrlicher, Annette Gerstenberg, Tobias Kraft, Ulrike Mühlschlegel, Nanette Rißler-Pipka, Christof Schöch, Harald Völker - Kontakt: Kontakt: digitaleromanistik@gmail.com", in: *DRV Deutscher Romanistenverband* . http://www.deutscher-romanistenverband.de/der-drv/ag-digitale-romanistik/ [letzter Zugriff 08. Januar 2016].

Schwerpunktthema "Langzeitarchivierung von Forschungsdaten" unter: **Arbeitsgruppe Digitale Romanistik** (2014-*): "Langzeitarchivierung von Forschungsdaten", in: *DRV Deutscher Romanistenverband*. http://www.deutscher-romanistenverband.de/der-drv/ag-digitale-romanistik/lza/ [letzter Zugriff 08. Januar 2016].

Schöch, Christoph (2014): “Zur Einrichtung einer DRV-Arbeitsgruppe Digitale Romanistik”, in: *Mitteilungsheft des DRV* https://zenodo.org/record/11807?ln=en [letzter Zugriff 08. Januar 2016].

Entwicklung einer digitalen Brief-Edition und eines Forschungsportals zu Theodor Fontane

Seifert, Sabine
sabine.seifert@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam, Deutschland

Die digitale Edition

Das Theodor-Fontane-Archiv konzipiert eine digitale, kritische Edition aller Briefe von und an Theodor Fontane. Hierbei handelt es sich um etwa 10.000 Briefe, die bislang unvollständig und nach heutigen editionswissenschaftlichen Standards unzureichend veröffentlicht sind. Ein Großteil des Briefnachlasses befindet sich im Archiv, doch sind weitere Bestände aus anderen Institutionen und Privatbesitz zu berücksichtigen. Somit wird der gesamte, stark verstreute Briefnachlass virtuell zusammengeführt und erstmalig als ein Korpus recherchierbar, wodurch die Voraussetzung für die systematische Erforschung dieses zentralen Werkbestandes geschaffen wird.

Verschiedene editorische Herausforderungen stellen sich bezüglich der Briefe. Neben den Originalhandschriften sind mitunter mehrfache Abschriften sowie bisherige Drucke einzubeziehen. Somit werden mehrere Überlieferungsträger eines Briefes verzeichnet und in der Edition präsentiert, wodurch sich die Frage einer sinnvollen Darstellung von Überlieferungsvarianten stellt. Daneben muss eine Lösung für die digitale Umsetzung materialspezifischer Besonderheiten gefunden werden, z. B. die häufig beschriebenen Briefränder. Diese Textbestandteile sind oft seitenübergreifend und stehen in den unterschiedlichsten Winkeln zur regulären Beschreibrichtung. In der Online-Präsentation werden neben den Digitalisaten und Transkriptionen die Metadaten, die Kommentierung sowie die XML / TEI P5-Auszeichnung verfügbar gemacht. Für die individuelle Benutzbarkeit soll die Anzeige der genannten Daten flexibel anzupassen sein. Aufgrund der besonderen Schriftbildlichkeit müssen die Digitalisate drehbar und im Falle von mehreren Textzeugen soll deren parallele Anzeige möglich sein. Auf die Verwendung von Standards (z. B. XML / TEI P5) und Normdaten (z. B. GND für Personen) und die Erstellung von projektinternen Indizes (zu Personen, Institutionen, Orten, Werken, Periodika) wird besonderer Wert gelegt. So können etwa personelle und institutionelle Netzwerke,

an denen Fontane teilhatte, rekonstruiert und intertextuelle Verbindungen nachvollzogen werden.

Das Fontane-Forschungsportal

Die Edition der Briefe steht im Zusammenhang mit dem ebenfalls in der konzeptionellen Entwicklung befindlichen Fontane-Forschungsportal. Dessen Ziel ist die Präsentation der Digitalen Sammlungen des Archivs und optional anderer Bestandshalter. Die aufbewahrten Handschriften liegen digitalisiert vor, die Verknüpfung der Digitalisate mit dem technischen und bibliographischen Metadatensatz wird über den METS / MODS- bzw. METS / EAD-Standard erfolgen. Die Handschriften- und Bibliothekskataloge des Archivs, bisher intern als allegro-C- und allegro-HANS-Datenbanken geführt, werden ebenfalls über das Portal als OPACs zugänglich gemacht. Somit stellt das Portal die Verknüpfung von archivalischen Quellen- und Erschließungsdaten und von Forschungsprimärdaten her.

Neben der virtuellen Zusammenführung des zerstreuten Nachlasses ist das zweite Ziel des Portals, alle Forschungsressourcen zu Fontane schnell zugänglich bereitzustellen. Die Fontane-Aktivitäten außerhalb des Archivs sollen hier gebündelt werden und das Portal als Kommunikationsplattform für die verschiedenen Akteure dienen. Dafür werden technische Schnittstellen für Datenaustausch sowie Kooperationen mit anderen Projekten geschaffen. Doch richtet sich das Portal nicht nur an die Wissenschaft, sondern auch an die breite Öffentlichkeit, der hier ein umfassender Zugang zu Fontane, seinem Leben und Werk ermöglicht werden soll.

FuD und CMS

Die technische Umsetzung der digitalen Edition und des Forschungsportals wird im Rahmen der virtuellen Forschungsumgebung FuD („Forschungsnetzwerk und Datenbanksystem“) erfolgen, die am Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier entwickelt wurde. Die in FuD bereitgestellten Tools werden an die projektspezifischen Bedürfnisse angepasst und weiterentwickelt, etwa für die Verzeichnung unterschiedlicher Textzeugen. Diese Weiterentwicklungen sollen von anderen Institutionen nachgenutzt werden können.

Im Hintergrund stehen eine Neustrukturierung der gesamten Datenstruktur des Theodor-Fontane-Archivs und die Überführung in ein Content Management System, das ebenfalls in Zusammenarbeit mit dem Kompetenzzentrum Trier erarbeitet wird. Die Datenmodellierung des CMS, auf das Edition und Portal gleichermaßen zugreifen, muss gewährleisten, dass unterschiedliche Arten von Daten in FuD zusammengeführt, angereichert und weiterverarbeitet werden können. Hierzu gehören bibliographische Metadaten von Handschriften und Bibliotheksbeständen des Archivs, bibliographische Metadaten anderer Institutionen, Digitalisate plus deren Metadaten, Normdaten, die Transkriptionen der Edition, die TEI-Textauszeichnung, editorischen Kommentare und die erstellten Indizes. Auch werden Schnittstellen zu anderen Projekten und zu Archivdatenbanken hergestellt und der Zugang per Open Access gewährleistet.

Edition und Portal als Forschungsumgebung

Die digitale Edition und das Forschungsportal werden auf Grundlage der gemeinsamen Datenbasis so konzipiert und miteinander verbunden, dass zwischen ihnen Datenaustausch möglich wird. So wird der jeweilige Rechercherahmen erweitert, eine umfassendere Kontextualisierung der Informationen erreicht und eine digitale Arbeits- und Forschungsumgebung geschaffen, die den Erfordernissen eines Archivs in seinen Aufgaben des Sammelns, Erschließens und Langzeitarchivierens sowie den Erfordernissen einer Forschungseinrichtung gleichermaßen gerecht wird. Dies und die Vernetzung mit verschiedenen Forschungsvorhaben ermöglicht nicht nur neue Erkenntnisse für die Fontane-Forschung selbst, sondern auch zur Literatur- und Geistesgeschichte des 19. Jahrhunderts. Für weitere, interdisziplinäre und neue Fragestellungen wird die Struktur offen und flexibel angelegt. Zudem soll die entwickelte Arbeitsumgebung in ihrer Datenstruktur auch Modellcharakter für kleinere Archive, Museen und Sammlungen haben.

Im Poster werden der derzeitige Konzeptionsstand von Edition und Portal vorgestellt, Probleme und offene Fragen aufgezeigt sowie Lösungsansätze zur Diskussion gestellt.

explore.bread.AT! Die österreichische Brotkultur dialektal

Siemund, Melanie
melanie.siemund@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Der Beitrag „explore.bread.AT! Die österreichische Brotkultur dialektal“ informiert über eine Pilotstudie, die synchron und diachron Lemmata für Brot und Gebäck betrachtet. Die Studie bettet sich in das Projekt „exploreAT! exploring austrias culture through the language glass“ (Wandl-Vogt 2015) am Austrian Centre for Digital Humanities der Österreichischen Akademie

der Wissenschaften (ACDH-ÖAW) ein. Das Projekt „exploreAT!“ baut auf dem „Wörterbuch der bairischen Mundarten des Österreichischen“ (WBÖ) und der dazugehörigen „Datenbank der bairischen Mundarten des Österreichischen“ (DBÖ) auf, welche Daten aus der Zeit von 1911-1998 der (ehemaligen) Habsburger Monarchie beinhalten. Die DBÖ umfasst u. a. auf bestimmte Fragen genannte Lemmata. Eine genaue Aufstellung der Ressourcen stellen Siemund et al. (in Vorbereitung) dar.

Für die Studie werden Fragen aus den Fragebögen „30. Brot backen (III)“ sowie „31. Weißgebäck“ der WBÖ verwendet. Hierin befinden sich unter anderem Fragen nach der Benennung bestimmter Brotsorten, ob handgebackenes Brot anders genannt wird als Bäckerbrot, welche sonstigen Brotsorten bekannt sind und Verwendung finden, wie einzelne Teile sowie Eigenschaften des Brotes genannt werden, aber auch zur Brotzubereitung sowie zu Redensarten, welche in irgendeiner Form im Zusammenhang mit Brot stehen. Die verschiedenen benutzten Wortstämme der genannten Lemmata werden miteinander verglichen und Tendenzen der Verbreitung – zunächst im österreichischen Raum – aufgezeigt. Ein weiterer Untersuchungsschwerpunkt liegt in der Ermittlung der regional vorkommenden Brot- und Gebäcksorten anhand des Auftretens ihrer entsprechenden Lemmata. Im historischen Verlauf kann daraus geschlossen werden, inwiefern sich die Brotkultur innerhalb von Österreich unterscheidet und im Laufe der Zeit angepasst beziehungsweise weiter voneinander abgegrenzt hat. Ähnlich geartete Fragen anderer Sprachatlanten wie beispielsweise dem Sprachatlas von Oberösterreich (SAO) werden analysiert, sodass die Lemmata der DBÖ mit ihnen in Verbindung gesetzt werden können. Außerdem werden weitere Lexika und Atlanten wie der österreichische Volkskundeatlas (Burgstaller et al. 1959-1981) zu Rate gezogen, um die Ergebnisse in einen kulturhistorischen Kontext einbetten zu können. Langfristiges Ziel ist es, die Betrachtungen auf den deutschsprachigen Bereich sowie deren Grenzregionen auszuweiten und somit Gemeinsamkeiten, aber auch Unterschiede zwischen den einzelnen Ländern sichtbar zu machen sowie die Entwicklung mit Migrationsströmen zu vergleichen.

Technisch werden die verwendeten Lemmata aus den Datenbanken exportiert und in XML modelliert. Dafür sollen die XML-Auszeichnungssprachen TBX und TEI genutzt werden. Es ist vorgesehen die genutzten Daten in die Eweiterung von Standards für die Kombination von TBX und TEI einfließen zu lassen. Die Ergebnisse sollen in Kooperation mit Roberto Theron (Universidad de Salamanca) visualisiert werden. Einblicke in die Arbeit, aber auch praktische Realia wie Brotrezepte sollen auf dem Projektblog http://brot.linguence.de (Siemund 2015-*) veröffentlicht werden, um die Forschung nicht der Wissenschaft exklusiv zu halten, sondern auch interessierte Laien partizipieren zu lassen und sich im Citizen Science-Kontext des Projekts „exploreAT!“ einzubetten. Das Blog ist im Rahmen des 10. World Bread Day am 16.10.2015 mit einem Rezept für Kletzenbrot online gegangen und hat sich bei einer Bloggerinitiative (zorra 25.09.2015) zu selbigem Tag beteiligt, um möglichst publikumswirksam zu starten. Es ist anvisiert, die Forschungsergebnisse in verschiedene europäische Initiativen und Netzwerke einzubetten. Eigentlich nicht erwähnt werden muss, dass die Forschungsergebnisse open access unter creative commons-Lizenz zur Verfügung gestellt werden sollen.

Im Fokus des Posters liegen zum einen die technischen Vorgehensweisen, zum anderen die Methoden des Citizen Science.

Bibliographie

Burgstaller, Ernst / Wolfram, Richard / Helbok, Adolf (eds.) (1959-1981): *Österreichischer Volkskundeatlas*. Österreichische Akademie der Wissenschaften.

Siemund, Melanie (2015-*): *explore.bread.AT!*. Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities http://brot.linguence.de/ [letzter Zugriff 08. Januar 2016].

Wandl-Vogt, Eveline (2015): *exploreAT! exploring austrias culture through the language glass* https://acdh.oeaw.ac.at/dha/de/node/78 [letzter Zugriff 17. Februar 2016].

zorra (25.09.2015): "World Bread Day 2015 – Invitation / Einladung", in: *1x umrühren aka kochtopf*. Food-Blog http://www.kochtopf.me/world-bread-day-2015-invitation-einladung [letzter Zugriff 08. Januar 2016].

Visuelle Möglichkeiten der Textkollation anhand des Beispiels eines Vergleiches von Erich Kästners "Fabian" und "Der Gang vor die Hunde"

Stange, Jan-Erik
stange@fh-potsdam.de
Fachhochschule Potsdam, Deutschland

Verschiedene Auflagen literarischer Werke zu unterschiedlichen Zeiten werden häufig durch Leser als originäres und statisches Erzeugnis von Autoren wahrgenommen. In der Realität führen unterschiedliche Einflussfaktoren, wie etwa das Einwirken eines repressiven staatlichen Zensurapparates auf die

Veröffentlichung von Werken, das Nachbearbeiten durch Autoren selbst zu späteren Zeitpunkten oder in besonderem Maße auch die Wandlung durch Übersetzungen aus anderen Sprachen zu einer Varianz in den Fassungen, die von geringen Unterschieden in der Zeichensetzung oder unterschiedlichen Schreibweisen bis zu gänzlich anderen Inhalten reicht. Beispielhaft hierfür sind etwa die Unterschiede zwischen der 1931 veröffentlichten Fassung des Romans "Fabian" von Erich Kästner und der erst kürzlich wiederentdeckten und veröffentlichten Originalfassung des Romans "Der Gang vor die Hunde". Für die Deutsche Verlags-Anstalt war Kästners Originalmanuskript an vielen Stellen zu provokativ und sie stimmte einer Veröffentlichung nur zu unter der Bedingung, dass Kästner diese Stellen entschärfte. Gut lässt sich die Varianz auch anhand der zahllosen Übersetzungen von Shakespeare-Werken in andere Sprachen feststellen, die über die Jahrhunderte entstanden sind.

In der Literaturwissenschaft erlauben sogenannte Kollationstools einen Vergleich verschiedener Fassungen eines Textes. Dieses Vorgehen bezeichnet man als textkritische Methode. Sie hat zum Ziel, oben genannte Einflussfaktoren auf die Gestalt eines Textes zu identifizieren, indem man verschiedene Textfassungen Satz für Satz miteinander vergleicht. Typische Tools dieser Art sind etwa "CollateX" oder "JuXta".

Hinsichtlich des Interfaces bleiben diese Tools recht nah an der Textebene. Ohne Frage ist diese Nähe zur Textebene für eine Detailanalyse unerlässlich, erschwert es aber durch das Fehlen einer Überblicksdarstellung Zusammenhänge zwischen weiter auseinanderliegenden Textfragmenten zu erkennen, Muster zu identifizieren, die sich auf der Detailebene nicht erschließen und als unterstützende Navigationsebene, das schnelle Bewegen zwischen voneinander entfernten Positionen im Text. Auch ermöglicht der Überblick eine schnelle Einschätzung, welche Stellen des Textes für die Analyse besonders interessant sein könnten.

Abb. 1: Überblicksdarstellung von beiden Fassungen mit Detaildarstellung rechts daneben

Hierfür bietet sich eine visuelle Kodierung des Textes an, eine navigierbare interaktive Datenvisualisierung, da sie durch die Komplexitätsreduktion und die Konzentration auf bestimmte Attribute der Wörter eine komprimiertere Darstellung erlaubt.

Der Vortrag zeigt anhand des oben genannten Kästner-Beispiels neuartige Visualisierungsmöglichkeiten für Textkollationen auf und baut hierbei auf bestehenden Visualisierungsansätzen auf. Hier ist vor allem Ben Frys Projekt zu Darwins Werk "On the Origin of Species: The Preservation of Favoured Traces" zu nennen. Die explorative Datenvisualisierung erlaubt es dem Betrachter, einen Gesamtüberblick über die Entstehung des Werkes zu erhalten. Durch eine Farbkodierung für die unterschiedlichen Ausgaben ist es auf einen Blick ersichtlich, welche Abschnitte in welcher Auflage hinzugefügt wurden und welche Änderungen in den verschiedenen Ausgaben vorgenommen worden sind. Zu jedem als farbigen Strich dargestellten Absatz lässt sich in dieser Ansicht die entsprechende Textstelle anzeigen, wenn man sich mit der Maus darüberbewegt.

In der auf dem Poster vorgestellten Visualisierung der beiden Kästner-Fassungen dient die Farbkodierung dazu, den Grad der Abweichung zwischen Originalfassung und angepasster Fassung anzuzeigen. Satz für Satz können so die beiden Ausgaben miteinander verglichen werden.

Als verwandtes Projekt, das Kollationen ebenfalls auf visuellem Wege zugänglich macht, ist hier außerdem die experimentelle Visualisierung "TransVis" zu nennen, entstanden innerhalb des Projektes "Version Variation Visualization", die es erlaubt, die insgesamt 37 deutschen Übersetzungen von Shakespeares Othello in einem visuellen Interface nebeneinanderzustellen und zu vergleichen. Ähnlich wie in diesem Projekt, gestattet es auch die Kästner-Visualisierung, Satz für Satz zu vergleichen, wobei visuell hervorgehoben wird, welche Teile des Satzes geändert wurden bzw. ob der Satz gänzlich ersetzt wurde.

Eine weitere Ansicht ist geplant, die für einen durch den Nutzer zu definierenden Teilbereich des Buches eine typographische Übersicht bietet über die in diesem Bereich vorgenommenen Änderungen. Auf diese Weise kann schnell eingeschätzt werden, ob die Änderungen in einem Abschnitt einen thematischen Fokus haben (In Kästners Werk sind es häufig Formulierungen mit erotischem Bezug).

Außerdem soll im Verlauf des Projektes eine weitere Variante entstehen, die sich nicht nur an Literaturwissenschaftler in Form einer analytischen Datenvisualisierung richtet, sondern auch an gewöhnliche Leser, bei denen zunächst der ununterbrochene Konsum des Textes im Vordergrund steht, für die aber zusätzliche Informationen, wie die Veränderung eines Textabschnittes über mehrere Fassungen hinweg auch von Interesse sein können.

Die Textdaten für die Visualisierung basieren auf den beiden E-Book-Fassungen der Texte "Der Gang vor die Hunde" (2013) und Fabian (2010). Unterschiede zwischen den beiden Fassungen wurden mithilfe des Levenshtein-Algorithmus in Java vorberechnet und dann als Datensatz in einer Webvisualisierung verwendet, die mit HTML, CSS, Javascript und der Visualisierungsbibliothek D3 erstellt wurde.

Bibliographie

Cheesman, Tom / Flanagan, Kevin / Thiel, Stephan (2013): "Translation Array Prototype 1: Project Overview" www.delightedbeauty.org/vvv [letzter Zugriff 15. Oktober 2015].

CollateX : http://collatex.net [letzter Zugriff 12. Februar 2016].

D3.js: http://d3js.org [letzter Zugriff 12. Februar 2016].

Fry, Ben (2009): *On the Origin of Species*. The Preservation of Favoured Traces http://benfry.com/traces/ [letzter Zugriff 12. Februar 2016].

Juxta Collation Software for Scholars: http://www.juxtasoftware.org/ [letzter Zugriff 12. Februar 2016].

Kästner, Erich (1931): *Fabian*. Die Geschichte eines Moralisten". Zürich: Atrium Verlag.

Kästner, Erich (1931): *Der Gang vor die Hunde*. Zürich: Atrium Verlag.

Digitale Dokumentation des Kulturerbes im internationalen Verbund. Das Projekt Forschungsinfrastruktur Kunstdenkmäler in Ostmitteleuropa (FoKO)

Stanicka-Brzezicka, Ksenia
ksenia.stanicka@herder-institut.de
Herder-Institut für historische Ostmitteleuropaforschung – Institut der Leibniz-Gemeinschaft

FoKO ist ein internationales Verbundprojekt, das den Aufbau einer interaktiven kunsthistorischen Forschungsinfrastruktur zum Ziel hat, mit welcher Methoden, Konzepte und Produkte der digitalen Kunstgeschichte angewendet und erprobt werden sollen. In den Fokus gerückt werden dabei die bislang noch unzureichend gewürdigten spezifischen Leistungen der Kunstproduktion im östlichen Mitteleuropa, einer Region von komplexer historischer Dynamik. Mit der transnationalen Zusammenführung von Dokumentationsdaten und Bildbeständen soll der internationalen Forschung ein Wissensportal zur Verfügung gestellt werden, in dem die vielfältigen Verflechtungen der Kunstentwicklung in Ostmitteleuropa im Zeitraum von 1000 bis 1800 ebenso deutlich werden, wie die konkurrierenden wissensgeschichtlichen Bezugnahmen. Die Forschungsinfrastruktur versteht sich als Basis, um Kunstdenkmäler mit ihren Funktionen sowie künstlerische Gattungen und Phänomene adäquat analysieren und sinnvoll im europäischen Gesamtzusammenhang verstehen zu können. Ermöglicht wird damit der vergleichende Blick etwa auf Aspekte der Stil- und Tradierungsgeschichte, des Kulturtransfers, der Auftraggeberschaft sowie der Netzwerkbildung von Künstlern und Baumeistern.

Das Vorhaben ist inhaltlich eng an das am Geisteswissenschaftlichen Zentrum Geschichte und Kultur Ostmitteleuropas an der Universität Leipzig angesiedelte Publikationsprojekt „Handbuch zur Geschichte der Kunst in Ostmitteleuropa“ angelehnt. Für Bebilderung des Handbuches wird in Rahmen des FoKO-Projektes ein neues, standardisiertes Bildmaterial über Fotokampagnen generiert. Zudem basiert FoKO auf den Bildmaterialien der Sammlungen der Partnerinstitutionen in Deutschland: des Herders-Instituts (Gesamtkoordination), des Deutschen Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg, sowie der Auslandspartner: die kunsthistorischen Institute der polnischen, slowakischen und ungarischen Akademien der Wissenschaften.

Das Projekt greift mehrere von Herausforderungen auf, die heutzutage vor den Dokumentationsinstitutionen wie Bildarchiven stehen: auf internationalem Forschungsstand beruhende Dokumentation von herausragenden Kunstdenkmälern, hochwertige Digitalisierung von historischen Bildmaterialien und Erzeugung aktueller Fotografien des Kulturerbes sowie Bereitstellung valider Daten im Internet in einem zeitgemäßen, nutzerorientierten Angebot für Wissenschaft und Öffentlichkeit. Dieses Vorhaben steht auch in Zusammenhang mit der transnationalen Vernetzung von Institutionen und der Kooperation bei der Erstellung von „traditionellen“ wie auch digitalen Erzeugnissen nach internationalen Standards und nicht zuletzt mit dem sich wandelnden Umgang mit Bildquellen in der Geschichts- und Kulturwissenschaft.

Ontologie CIDOC CRM (ISO 21127) als semantisches Rückgrat:

Das CIDOC Conceptual Reference Model ist definiert als formale Ontologie, die die Integration, Vermittlung und den Austausch verschiedenartig strukturierter Informationen des kulturellen Erbes unterstützt. Dies entspricht dem Ziel des Projektes: der Integration der Daten aus verschiedenen Einrichtungen. Folglich definiert

das CRM im Wesentlichen die zu Grunde liegende Semantik von Datenbankschemata und Strukturen von Dokumenten, d. h. es beschreibt die expliziten und impliziten Begriffe, die zur Kulturerbe- und Museumsdokumentation benutzt werden, sowie deren Beziehungen. Die Modellierung dieser Begriffe und Beziehungen erfolgt ereignisorientiert. Darüber hinaus als semantische Grundgerüst kann die Ontologie von Domain- Ontologien und Normdateien erweitert werden (wie GND , Getty AAT etc.)

System: Wissenschaftliche Kommunikations-Infrastruktur

WissKI ist im Projekt im Einsatz als Datenbank und als virtuelle Forschungsumgebung. Das Datenmodel berücksichtigt die Kunstobjekte, wie auch die Fotografien, die als separate Entitäten aufgefasst und gleichermaßen in ihrer Gegenständlichkeit, unter Berücksichtigung ihres Spezifikums der technischen Vervielfältigung beschrieben werden. In der Datenmodellierung wurden – mit CRM semantischen Definitionen und Begriffserklärungen eindeutige und trennscharfe Beziehungen zwischen Fotografien und dargestellten Entitäten gebildet. Auch weitere Masken wurden entworfen – z. B. für die Personen, historische Ereignisse, die in einem Bezug zu den Kunstobjekten oder Fotografien stehen. Für alle diese Beziehungen wurden semantische Pfade nach CIDOC CRM entworfen, wobei teilweise konnte man aus den anderen WissKI-Anwendungen schöpfen, teilweise müssten die Pfade neu konfiguriert werden. Benutz werden auch weitere WissKI-Tools: differenzierte Text und Kommentarfelde, semantische Annotation der Texten, zoom. Bei der Erfassung von verschiedenen Kategorien von Entitäten sind solche Funktionalitäten wie automatische Verknüpfungserzeugung, Datenkapselung, Link-Block oder automatische Titelgenerierung sehr hilfreich. Eingebunden wurden Normdaten (AAT, GND, polnischer Thesaurus für kulturelles Erbe – das Portal wird in den wichtigsten Kategorien in mehreren Sprachen durchsuchbar). Die Fotos werden in der Auflösung von 72 dpi und nach CC BY SA 3.0 publiziert. Für höhere Auflösung wird der Nutzer automatisch an das zuständige Institution umgeleitet.

Ein wichtiges und aktuelles Thema im Projekt ist ebenfalls die Auslegung von Urheber- und Nutzungsrechten der genutzten und erzeugten Bildmaterialien, der Deriverate, digitalen Inhalte und Metadaten, und zwar zwischen den Partnern wie auch im Hinblick auf die Onlinestellung für die allgemeine Nutzung und geplanter Transfer der Bilder und der Metadaten in die Europeana und die Deutsche Digitale Bibliothek am Ende der Projektlaufzeit 2017.

Erste Projektergebnisse:

Entwicklung der Datenbankstruktur

Aufbau der Datenbank

Klärung von Rechten und Lizenzen für die Veröffentlichung der digitalen Fotografien, Metadaten und anderen digitalen Inhalten

Erstellung von Objektlisten (Denkmäler)

Vorläufige Dokumentation in der Datenbank von 1500 Denkmälern, 400 Fotografien (mit circa 800 Manifestationen wie Negative, Abzüge, Dias), 160 Personen , 500 Ortschaften

Verwaltung eines mehrsprachigen Thesaurus mit den ersten 200 Termini gemappt auf AAT sowie polnischen Thesaurus

Bibliographie

Bentkowska-Kafel, Anna / Denard, Hugh / Baker, Drew (2012): *Paradata and Transparency in Virtual Heritage*. Farnham / London: Ashgate.

Cameron, Fiona / Kenderdine, Sarah (eds.) (2007): *Theorizing Digital Cultural Heritage*. A Critical Discourse. London / Cambridge: MIT Press.

Caraffa, Constanza (ed.) (2011): *Photo Archives and the Photographic Memory of Art History*. Berlin-München: Deutscher Kunstverlag.

Caraffa, Constanza (2011): "'Wenden!' Fotografien in Archiven im Zeitalter ihrer Digitalisierbarkeit: ein 'materialturn'", in: *Rundbrief Fotografie* 18, 3: 8-15.

Caraffa, Constanza (ed.) (2009): *Fotografie als Instrument und Medium der Kunstgeschichte*. Berlin-München: Deutscher Kunstverlag.

Edwards, Elizabeth / Hart, Janice (2004): *Photographs Objects Histories*. On the Materiality of Images. London / New York: Routledge.

Herden, Elżbieta / Seidel-Grzesińska, Agnieszka / Stanicka-Brzezicka, Ksenia (eds.) (2012): *Dobra kultury w Sieci*. Wrocław: Wydaw. Uniwersytetu Wrocławskiego.

Jäger, Jens (2009): *Fotografie und Geschichte*. Frankfurt am Main / New York: Campus.

Jäger, Jens / Knauer, Martin (eds.) (2009): *Bilder als historische Quellen?* Dimension der Debatten um historische Bildforschung. Paderborn: Wilhelm Fink.

Jagschitz, Gerhard (1991): "Visual History", in: *Das audiovisuelle Archiv* 29 / 30: 23-51.

Kalay, Yehuda E. / Kvan, Thomas / Affleck, Janice (2008): *New Heritage. New Media and Cultural Heritage*. London: Routledge.

Miller, Maria / Wornbard, Małgorzata (2009): "Fotografie w zbiorach cyfrowych. Problemy z opracowaniem formalnym i rzeczowym na przykładzie Biblioteki Cyfrowej Politechniki Warszawskiej", in: *Przegląd Biblioteczny* 77, 2: 201-218.

Paul, Gerhard (ed.) (2006): *Visual History*. Ein Studienbuch. Göttingen: Vandenhoeck & Ruprecht.

Płoszajski, Grzegorz (ed.) (2008): *Standardy w procesie digitalizacji obiektów dziedzictwa kulturowego*. Warszawa: Biblioteka Główna Politechniki Warszawskiej.

Seidel-Grzesińska, Agnieszka / Stanicka-Brzezicka, Ksenia (eds.) (2014): *Obraz i metoda* (= Cyfrowe Spotkania z Zabytkami 4). Wrocław.

Seidel-Grzesińska, Agnieszka / Stanicka-Brzezicka, Ksenia (2015): „Wielojęzyczne słowniki hierarchiczne w dokumentacji muzealnej w Polsce", in: *Muzealnictwo*.

Narodowy Instytut Muzealnictwa i Ochrony Zbiorów 56: 169-181.

Sztompka, Piotr (2006): *Socjologia wizualna. Fotografia jako metoda badawcza*. Warszawa: Wydawnictwo Naukowe PWN.

Zeidler-Janiszewska, Anna (2006): „Visual Culture Studies czy antropologicznie zorientowana Bildwissenschaft? O kierunkach zwrotu ikonicznego w naukach o kulturze", in: *Teksty Drugie* 100, 4: 9-30.

Kein Gedanke ohne Gedächtnis: Aspekte der Kooperation zwischen digitaler Geisteswissenschaft und BAM-Institutionen

Steiner, Elisabeth
elisabeth.steiner@uni-graz.at
Universität Graz, Österreich

Koch, Carina
carina.koch@uni-graz.at
Universität Graz, Österreich

Einleitung

Seit 2014 arbeitet das Projekt „Repositorium Steirisches Wissenschaftserbe" an der digitalen Erschließung, Archivierung und Veröffentlichung von für den Regionalraum Steiermark bedeutsamem Quellenmaterial. Das Projekt vereint nicht nur zahlreiche Konsortialpartner sondern auch unterschiedliche Quellengattungen, von Museumsobjekten, über Handschriften, bis zu Postkarten oder historischen Glasdias. Die homogene Beschreibung dieser Mischung aus objekt-, bild- und textzentrierten Ressourcen ist eine Herausforderung. Das Poster zeigt die ersten Zwischenergebnisse des Projektes. Dabei wird nicht nur der Kernteil der digitalen Erschließung, Langzeitarchivierung und Dissemination vorgestellt, sondern auch Aspekte, denen im Vorhinein oft weniger Beachtung geschenkt wird, die aber für einen erfolgreichen Projektverlauf ebenso zentral sind.

Fachlich-inhaltliche Aspekte

In einem ersten Schritt wurden disziplinenspezifische Fragestellungen an Einzelbestände eruiert, sowie die Funktionalitäten der geplanten gemeinsamen Webplattform und die dafür benötigten Daten mit den Partnern festgelegt. Als größte Herausforderung aus der Sicht der Digital Humanities ergab sich dabei die Homogenisierung der Metadaten: Die Inhalte der unterschiedlichen Objekttypen werden im Projekt zunächst mit domänenspezifischen und international anerkannten XML-Metadatenstandards (etwa TEI, LIDO, EAD) beschrieben. Die Datenerfassung und Erschließung erfolgt dabei nach eigens entworfenen bestands- und disziplinenübergreifenden Richtlinien, um eine möglichst konsistente Datenbasis zu schaffen. Das umfasst auch die Verwendung von kontrollierten Vokabularien und Thesauri (z. B. Geonames, AAT, GND) für die semantische Anreicherung der Quellen. Für das gemeinsame Portal werden aus dieser Datenbasis festgelegte Metadaten-Kernkategorien (Person, Ort, Zeit, Objekttyp, Medientyp und Kurzbeschreibung) extrahiert und auf Dublin Core und Europeana Data Model-Kategorien gemappt. Dieser Ansatz soll der Diversität der Quellen Rechnung tragen und eine qualitativ hochwertige Einzelbeschreibung mit generischen Beschreibungskategorien verbinden, die die Grundlage für einen gemeinsamen Suchraum bilden. Die Langzeitarchivierung der digitalen Forschungsdaten erfolgt in bereits vorhandenen institutionellen Repositorien der universitären Partner. [1]

Finanzielle, organisatorische und rechtliche Aspekte

Kooperationen zwischen Universitäten und Gedächtnisinstitutionen sind oft schwierig formal abzusichern. Neben der inhaltlichen Bereitschaft zur Partnerschaft müssen sich alle Beteiligten zunächst auch auf ein gemeinsames Geschäftsmodell einigen. Während an den wissenschaftlichen Forschungseinrichtungen immer mehr auf Open Access gesetzt wird, sind Gedächtnisinstitutionen traditionell stärker auf persönliche BesucherInnen fokussiert, und sehen die Onlinepräsentation ihrer Sammlungen als potentielle Konkurrenz zu analogen Ausstellungen. Da die meisten Partnerinstitutionen öffentlich gefördert werden und öffentliches Gut verwahren, ist ein freier Zugang zu den Online-Ressourcen jedoch wünschenswert. In das entstehende gemeinsame Projektportal werden daher per Definition nur kostenfrei zugängliche digitale Objekte aufgenommen.

Doch nicht nur finanzielle Interessen können für einen beschränkten Zugang zu Ressourcen verantwortlich sein sondern auch rechtliche Aspekte. Archive müssen gesetzliche Sperrfristen beachten, Fragen des Datenschutzes sowie Persönlichkeitsrechtes spielen oft bei Korrespondenzen und Bildsammlungen eine Rolle. Selbstverständlich müssen alle Institutionen auf die Urheberrechte der WerkschöpferIn Rücksicht nehmen.

Während fachlich-inhaltliche Nachhaltigkeit durch bereits vorhandene Infrastruktur und entsprechendes Know-How bei den jeweiligen Partnern relativ mühelos umgesetzt werden kann, ist finanzielle und organisatorische Nachhaltigkeit schwieriger sicherzustellen. Gerade bei institutionsübergreifenden Projekten mit begrenzter Laufzeit stellt dies ein Problem dar. In diesem Fall kann die Projektarbeit nur als Anschubfinanzierung für den Aufbau einer Infrastruktur und von Arbeitsabläufen verstanden werden, die darauffolgend in den Regelbetrieb übergehen sollten.

Kooperativ-kommunikative Aspekte

Differenzen zwischen Wissenschafts- und Gedächtnisinstitutionen können nur durch kontinuierliche Gespräche überbrückt, Ziele und Lösungsansätze nur gemeinsam erarbeitet und umgesetzt werden. Spezielle Begrifflichkeiten, die auf divergierende disziplinäre Hintergründe zurückzuführen sind, sind von Beginn an explizit zu machen. Vorstellungen und Erwartungen bezüglich des Projekt-Ergebnisses oder der visuellen Umsetzung unterscheiden sich häufig gravierend. Das zentrale Augenmerk der Digital Humanities muss hier auf der Vermittlerrolle liegen: Die Kommunikation mit den unterschiedlichen Institutionen, Disziplinen und FachwissenschaftlerInnen erfordert oft Fingerspitzengefühl, da nicht nur die Kooperation selbst hinterfragt wird, sondern auch die Vorteile der digitalen Komponente des Projektes.

Ausblick

Speziell in den Geisteswissenschaften darf nicht vergessen werden, dass viele der zentralen Forschungsobjekte in BAM-Institutionen gepflegt und aufbewahrt werden. Unter diesem Gesichtspunkt ist die enge Kooperation zwischen der (digitalen) Wissenschaft und den Gedächtnisinstitutionen nicht nur wünschenswert sondern unabdinglich. Die inhaltliche, organisatorische und technische Vernetzung birgt Synergien, die im Idealfall über den begrenzten Zeitraum eines Projektes hinausgehen und in nachhaltigere Formen der Kooperation überführt werden. Beide Bereiche profitieren von dieser Zusammenarbeit und können damit nicht zuletzt der Öffentlichkeit auch einen Teil der eigenen Geschichte und Kultur besser vermitteln und zugänglich machen.

Nachdem im ersten Projektabschnitt die Grundlagen für eine erfolgreiche Zusammenarbeit gelegt und die Daten der Partner nach den festgelegten Richtlinien erfasst und angereichert wurden, folgt in der nächsten Phase die technische Umsetzung der Projektplattform und des Suchportals.

Notes

1. Die Karl-Franzens-Universität Graz und die Kunstuniversität Graz betreiben jeweils eine auf der open source Software FEDORA Commons basierende Archivierungslösung: GAMS (http://gams.uni-graz.at) und PHAIDRA (http://phaidra.kug.ac.at).

Bibliography

GAMS (2014-*): *GAMS: Geisteswissenschaftliches Asset Management System*. Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Karl-Franzens-Universität Graz http://gams.uni-graz.at/ [letzter Zugriff: 08. Januar 2016].

Phaidra (o. J.): *Phaidra: Permanent Hosting, Archiving and Indexing of Digital Resources and Assets*. Universität Wien, Bibliotheks- und Archivwesen https://phaidra.kug.ac.at/ [letzter Zugriff: 08. Januar 2016].

Digital Zusammenwachsen: Forschungsdaten-management im Forschungsverbund MWW

Steyer, Timo
steyer@hab.de
Forschungsverbund MWW

Koglin, Lydia
lydia.koglin@klassik-stiftung.de
Forschungsverbund MWW

Fritz, Steffen
fritz@dla-marbach.de
Forschungsverbund MWW

Das Poster stellt ein Konzept für das digitale Zusammenwachsen heterogener Datenbestände dreier Gedächtnisinstitutionen zur Diskussion vor n. Im Zentrum steht das Forschungsdatenmanagement, welches neben Aspekten der Datenmodellierung und Beschreibung von heterogenen Datenbeständen über Metadaten, auch die Präsentation, Benutzung und Langzeitarchivierung der digitalen Beständen umfasst.

Der durch das BMBF geförderte Forschungsverbund Marbach Weimar Wolfenbüttel , bestehend aus dem Deutschen Literatur Archiv Marbach, der Klassik Stiftung Weimar und der Herzog August Bibliothek Wolfenbüttel, entwickelt seit 2014 einen gemeinsamen

virtuellen Forschungsraum. Auf diesem sollen die einzigartigen wissenschaftlichen Bestände der drei Einrichtungen gemeinsam präsentiert und der Forschung zur Verfügung gestellt werden. Durch Methoden der Digital Humanities werden neue Zugänge zu den digitalen Forschungsdaten des Verbundes geschaffen und die Implementierung innovativer Tools wird neue Forschungsfragen an das Material ermöglichen. Dabei wird die Sicherung der wissenschaftlichen Erträge durch die Etablierung eines verlässlichen Speichers garantiert, der überdies die Langzeitarchivierung digitaler Objekte übernehmen wird. Langfristig soll der Forschungsraum zu einem universellen Rechercheinstrument, einem virtuellen Arbeitsplatz und einem Repositorium für Forschungsergebnisse für die Verbundeinrichtungen ausgebaut werden.

Die MWW-Forschungsumgebung ist dabei als Teil der digitalen Infrastrukturen der drei Einrichtungen zu sehen, über den die verteilten Ressourcen aggregiert werden. Der Forschungsraum soll dezidiert so gestaltet werden, dass er mit anderen Forschungsumgebungen kompatibel ist, aber trotzdem die Spezifika der Bestände der Verbundeinrichtungen berücksichtigt und den digitalen Geisteswissenschaften zur Verfügung stellt.

Funktionsumfang und Schnittstellen des Forschungsraums orientieren sich dabei an den historischen Sammlungsschwerpunkten der Verbundeinrichtungen. Im Gegensatz zu universell ausgelegten Forschungsinfrastrukturen handelt es sich daher bei dem MWW-Forschungsraum um eine bestandsbezogene Infrastruktur. Dies hat zur Folge, dass nicht die Produktion neuer Forschungsdaten und Workflows im Vordergrund steht, sondern die Integration von bereits vorhandenen Daten und die Verknüpfung bereits bestehender Arbeitsprozesse. Für die Verwirklichung dieser Ziele bedarf es eines komplexen Forschungsdatenmanagements, um den gemeinsamen Zugang zu ermöglichen. Die Herausforderung besteht in der Bildung eines interoperablen Datenpools, welcher die technische Heterogenität der Daten und die unterschiedlichen Workflows bei der Produktion, Präsentation und Sicherung der jeweils eigenen Forschungsdaten für den Forschungsraum meistert. Die Datenbasis bildet die Modellierung und Beschreibung der digitalen Informationseinheiten über eine Metadatengrammatik. Neben den zentralen Nachweissystemen mit standardisierten Metadaten gibt es in den Einrichtungen diverse Fach- und Spezialdatenbanken, die technisch und inhaltlich häufig Insellösungen darstellen: Daten wurden und werden kontinuierlich erhoben oder stellen Ergebnisse von abgeschlossenen Projekten dar. Das Ziel ist es, aus diesen inhaltlich hochwertigen aber technisch nicht aktuellen Standards entsprechenden digitalen Daten der Verbundeinrichtungen, Forschungsdaten zu generieren, also gut strukturierte, über etablierte Metadatenstandards erschlossene und somit auch leicht zu prozessierende Daten.

In der Regel ist aber bei der Erhebung dieser Forschungsdaten nicht ein Data-Reuse-Szenario konzipiert worden ist. Zu diesem Szenario gehört aber nicht nur die Aufbereitung der Forschungsdaten für den MWW-Forschungsumgebung, sondern auch deren Bereitstellung für externe Dienste über standardisierte Schnittstellen. Des Weiteren wird ein aktiver Austausch mit anderen Forschungsumgebungen angestrebt.

So sind auch die langfristige Sicherung und Nachhaltigkeit dieser Forschungsdaten ein Bestandteil der digitalen Infrastruktur. Dabei werden nicht nur die Daten des Forschungsraums, sondern alle langzeitarchivwürdigen Daten der Verbundeinrichtungen in einem verlässlichen Speicher archiviert.
Das Modell eines verteilten Speichers von drei Gedächtniseinrichtungen und die sich dadurch ergebenen Synergieeffekte sollen hier verdeutlicht werden. Damit begegnen die Verbundeinrichtungen auch der Herausforderung der stetig wachsenden Zahl an Forschungsdaten und dem Wunsch der Wissenschaftler nach langfristiger Verfügbarkeit ihrer Daten.

Ergänzt wird der virtuelle Forschungsraum durch eine „Workbench“ zur Bearbeitung und Analyse der Forschungsdaten. Dabei wird jedoch nicht primär eine Neuentwicklung von Tools das Ziel sein. Vielmehr wurden durch Umfragen mit Digital Humanists und „traditionellen“ Geisteswissenschaftlern der Bedarf an möglichen DH-Tools und -Funktionalitäten ermittelt. Die Auswertung zeigte, dass bereits eine Vielzahl an Tools existiert, welche die Wissenschaftler für ihrer Arbeit unterstützen. Allerdings scheint es, dass diese Tools entweder nicht bekannt sind oder ihre Nutzung als zu kompliziert erachtet wird, um diese nebenbei zu erlernen. Daher konzentriert sich der MWW-Forschungsraum auf die Integration bereits vorhandener und etablierter freier DH-Dienste bzw. Funktionalitäten, welche zusammen mit umfangreichen Tutorials und Best-Practice-Beispielen im Forschungsraum angeboten werden sollen. Gerade im Bereich der Tools ergeben sich umfangreiche Kooperationspotentiale mit anderen Forschungsumgebungen, da der Forschungsverbund über eine große Menge an wissenschaftlich hochwertigen Datenmaterial verfügt, dafür über bisher über nur sehr wenige Tools für die Analyse und Visualisierung der Daten.

Im Zentrum der Präsentation steht das Zusammenspiel der soeben skizzierten drei Komponenten Metadatengrammatik, virtueller Forschungsraum und verlässlicher Speicher. Die Projekte haben die Konzeptionsphase hinter sich und würden die DHd-Tagung dazu nutzen, das entwickelte Konzept für die erste Förderphase, die Mitte 2018 endet, zu präsentieren und zur Diskussion zu stellen.

MMW Forschungsverbund (2013-*): *Forschungsverbund Marbach Weimar Wolfenbüttel* http://

www.mww-forschung.de [letzter Zugriff 15. Februar 2016].

Wie verhalten sich Aktionäre bei Unternehmenszusammenschlüssen? Modellierung sprachlicher Muster zur Analyse treibender Faktoren bei der Berichterstattung

Stotz, Sophia
stotz@hni.upb.de
Universität Paderborn, Deutschland

Geierhos, Michaela
geierhos@hni.upb.de
Universität Paderborn, Deutschland

Welche Informationen über Unternehmenszusammenschlüsse werden in Zeitungsnachrichten vermittelt, und wie können diese Informationen automatisch extrahiert werden? Dies soll am Beispiel des Verhaltens von Aktionären während eines Zusammenschlusses ermittelt werden. Dazu werden die wichtigsten Aussagen über das Votum der Aktionäre im Hinblick auf eine automatische Erkennung sprachlich analysiert. Im Fokus stehen dabei die Berichte über Aktionärsabstimmungen hinsichtlich der Annahme bzw. Ablehnung eines Übernahmeangebots. Dabei gilt es, die folgenden beiden Herausforderungen zu meistern:

Identifikation der Treiber

Bei der vorliegenden Fragestellung geht es darum, die sprachliche Gestaltung der Rolle der Aktionäre bei Unternehmenszusammenschlüssen in der Presse zu analysieren. Die Aktionärsabstimmung ist eingebettet in den Kontext unterschiedlicher Ereignisse, die Teil eines Zusammenschlussversuchs sind. Da Unternehmenszusammenschlüsse erhebliche Auswirkungen sowohl auf die Beschäftigten, als auch auf andere Unternehmen (insbesondere Konkurrenten), Aktionäre und die Verteilung von Spitzenposten haben, wird in Zeitungen ausführlich darüber berichtet. Medien haben zwar keinen direkten Einfluss auf wirtschaftliche Prozesse, sie können jedoch die Meinung von primären Adressaten, wie z. B. Aktionären, beeinflussen und nehmen daher dennoch eine wichtige Rolle ein (vgl. Palmieri 2014: 71f.). In erster Linie veröffentlichen sie Mitteilungen von Unternehmen und andere relevante Dokumente und tragen so zur Verbreitung von Schlüsselinformationen bei. Im Vergleich zu Eigenwerbung oder selbst verfassten Mitteilungen wirken Berichte von Journalisten unabhängiger und damit glaubwürdiger. Der „Vertrauenswürdigkeitsvorsprung journalistischer Berichterstattung“ (Hoffjann 2014: 674) führt dazu, dass diese besonders wichtig für die Reputation eines Unternehmens ist (vgl. Hoffjann 2014: 673 f.).

Modellierung des sprachlichen Variantenreichtums

Verschiedene Akteure beeinflussen den Verlauf eines Zusammenschlussversuchs und bilden untereinander ein komplexes Beziehungsnetzwerk. So kann beispielsweise im Fall einer feindlichen Übernahme der Aufsichtsrats des Zielunternehmens seinen Aktionären empfehlen, keine Aktien zu verkaufen und dadurch die Übernahme gefährden. Die Haltung der Aktionäre, aber auch die des Kartellamts, kann einen Zusammenschluss zum Scheitern bringen. Um diese Prozesse zu modellieren, muss das Korpus im Hinblick auf die Akteure, ihre Entscheidungen und Meinungen untersucht werden. Zur Extraktion von Ereignissen besteht bereits eine große Bandbreite an Fachliteratur, die grob in maschinelle, musterbasierte und hybride Vorgehensweisen eingeteilt werden kann (Hogenboom et al. 2011). Auch zur automatischen Bestimmung des faktischen Status von Ereignissen sowie zur automatischen Analyse von Wirtschaftsnachrichten hat es bereits einige Ansätze gegeben (z. B. Saurí / Pustejovsky 2012; Nassirtoussi 2014). Im vorliegenden Beitrag soll am Beispiel von Aktionärsabstimmungen gezeigt werden, wie semi-automatisch ermittelte morpho-syntaktische Muster den Kontext des Zusammenschlusses semantisch mittels lokaler Grammatiken (Gross 1997) modellieren können. Die mithilfe der Muster gewonnenen Informationen können z. B. für eine semantische Suchmaschine genutzt werden, um Fragen der Art „Wie oft waren Aktionäre für das Scheitern einer Fusion in den letzten 2 Jahren verantwortlich?“ oder „Wer trug zum Scheitern der Übernahme von Tele Columbus durch Kabel Deutschland bei?“ beantworten zu können.

Datenbasis

Um die von deutschsprachigen Wirtschaftsnachrichten erwähnten Einflüsse der Aktionäre auf einen Zusammenschluss (und umgekehrt) zu erfassen, wurden mithilfe des COSMAS-Tools (vgl. Institut für Deutsche Sprache) 6784 Sätze zusammengestellt, die jeweils die Schlüsselwörter („Übernahme“ ODER „Fusion“) sowie

„Aktionäre“ enthalten. Die über das COSMAS-Tool verfügbaren Korpora bestehen hauptsächlich aus den Archiven regionaler und überregionaler deutschsprachiger Zeitungen. Mit der Keyword-Auswahl sollen die wichtigsten Ereignisse identifiziert werden, die während eines Zusammenschlusses im Zusammenhang mit den Aktionären stehen. Folgende Tabelle zeigt einen Auszug aus den manuell erstellten relevanten Themen sowie die dazugehörigen sprachlichen Muster:

Phasen eines Unternehmens-zusammen-schlusses	**Schlüsselwort (in Kombination mit „Aktionäre“ und „Fusion“ bzw. „Übernahme“)**	**Auftretens-häufigkeit**	**Korrekt bezüglich der Phase des Zusammenschlusses**	**Konfidenzmaß**
Zustimmung	*zugestimmt*	425	389	91,53%
	gebilligt	78	72	92,30%
Geplante Abstimmung	*zustimmen*	349	336	96,28%
Absage der Aktionäre	*abgelehnt*	25	16	64,00%

Tab. 1: Konfidenzmaß ausgewählter Keywords bzgl. bestimmter Phasen

Korpusverarbeitungssystem und Modellierungswerkzeug

Da Medienberichte in Wirtschaftsnachrichten wiederkehrende Sprachmuster für die Ankündigung gescheiterter und erfolgter Zusammenschlüsse benutzen, ist der Ansatz der lokalen Grammatiken für diese Aufgabe vielversprechend (Gross 1997). Lokale Grammatiken erlauben es, morpho-syntaktische sowie semantische Eigenschaften der Sprache zu berücksichtigen. Zur Implementierung verwende ich das Korpusverarbeitungssystem Unitex , das an der Université Marne-la-Vallée entwickelt wurde.

Am Centrum für Informations- und Sprachverarbeitung wurde in München ein sehr umfangreiches deutschsprachiges Lexikon aufgebaut, das zahlreiche syntaktische und semantische Informationen mit einschließt und sich dadurch erheblich von vergleichbaren Systemen absetzt (Guenthner / Maier 1994). Neben der Verfügbarkeit dieses Lexikons bietet Unitex den Vorteil, dass der Benutzer selbst Regeln zur Erkennung von Named Entities schreiben und dabei eine Vielzahl an morphologischen und syntaktischen sowie semantischen Einschränkungen einbauen kann. Für die vorliegende Studie wurde z. B. ein Lexikon mit Organisationsnamen mit über 335000 Einträgen eingesetzt (Mallchok 2004).

Lokale Grammatiken zur Differenzierung bestimmter Phasen

Für die Auswahl geeigneter und besonders charakteristischer Schlüsselpassagen werden zunächst die einschlägigen verbalen Wendungen (z. B. „zustimmen“, „grünes Licht geben“) im Textkorpus ermittelt. Da die isolierte Erkennung der Prädikate allein oft nicht ausreicht, um eine Aussage korrekt zu extrahieren, muss der jeweilige Kontext berücksichtigt werden. Es geht also darum, die syntaktische Distributionsklasse des Prädikats (Anzahl und Form der Komplemente) zu ermitteln (vgl. Nagel 2005: 16). Das Verb „genehmigen“ erfordert im untersuchten Bereich beispielsweise ein Subjekt, das die Aktionäre beschreibt, und ein Objekt, den Zusammenschluss. Sie werden anschließend in Unitex-Graphen eingearbeitet, die auf neuen Korpora den Status eines Unternehmenszusammenschlusses bestimmen können. Die Graphen können durch die Einbindung von Lexika und Kontextmodellierung zudem relevante Entitäten wie z. B. den Unternehmensnamen erkennen. Lokale Grammatiken eignen sich in besonderem Maße für das in diesem Beitrag behandelte Thema, da sie die syntaktische Struktur des Satzes berücksichtigen und somit der Akteur und im Zusammenhang dazu seine Handlungen oder Meinungen extrahiert werden können. Kleinere Entitäten können einzeln erkannt und in den Kontext eines größeren Graphs eingebettet werden. Folgender (etwas vereinfachter) Graph erkennt z. B. das Abstimmungsergebnis:

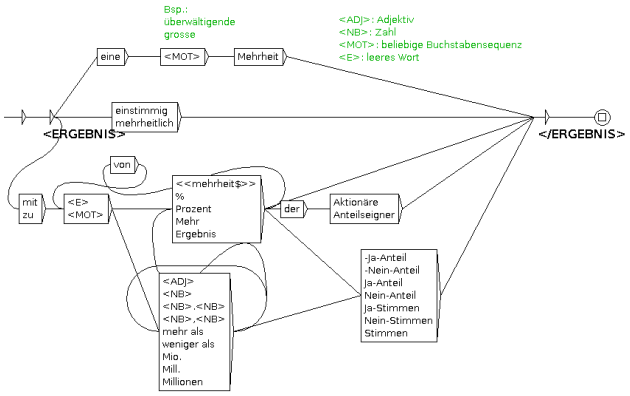


Abb. 1: Graph zur Erkennung des Abstimmungsergebnisses

Die erkannten Textstellen werden mit XML-Tags versehen und im Textformat abgespeichert, sodass sie leicht visualisiert und weiterverarbeitet werden können.

Beispiele für mit diesem Graphen erkannte Textstellen sind:

1 Die Aktionäre von Rhône-Poulenc haben auf ihrer Hauptversammlung in Paris <ERGEBNIS>mit einer Mehrheit von 97,1 Prozent</ERGEBNIS> die Kapitalerhöhung ihres Unternehmens gebilligt, die Voraussetzung für den Aktientausch und die Fusion mit Hoechst zu Aventis ist.

2 In Pittsburgh stimmten die Aktionäre der Soßen-Firma <ERGEBNIS>mit großer Mehrheit</ERGEBNIS> für die Übernahme.

Abb. 2: Auszug aus der Konkordanz der erkannten Textstellen zum Abstimmungsergebnis

Folgende Beispiele zeigen einen Auszug aus der Konkordanz des Graphen zur Erkennung des Abstimmungsergebnisses von Aktionären:

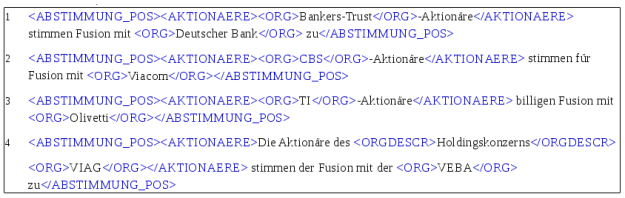


Abb. 3: Auszug aus der Konkordanz der erkannten Textstellen zur Aktionärsabstimmung

Durch Ersetzen der Prädikate können Synonyme gefunden werden, wie z. B. „erlauben“, oder aber auch Prädikate wie „vereiteln“ und „verhindern“, die die Ablehnung des Zusammenschlusses ausdrücken. Hinsichtlich der Argumentstruktur unterscheiden sich die Prädikate nicht, sie müssen jedoch aufgrund ihrer Bedeutung verschieden annotiert werden. Ebenso können durch Ersetzen des Subjekts die Entscheidungen anderer Akteure, beispielsweise des Kartellamts, erkannt werden.

Oft ist auch nur von dem Plan einer Aktionärsabstimmung die Rede. Dieser wird häufig in Form von Modalverbkonstruktionen zum Ausdruck gebracht:

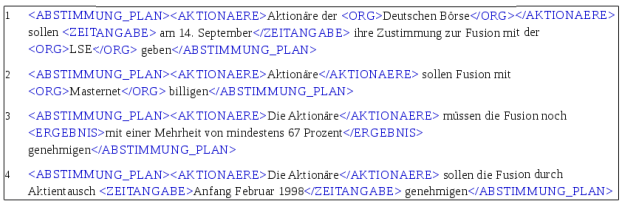


Abb. 4: Auszug aus der Konkordanz der erkannten Textstellen zu einer geplanten Aktionärsabstimmung

Die Extraktion der analysierten Prädikat-Argument-Strukturen mithilfe von lokalen Grammatiken ermöglicht die Kategorisierung eines Ereignisses, im vorliegenden Fall der Aktionärsabstimmung, das für die Einschätzung des Status eines Zusammenschlusses von zentraler Bedeutung ist.

Evaluation

Tabelle 2 zeigt die Ergebnisse einer Evaluation bezüglich der erfolgten Zustimmung der Aktionäre zu einem Zusammenschluss. Sie wurde auf einem mithilfe der COSMAS-Datenbank erstellten Testkorpus mit 623 Sätzen (Keywords: („Fusion“ ODER „Übernahme“) sowie „Aktionäre“) durchgeführt. Hierbei wurden sowohl die Erkennung der Relation als auch der daran beteiligten Entitäten wie z. B. Unternehmensnamen, Zeit- und Ortsangaben berücksichtigt.

	Zustimmung der Aktionäre zu einem Zusammenschluss	Entitäten
Precision	100/101=99,0%	117/120=97,5%
Recall	66/100=66,0%	120/197=60,9%
F-Score	79,2%	75,0%

Tab. 2: Ergebnisse der Evaluation bezüglich der erfolgten Zustimmung der Aktionäre zu einem Zusammenschluss

Fazit und Ausblick

Im vorliegenden Beitrag wird gezeigt, wie wiederkehrende sprachliche Muster dazu genutzt werden können, die Rolle von Aktionären bei einem Unternehmenszusammenschluss am Beispiel von Aktionärsabstimmungen zu modellieren. Mithilfe von lokalen Grammatiken kann der Ausgang der Abstimmungen automatisch extrahiert werden. Wichtige Entitäten wie Zeit- und Ortsangaben, Akteure und Organisationsnamen werden ebenfalls erkannt. Die so strukturierten Informationen können anschließend in eine semantische Suchmaschine eingebettet werden. Mit dieser Methode und durch Erweiterung der vorhandenen Muster kann in den nächsten Schritten ein System zur Erkennung der relevanten Phasen eines Zusammenschlusses sowie das Zusammenwirken der unterschiedlichen Akteure erstellt werden. Nach der Fertigstellung des Systems sollen bezüglich neuer Zeitungsnachrichten Aussagen zum derzeitigen Stand eines Zusammenschlusses getroffen werden können. In einem nächsten Schritt wäre auch eine Analyse hinsichtlich sprachlicher Indikatoren interessant, die nicht unmittelbar an ökonomische Schritte geknüpft ist: Inwiefern kündigen Passagen wie „droht zu scheitern“ tatsächlich das Scheitern des Prozesses an? Darüber hinaus könnte auch die sprachliche Modellierung von Gerüchten („Der Konsumgüterhersteller Henkel ist Kreisen zufolge Favorit im Rennen um den Haarpflegespezialisten Wella“, vgl. Focus 25.05.2015) mit lokalen Grammatiken implementiert werden, um die Grenze zwischen klaren Fakten und unsicheren Aussagen in Zeitungsnachrichten zu markieren. Die Methode ist in thematischer Hinsicht nicht auf ein bestimmtes Textkorpus beschränkt, die Graphen müssen jedoch bei Wechsel der Textdomäne angepasst werden.

Bibliographie

Focus (25.05.2015): "Henkel Favorit für Wella-Übernahme - Wert: 5,5 bis 7,0 Milliarden Dollar" http://www.focus.de/finanzen/news/wirtschaftsticker/kreise-henkel-favorit-fuer-wella-uebernahme-wert-5-5-bis-7-0-milliarden-dollar_id_4705463.html [letzter Zugriff 05. Oktober 2015].

Gross, Maurice (1997): "The Construction of Local Grammars", in: Roche, Emmanuel / Schabès, Yves (eds.): *Finite-State Language Processing*. Cambridge, Massachusetts, USA: MIT Press 329–354.

Guenthner, Franz / Maier, Petra (eds.) (1994): *Das CISLEX Wörterbuchsystem*. München.

Hoffjann, Olaf (2014): "Presse- und Medienarbeit in der Unternehmenskommunikation", in: Zerfaß, Ansgar / Piwinger, Manfred (eds.): *Handbuch Unternehmenskommunikation*. Wiesbaden: Springer 671-690.

Hogenboom, Frederik / Frasincar, Flavius / Kaymak, Uzay / de Jong, Franciska (2011): "An overview of event extraction from text", in: *Proceedings of Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011)*. Workshop in conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011), Bonn, Germany, October 23, 2011: 48–57.

Institut für Deutsche Sprache (IDS) (o.J.): *Cosmas II*. Corpus Search, Management and Analysis System http://www.ids-mannheim.de/cosmas2 [letzter Zugriff 05. Oktober 2015].

Mallchok, Friederike (2004): *Automatic Recognition of Organization Names in English Business News*. München: Ludwig-Maximilians-Universität München.

Nagel, Sebastian (2008): *Lokale Grammatiken zur Beschreibung von lokativen Sätzen und ihre Anwendung im Information Retrieval*. München: Ludwig-Maximilians-Universität München.

Nassirtoussi, Arman Khadjeh / Aghabozorgi, Saeed / Wah, Teh Ying / Chek Ling Ngo, David (2014): "Text mining for market prediction: A systematic review", in: *Expert Systems with Applications* 41,16: 7653–7670.

Palmieri, Rudi (2014): *Corporate argumentation in takeover bids*. Amsterdam / Philadelphia: John Benjamins.

Saurì, Roser / Pustejovsky, James (2012). "Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text", in: *Computational Linguistics* 35, 1: 1–39.

Université Paris-Est Marne-la-Vallée (o.J.): *Unitex* http://igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf [letzter Zugriff 05. Oktober 2015].

Digitales Publizieren. Bedingungen - Optionen - Empfehlungen

Stäcker, Thomas
staecker@hab.de
Herzog August Bibliothek Wolfenbüttel

Baum, Constanze
baum@hab.de
MWW-Forschungsverbund / Herzog August Bibliothek Wolfenbüttel

Steyer, Timo
steyer@hab.de
MWW-Forschungsverbund / Herzog August Bibliothek Wolfenbüttel

Kleineberg, Michael
michael.kleineberg@ub.hu-berlin.de
Humboldt Universität zu Berlin

Baillot, Anne
anne.baillot@gmail.com
Humboldt Universität zu Berlin

Kaden, Ben
ben.kaden@ub.hu-berlin.de
Humboldt Universität zu Berlin

Chen, Esther
echen@mpiwg-berlin.mpg.de
Max-Planck-Instiut für Wissenschaftsgeschichte

Walkowski, Nils-Oliver
walkowski@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Schwaderer, Christian
christian.schwaderer@uni-tuebingen.de
Universität Tübingen

Ernst, Thomas
thomas.ernst@uni-due.de
Universität Duisburg-Essen

Die AG Publikationen der DHd hat sich eingehend mit der Frage der digitalen Publikation auseinandergesetzt und will in Form eines Posters ihre Arbeitsergebnisse, die zeitgleich als *working paper* auf http://dhd-wp.hab.de/ erscheinen werden, vorstellen. Es sollen Empfehlungen entwickelt werden, die sowohl zu einem besseren Verständnis digitaler Publikationen beitragen, als auch Entscheidungsträgern Hinweise zur Entwicklung von Kriterien für eine gute digitale Praxis an die Hand geben. Die im Zusammenhang mit dem Begriff der digitalen Publikation behandelten Themen gliedern sich in fünf Arbeitsschwerpunkte, die jeweils mit einem Kurzessay gewürdigt werden.

Das erste Essay widmet sich der Frage der Definition *digitaler wissenschaftlicher Publikationen* und versucht zu klären, welche besonderen Merkmale für digitale wissenschaftliche Publikationen ausschlaggebend sind. Es behandelt nicht nur das Verhältnis von klassischen zu neuen Publikationsformaten und sucht zu bestimmen, welche Änderungen die digitale Präsentation mit sich bringt, sondern fragt auch nach den Konsequenzen, die sich aus der Prozessierbarkeit von elektronischen Texten und der mittlerweile verbreiteten Nutzung von mit deskriptiven bzw. semantischem Markup versehenen Texten ergeben und die bei einem unverändert hohen wissenschaftlichen Qualitätsanspruch zu einem völlig veränderten Text- und Dokumentverständnis führen, was auch mit Blick auf die Langzeitarchivierung Konsequenzen hat.

Ein weiterer Komplex nimmt sich der Frage an, was *digitale Autorschaft* kennzeichnet. Hier stehen einerseits neue und differenzierte Modelle der Urheber- und Beiträgerschaft im Fokus, andererseits aber auch Phänomene der kollaborativen oder auch anonymen Autorschaft in einem 'Schwarm'. Direkt damit verbunden sind veränderte Möglichkeiten der Zuschreibungen von Reputation. Förderinstitutionen wird empfohlen, Verfahren zu entwickeln, wie sie differenzierte Autorschafts- und Beiträgerrollen als Teil ihrer Vergabepraxen nutzen können.

Der dritte Teil widmet sich dem Thema *Begutachtung wissenschaftlicher Veröffentlichungen*. Diskutiert werden neue Modelle des *peer reviewing*, z. B. *blind review*, *double blind review* oder *open peer review*, aus denen Empfehlungen entwickelt werden, die zur Konsolidierung des WWW als verlässlichem Publikationsort und -archiv wissenschaftlicher Arbeiten insgesamt beitragen können. Die verschiedenen operativen Optionen im Review-Verfahren zeigen, dass es eine vielfältige Palette an Gestaltungsmöglichkeiten gibt, wobei vor allem solche Verfahren favorisiert werden, die einer Liberalisierung von Wissens- und Wissenschaftsdiskursen Vorschub leisten, die Transparenz von Ideen fördern und Exklusionsmechanismen vermeiden helfen.

Die beiden abschließenden Essays befassen sich einerseits mit Fragen der *Versionierung und Zitation*, andererseits mit den *Anforderungen und Bedingungen von Open Access*. Die Veränderbarkeit digitaler Dokumente wirft Fragen der Abschließbarkeit einer elektronischen Publikation ebenso auf wie zu deren Status als verlässlichem Referenzobjekt. Die Lösung solcher Fragen liegt einerseits in einer Versionierung, die verschiedene Zustände des digitalen Dokumentes reproduzierbar macht, zum anderen aber auch in einer Verständigung darüber, was als abgeschlossenes Dokument gelten kann. Eng verschränkt damit ist die Frage der Zitierbarkeit des Textes, wobei dabei nicht nur die Beziehung der verschiedenen Versionen untereinander zu problematisieren ist, sondern auch die sich aus dem digitalen Medium ergebenden neuen Möglichkeiten einer feineren Zitiergranularität, die theoretisch bis auf den einzelnen Buchstaben hinunter reicht. Mit dem Thema *Open Access* wird noch einmal die aus Sicht der DHd-Community zu formulierende *conditio sine qua non* für digitale Publikationen aufgegriffen und verschiedene Modelle der offenen Publikation und des mit ihr verbundenen Rechtsstatus erörtert.

Das Poster soll dazu dienen, diese fünf Schwerpunktthemen der DH-Community nahezubringen. Es soll durch repräsentative Schlagworte dazu anregen, sich mit den Themen der AG zu beschäftigen und gemeinsam mit den AG-Mitgliedern zu diskutieren. Ergebnisse aus den Postergesprächen sollen möglichst als Annotationen und Kommentare direkt in ein assoziiertes Dokument eingetragen werden, dessen Adresse als QR-Code auf dem Poster angeboten wird.

Bibliographie

Herzog-August-Bibliothek Wolfenbüttel (2015): *DHd Working Papers*. http://dhd-wp.hab.de/ .

Digital Humanities und Linguistik: Herausforderungen und ihre Potenziale am Beispiel der Annotation multimodaler Daten

Trevisan, Bianka
b.trevisan@tk.rwth-aachen.de
RWTH Aachen, Deutschland

Reimer, Eva
e.reimer@tk.rwth-aachen.de
RWTH Aachen, Deutschland

Digmayer, Claas
c.digmayer@tk.rwth-aachen.de
RWTH Aachen, Deutschland

Ullrich, Anna
a.ullrich@tk.rwth-aachen.de
RWTH Aachen, Deutschland

Jakobs, Eva-Maria
e.m.jakobs@tk.rwth-aachen.de
RWTH Aachen, Deutschland

Einführung

Das Abstract fokussiert die Frage, welchen Beitrag computergestützte Methoden für die Untersuchung multimodaler Daten in der angewandten Linguistik leisten können, welche Herausforderungen mit der Entwicklung verbunden sind und wie sich „traditionelle“ und computergestützte Verfahren wechselseitig bereichern. Einer der großen Vorteile computergestützten Arbeitens ist, dass der Forscher oder die Foscherin wesentlich größere Datenbestände analysieren kann und teilweise zu Aussagen kommt, die mit händischen Verfahren zeitlich wie personell kaum in Forschungsprojekten zu leisten sind.

Der Beitrag stützt sich auf Daten und Fragestellungen des DFG-geförderten Projektes ModiKo (2014-2017, GZ: JA1172/3-1). Ziel des Projektes ist die Entwicklung von Ansätzen und Methoden, die es erlauben, Formen und Funktionen von Modalitätsinterdependenzen (MID) in ihrer Musterhaftigkeit systematisch zu beschreiben und zu analysieren (Reimer et al. 2015; Ullrich et al. im Druck). Das Forschungsprogramm basiert auf gesprächsanalytischen Ansätzen, die gegenstandsbezogen erweitert werden durch korpus- und texttechnologische sowie computerlinguistische Ansätze. Teil des Projektes ist die Entwicklung eines Annotationstools für heterogene Datenbestände, mit dem Datenformate über mehrere Ebenen annotiert, Annotationen datenformatübergreifend in Bezug gesetzt und in ihrem Bezug dargestellt werden können.

Bisher fehlen für die systematische Beschreibung und Analyse von MID-Formen und Funktionen geeignete Ansätze, Methoden und Tools (Jakobs et al. 2011). Der vorliegende Beitrag gibt einen Einblick in die laufende Projektarbeit. Er diskutiert methodische Herausforderungen anhand der Frage, wie sich mit computergestützten Methoden verbale Thematisierungen von MID (MID-anzeigende Lexeme und Mehrwortlexeme) über große Datenbestände hinweg ermitteln lassen. Verbale Thematisierungen liefern Hinweise auf das Auftreten von Modalitätsinterdependenzen. Zwar können keine exakten Angaben über das Auftreten von MID gemacht, händische Analysen so jedoch vereinfacht werden.

Im Folgenden werden der Stand der Forschung, das im Projekt untersuchte Fallbeispiel und das Korpus beschrieben. Im Anschluss wird beispielhaft gezeigt, wie die Auseinandersetzung mit den Forschungsgegenständen (*hier*: verbale Thematisierung von MID) die Entwicklung von Methoden vorantreibt. Es werden Ergebnisse der Arbeiten in ModiKo aufgezeigt und ein Fazit gezogen.

Stand der Forschung

Modalitätsinterdependenzen (MID) sind definiert als Zusammenspiel komplexer Ausdrucksressourcen wie Sprechen, Schreiben und graphisch-symbolisches Visualisieren (Fiehler 1980), wie sie in professionellen Interaktionssituationen von den Interaktionsteilnehmer_innen zu bestimmten Zwecken genutzt und situationsabhängig kombiniert werden (Ullrich et al. im Druck). Das Eintreten einer Modalitätsänderung wird von den Interaktionsbeteiligten häufig verbal thematisiert, etwa durch MID-anzeigende Einzellexeme (z. B. *schreiben*) oder Mehrwortlexeme (z. B. *ich schreib das mal hier rein*). Die Thematisierung bezeichnen wir als *Modality-taking*.

Fallbeispiel

Die untersuchten MID sind Teil eines Fallbeispiels, das in einem Vorgängerprojekt (IMIP: Interdisziplinäre Methoden industrieller Prozessmodellierung, BMBF, 2008-2011; Jakobs et al. 2011) in der sachgüterproduzierenden Industrie erhoben wurde. Im Fallbeispiel werden Prozesse im Unternehmen von Prozessmodellierern im Gespräch mit Mitarbeitern erhoben und modelliert (Eraßme et al. 2015). Interaktionsbegleitend machen sich die Beteiligten (Prozessmodellierer und Unternehmensmitarbeiter) Notizen oder fertigen Skizzen an. Sie nutzen diese als intermediäre Objekte (Jeantet 1998) für die interaktive Rekonstruktion der Gesprächsinhalte.

Korpus

Die in ModiKo genutzten Daten stützen sich – wie oben erwähnt – auf das Vorprojekt IMIP. Der aus IMIP übernommene (Teil-)Datensatz umfasst 548 Minuten Videoaufzeichnung und 89 gescannte Dokumente sowie 266 Transkriptseiten.

Die Analyse der Daten erforderte eine Reihe methodischer Anpassungen, die im Folgenden beschrieben werden.

Methodenentwicklung

Für das übergeordnete Ziel der musterhaften Beschreibung von MID konzentrieren sich die Analysen in ModiKo auf die Textdokumente des IMIP-Datensatzes (Transkripte). Es stellte sich heraus, dass zahlreiche Anpassungen und Überarbeitungen der aus dem Vorgänger-Projekt stammenden Datensätze erforderlich waren. So zeigte sich zum Beispiel, dass die ursprünglich nach GAT 2 (Selting et al. 2009) erstellten Transkripte zu statisch waren für eine adäquate Erfassung und Notation interaktionsbegleitender Phänomene (z. B. die Erfassung genutzter Objekte, Kontextinformationen). Um das Problem der mehr oder weniger statischen Beschreibung verbaler Interaktionen in Textdokumenten (Transkript) zu lösen, werden die Transkripte in das Tool EXMARaLDA

(Schmidt / Wörner 2014) eingelesen. EXMARaLDA ermöglicht eine Mehrebenen-Annotation. Eine geeignete Annotation von MID in den Interaktionsausschnitten erfordert jedoch eine Erweiterung von EXMARaLDA. Die Erweiterung zielt auf eine größtmögliche Flexibilität in der Annotation von MID-bezogenen Phänomenen als Voraussetzung für die Identifizierung von Mustern (z. B. für MID-anzeigende verbale Thematisierungen).

Eine weitere methodische Neuerung ergibt sich mit der Unterscheidung von drei Typen von Dokumenten: Primär-, Sekundär- und Tertiärdokumente.

Primärdokumente sind Videodateien der erhobenen professionellen Interaktionen sowie Scans der Skizzen, die von den beteiligten Akteuren in der Interaktion angefertigt werden (Berg / Milmeister 2008).

Sekundärdokumente sind multimodale Transkripte (*auch*: Verbaltranskripte) der Primärdokumente (s. auch Schmitt / Dausendschön-Gay 2015). Multimodale Transkripte erfassen die Komplexität verschiedener Ausdrucksressourcen, die die Interaktionsbeteiligten in der Interaktion nutzen (wie praktische Handlung, Verbales, Mimik, Blickrichtung, Gestik oder die Position im Raum).

Die Kategorie *Tertiärdokument* wurde in ModiKo geprägt, um einen dritten Typ von Dokumenten terminologisch fassen zu können – die Mehrebenen-Annotation von Sekundärdokumenten (Reimer et al. 2015). Tertiärdokumente geben dem Forscher die Möglichkeit, Sekundärdokumente durch die Notation verschiedener Phänomene wie sprachbegleitende Gesten (z. B. *auf etw. zeigen*), (materielle) Objekte (z. B. Klemmbrett, Kugelschreiber) sowie kontextuelle und verbale Informationen zu ergänzen. Das für MID-bezogene Phänomene zu entwickelnde Annotationssystem orientiert sich an dem in Trevisan (2014) entwickelten Mehrebenen-Annotationsansatz und adaptiert ihn gegenstandsspezifisch.

Im Laufe des Projektes sollen die Erweiterungen von EXMARaLDA erlauben, alle drei oben genannten Typen von Dokumenten in ein und dem selben Tool zu erfassen und bezogen aufeinander zu analysieren (bisher fehlt die Integration der Videos und der Scans als Teil der Primärdokumente).

Zu den Langzeitzielen der Toolentwicklung gehört, dass das Tool Forscher digital dabei unterstützt, MID-bezogene Muster zu identifizieren und abzubilden. Die Identifizierung und Abbildung dieser Muster soll durch das in Beziehung setzen von Einträgen verschiedener Dokumenttypen ermöglicht werden (vgl. Abbildung 1).

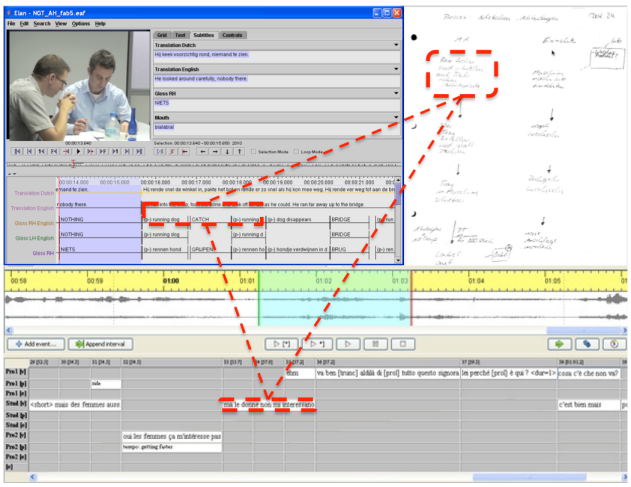


Abb. 1: Beispielhafte Toolabbildung

Zukünftige Arbeiten betreffen die Visualisierung und Umsetzung des Zusammenspiels der unterschiedlichen Datenformate und Modalitäten (Videos, Scans der Skizzen, Transkripte, Mehrebenen-Annotation). Die Toolentwicklung soll es ermöglichen, dem Forscher verschiedene Datenformate in einem visuellen Bezugsfeld (Screen) anzuzeigen und sie dort in Bezug zu setzen (in Abbildung 1 beispielhaft rot markiert).

Ergebnisse

Für die Ermittlung verbaler Thematisierungen von MID wurde im Projekt ModiKo ein Analyseverfahren entwickelt, das händische und computergestützte Methoden kombiniert. Im ersten Schritt (händische Analyse) wurden alle Sekundärdokumente in ihrer Erhebungslogik händisch auf MID-anzeigende Lexeme durchsucht. Die identifizierten MID-anzeigenden Lexeme wurden extrahiert und systematisch als Lexikon aufbereitet. Im zweiten Schritt (computergestützte Analyse) wurde die Auftretenshäufigkeit verbaler Thematisierungen mit dem Tool AntConc [1] ermittelt. Zu diesem Zweck wurden die Transkripte in ein AntConc-kompatibles Format (.txt) überführt, in AntConc eingelesen und quantitativ analysiert.

Die händische Auswertung ergab, dass primär Verben (hier: *malen*, *schreiben*) aber auch Substantive (z. B. *Blatt, Bleistift*) und Adverbien (z. B. *hier*, *da*) Modalitätsänderungen anzeigen. Dies bestätigte sich in der quantitativen Analyse: Eine außerordentliche hohe Anzahl an Fundstellen konnte für die Verben *schreiben* (n = 157) und *machen* (n = 531) sowie für die Adverbien *hier* (n = 840) und *mal* (n = 149) identifiziert werden.

Besonders hervorzuheben ist die Verwendung von verbalen Thematisierungen von MID in Form von Mehrwortlexemen (z. B. *ich setz das mal hier vor*). Bezogen auf das Gesamtmaterial lassen sich für interaktionsspezifische Aufgaben Trigramme (z. B. *ich mach mal*) bestimmen, die das Eintreten einer

Schreibhandlung oder etwa das Skizzieren von Prozessen andeuten.

Fazit

Die Verbindung qualitativ-händischer Verfahren mit quantitativ-computergestützten Verfahren bietet neuartige, sehr vielversprechende Forschungsergebnisse, die mit händischen Verfahren allein so im normalen Forscheralltag nicht zu erreichen wären. Im vorliegenden Beitrag wurde aufgezeigt, wie im Projekt ModiKo computergestützte Verfahren genutzt werden, um verbale Thematisierungen von MID im Korpus identifizieren und analysieren zu können. Durch die Einschränkung der zu untersuchenden Datenmenge mittels computergestützter Verfahren wird die notwendige händische Analyse vereinfacht. Zukünftig soll das Verfahren durch zusätzliche Analysen verfeinert werden, um eine exaktere Bestimmung / Identifikation von MID zu erreichen, etwa durch die Einschränkung des zu analysierenden Textfensters oder die Bestimmung zusätzlicher MID-anzeigender Indikatoren.

Die Integration von computergestützten Verfahren in linguistische Analysen erfordert andererseits ein erhebliches computerlinguistisches Know-how für die Adaption und Weiterentwicklung existierender Tools, das bislang kaum Teil der Ausbildung von Linguisten ist und die Zusammenarbeit mit Spezialisten erfordert. Auf längere Sicht erfordert die digitale methodisch-theoretische Weiterentwicklung von Disziplinen wie der Linguistik auch ein Umdenken in den universitären Ausbildungsprogrammen.

Notes

1. Das Programm wurde von Laurence Anthony entwickelt und funktioniert auf allen gängigen Betriebssystemen. Die aktuelle Programmversion 3.5.0 ist unter http://www.laurenceanthony.net/software/antconc/ frei erhältlich.

Bibliographie

Anthony, Laurence (2015): *AntConc Homepage* http://www.laurenceanthony.net/software/antconc/ [letzter Zugriff 12. Februar 2016].

Berg, Charles / Milmeister, Marianne (2008): „Im Dialog mit den Daten das eigene Erzählen der Geschichte finden. Über die Kodierverfahren der Grounded-Theory-Methodologie", in: *Forum Qualitative Social Research FQS* 9, 2: Nr. 13.

Eraßme, Denise / Trevisan, Bianka / Reimer, Eva / Jakobs, Eva-Maria (2015): „Kooperative Konzeptgenesen in professionellen Interaktionen (Poster)", in: *Tagung der Gesellschaft für Angewandte Linguistik 2015.*

Fiehler, Reinhard (1980): *Kommunikation und Kooperation*. Theoretische und empirische Untersuchungen zur kommunikativen Organisation kooperativer Prozesse. Berlin: Einhorn.

Jakobs, Eva-Maria / Fiehler, Reinhard / Eraßme, Denise / Kursten, Anne (2011): „Industrielle Prozessmodellierung als kommunikativer Prozess. Eine Typologie zentraler Probleme", in: *Gesprächsforschung* 12: 223-264.

Jeantet, Alain (1998): „Les objects intermédiaires dans la conception. Eléments pour une sociologie des processus de conception", in: *Sociologie du Travail* 3: 291-316.

Reimer, Eva / Trevisan, Bianka / Eraßme, Denise / Schmidt, Thomas / Jakobs, Eva# Maria (2015): „Annotating Modality Interdependencies", in: *Proceedings of the GSCL 2015*: 110-111.

Schmidt, Thomas / Wörner, Kai (2014): „EXMARaLDA", in: Durand, Jacques / Gut, Ulrike / Kristoffersen, Gjert (eds.): *Handbook on Corpus Phonology.* Oxford University Press: 402-419.

Schmitt, Reinhold / Dausendschön-Gay, Ulrich (2015): „Freiraum schaffen im Klassenzimmer: Fallbasierte methodologische Überlegungen zur Raumanalyse", in: *SpuR - Arbeitspapiere des UFSP Sprache und Raum* 4. Universität Zürich http://www.spur.uzh.ch/research/publications/SpuR_Arbeitspapier_Nr04_150711.pdf [letzter Zugriff 12. Februar 2016].

Selting, Margret / Auer, Peter / Barth-Weingarten, Dagmar et al. (2009): „Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)", in: *Gepsrächsforschung* 10: 353-402. http://www.gespraechsforschung-ozs.de/fileadmin/dateien/heft2009/px-gat2.pdf [letzter Zugriff 12. Februar 2016].

Trevisan, Bianka (2014): *Bewerten in Blogkommentaren*. Mehrebenenannotation sprachlichen Bewertens. PhD, RWTH Aachen University.

Ullrich, Anna Valentine / Jakobs, Eva-Maria / Eraßme, Denise (im Druck): „„ich schreib das mal hier rein ähm". Modality-taking – Schreibhinweise in professionellen mündlichen Interaktionssituationen", in: *Glottotheory.*

Die Schule von Salamanca. Ansätze für vernetzte und visualisierbare Daten

Wagner, Andreas
andreas.wagner@em.uni-frankfurt.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Caesar, Ingo
caesar@rg.mpg.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Das 2013 gestartete Projekt *Die Schule von Salamanca. Eine digitale Quellensammlung und ein Wörterbuch ihrer juristisch-politischen Sprache* umfasst eine Sammlung von über 100 Texten iberischer Theologen und Juristen des 16. und 17. Jahrhunderts zu politischen und rechtlichen Themen. Diese Sammlung wird ergänzt durch ein Handbuch, in dem neben biografischen Informationen zu den in der Edition vertretenen Autoren die Entwicklung zentraler Begriffe der europäischen Rechts- und politischen Ideengeschichte in diesem Diskussionszusammenhang erschlossen wird. Beide Teile werden vollumfänglich und frei online zugänglich sein; zu Beginn des Jahres 2016 werden die ersten Daten der Quellensammlung und des Wörterbuchs zur Verfügung gestellt. Dies sind insbesondere die TEI-XML Dateien der Volltexte der ersten Werke sowie die dazugehörigen Digitalisate. Parallel besorgen wir die Bereitstellung der Daten als Linked Open Data in Verbindung mit einem SPARQL-Endpoint.

Zu Beginn des DHd-Vortrags wird das Projekt und seine Website kurz vorgestellt. Eine wichtige Facette besonders der künftigen Projektentwicklung sehen wir aber in der Einbindung unserer Daten in eine Linked Data Infrastruktur. Der Vortrag soll daher vor allem dafür genutzt werden, um einen Einblick in die Arbeitsschritte, Entscheidungen und Implementierungen zu geben, die in dieser Hinsicht erfolgt sind und noch ausstehen.

1. Wie werden die TEI-Daten in einer Linked-Data-Umwelt angeboten: welche Elemente oder Attribute werden welchen Objekten und Prädikaten welcher Ontologien zugeordnet? Wie wird diese Zuordnung in Anschlag gebracht, um die Daten als semantische Daten anzubieten? Wie ist die zeitliche Dimension vieler biographischer Angaben abzubilden? Wie ist mit alternativen Werten, z. B. konkurrierenden Angaben zu Geburtsdaten, umzugehen? Gibt es Rückwirkungen auf unser „Kerngeschäft“, i. e. das TEI Schema und die Datenerfassung?

Wir werden semantische Daten zu den Werken und zu den in der Edition vertretenen Autoren anbieten, dabei hauptsächlich auf die foaf-, bio-, relationship-, und SPAR-Ontologien zurückgreifen und die TEI Daten über den xtriples Webservice (xTriples 2015) in RDF umwandeln.

2. Wie werden die Daten in unseren eigenen Diensten vernetzt genutzt, wie sollen sie zur externen Nachnutzung und Vernetzung angeboten werden: Welche Dienste und Ressourcen sollen – direkt oder indirekt – über das Projekt angeboten werden? [1] Welche Möglichkeiten ergeben sich durch die Vernetzung unserer Daten mit denen anderer Anbieter, Forschungsfragen in neuer Weise zu bearbeiten? In welcher Form können oder sollen solche erweiterten Möglichkeiten auf der Projektsite öffentlich angeboten werden, wie verhält es sich mit Rechtemanagement und Qualitätssicherung in föderierten Abfragen/Daten? Welche anderen Datenanbieter sind interessante Kooperationspartner, welches sind die einschlägigen Normdatenbanken? [2] Was sind die Erfahrungen mit den verfügbaren Schnittstellen dieser Anbieter?

Wir werden sowohl auflösbare Entitäten URIs als auch einen SPARQL Endpoint anbieten. Föderierte Abfragen zur Bearbeitung spezieller Forschungsfragen werden im Vortrag vorgestellt. Auch für uns bislang noch ungelöste Probleme, die sich z. B. aus der heterogenen Struktur und Qualität der Daten ergeben, werden angesprochen.

Über die infrastrukturelle Seite dieser Vernetzung hinaus gilt es hier die Benutzerperspektive nicht aus den Augen zu verlieren: Wie und wo sollen welche Art von Benutzern Vernetzungen und Abfragen selbständig definieren können? Weit mehr als eine Frage der technischen Möglichkeiten ist dies ein Problem der Widmung und des Einsatzes von Entwicklungsressourcen, denn sie betrifft die Entwicklung von Oberflächen, deren Nutzung weit über die Kernaufgaben der meisten Projekte hinausgeht. [3]

3. Visualisierungsmöglichkeiten, aufsetzend auf verschiedene Services: Ein Problem bei der Gewinnung vieler Daten aus dem Semantic Web ist ihre Aufbereitung in einer Form, die den Anwendenden dazu befähigt, neue Entdeckungen zu machen, vielleicht sogar neue wissenschaftliche Erkenntnisse zu generieren. Je größer dabei die Quantität und die Heterogenität der Daten ist, desto anspruchsvoller wird die Aufgabe, eine gute Visualisierung zu erzeugen.

Es gibt bereits verschiedene Open Source Programme, die sich als Visualisierungstools eignen und bei denen es sich lohnt, über einen Einsatz nachzudenken. Neben generischen Visualisierungs-Tools (z. B. NodeBox 2015, Sgvizler 2015, d3SPARQL 2015), *Google Heatmaps* (Heatmaps 2015) und Zeitverlaufsanzeigen (z. B. über Timeline 2015; HDAT 2015) etwa durch Abfragen von Koordinaten aus dem TGN, lohnt es auch, einen Blick auf *RelFinder* (2015) zu werfen. RelFinder visualisiert die Beziehungen zwischen zwei Ressourcen und gibt auch den jeweiligen Inhalt der Ressource aus. Für alle diese Werkzeuge gilt es jedoch das mögliche Einsatzszenario sorgfältig zu entwerfen. Überlegungen und Kriterien für ein allgemeines Verfahren solcher Entwürfe sollen vorgestellt werden.

Die Zukunftsperspektiven für die Bereiche Vernetzung und Visualisierung werden im Salamanca-Projekt stets mitbedacht, und so umreißt der Vortrag bereits erbrachte Leistungen und Implementierungen, getroffene Entscheidungen aber auch noch offene und zu erprobende Themen: Insbesondere in Punkt 1 gehen viele Aspekte schon im Frühjahr 2016 in „production use“, in Punkt 2 werden einige konzeptuelle Fragen angesprochen, die wir für uns noch nicht gelöst haben, in Punkt 3 werden wir beschreiben, wie wir von einem eher zufälligen Spiel

mit verschiedensten technischen Möglichkeiten zu einem systematischen Prozess der Reflexion über aktuelle und neue Entwicklungsziele übergehen.

Notes

1. Z.B. Auflösung von Ressourcen-URLs, content negotiation, triple-store dump, Linked Data Fragments, SPARQL Endpoint usw.
2. Projekte: z. B. *Scholasticon* (Schmutz 2009) mit biographischen Informationen, die Sammlung *Post-Reformation Digital Library Scholastica* (PRDL 2013) mit Informationen zu Universitäten, Fakultäten und Lehrstühlen der Frühen Neuzeit oder *Early Modern Letters Online* (EMLO 2015) mit Daten zu frühneuzeitlichen Korrespondenzen. Normdatenbanken: z. B. *Getty Thesaurus of Geographic Names* (TGN 2015), *Gemeinsame Normdatenbank* der Deutschen Nationalbibliothek (GND 2015) oder der *CERL Thesaurus* (CERL 2015).
3. Die Anwendung *Linked Data Fragments Client* (LDFC 2015) zeigt etwa Beispiele, wie sich Nutzende Ressource(n) und Abfragen zusammenklicken können. Eine graphische Oberfläche für (begrenzte) Abfragen ermöglicht einen effizienten Umgang mit Linked Open Data, ohne dafür die Abfragesyntax beherrschen zu müssen. Auch über *LodLive* (LodLive 2012) ließe sich ein entsprechendes Abfrage-Interface realisieren.

Bibliography

CERL (2015): *CERL Thesaurus* http://thesaurus.cerl.org [letzter Zugriff 14. Oktober 2015].

d3SPARQL (2015): *d3sparql.js* Utilities for visualizing SPARQL results with the D3 library http://biohackathon.org/d3sparql/ [letzter Zugriff 13. Oktober 2015].

EMLO (2015): *Early Modern Letters Online* http://emlo.bodleian.ox.ac.uk/ [letzter Zugriff 14. Oktober 2015].

GND (2015): *Gemeinsame Normdatei* http://www.dnb.de/DE/Header/Hilfe/kataloghilfe.html#doc207526bodyText73 [letzter Zugriff 14. Oktober 2015].

HDAT (2015): *Historical Dutch-Asiatic Shipping* http://app.thebrownmap.nl/ [letzter Zugriff 13. Oktober 2015].

Heatmaps (2015): *Google Maps Javascript API* https://developers.google.com/maps/documentation/javascript/heatmaplayer [letzter Zugriff 13. Oktober 2015].

LDFC (2015): *Linked Data Fragments client* http://client.linkeddatafragments.org/ [letzter Zugriff 13. Oktober 2015].

LodLive (2012): *LodLive* Browsing the Web of Data http://en.lodlive.it/ [letzter Zugriff 13. Oktober 2015].

NodeBox (2015): *Meet NodeBox 3* https://www.nodebox.net/node/ [letzter Zugriff 13. Oktober 2015].

PRDL (2013): *Scholastica* Early Modern Academies & Universities http://www.prdl.org/schools.php [letzter Zugriff 14. Oktober 2015].

RelFinder (2015): *RelFinder* Interactive Relationship Discovery in RDF Data http://www.visualdataweb.org/relfinder.php [letzter Zugriff 13. Oktober 2015].

Schmutz, Jacob (2009): *Scholasticon* Ressources en ligne pour l'étude de la scolastique moderne (1500-1800): auteurs, sources, institutions http://scholasticon.ish-lyon.cnrs.fr [letzter Zugriff 14. Oktober 2015].

Sgvizler (2015): *Sgvizler* http://dev.data2000.no/sgvizler/ [letzter Zugriff 13. Oktober 2015].

TGN (2015): *The Getty Thesaurus of Geographic Names* (TGN) http://vocab.getty.edu/ [letzter Zugriff 13. Oktober 2015].

Timeline (2015): *Leaflet.timeline* https://github.com/skeate/Leaflet.timeline [letzter Zugriff 13. Oktober 2015].

xTriples (2015): *Xtriples* A generic webservice to extract RDF statements from XML resources http://xtriples.spatialhumanities.de/index.html [letzter Zugriff 13. Oktober 2015].

Briding the GAP: 100 Jahre Dialektlexikographie als Cloud Service. Der SADE Use Case im DARIAH Competence Centre

Wandl-Vogt, Eveline
eveline.wandl-vogt@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austrian Centre for Digital Humanities; AT

Barbera, Roberto
roberto.barbera@ct.infn.it
Institutio Nationale de Fisica Nucleare; IT

La Rocca, Guiseppe
guiseppe.larocca@ct.infn.it
Institutio Nationale de Fisica Nucleare; IT

Calanducci, Antonio
antonio.calanducci@ct.infn.it
Institutio Nationale de Fisica Nucleare; IT

Kalman, Tibor

tkalman@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mgH Göttingen; DE

Intro

In den Naturwissenschaften ist es seit mehreren Jahrzehnten undenkbar, exzellente Forschung nicht auf Basis elektronischer Forschungsinfrastrukturen zu betreiben bzw. sich diese zu Nutze zu machen. Anders ist dies in den Geisteswissenschaften, wo es nach wie vor zu den Ausnahmen gehört, sich elektronische Forschungsinfrastrukturen für die eigene Forschungsarbeit einzusetzen.

Um diese Lücke zu schließen, wurde das Projekt EGI ENGAGE – DARIAH Competence Centre gestartet (Näheres s.u.).

In diesem Paper diskutieren die Autor.innen beispielhaft den Einsatz und die Zusammenarbeit zweier sehr unterschiedlicher Anbieter von europäischen Forschungsinfrastrukturen, nämlich Digital Research Infrastructures for the Arts and Humanities European Research Infrastructure Consortium (DARIAH-ERIC (cf. DARIAH) und European Grid Infrastructure (EGI 2010-*).

Die Ausgangsprojekte: EGI ENGAGE, DARIAH CC, exploreAT! und COST ENEL

Das Projekt EGI ENGAGE – Engaging the Research Community towards an Open Science Commons (EGI 2015-*) – startete am 1. März 2015 unter Mitfinanzierung der Europäischen Kommission (Horizon2020) mit einer Projektlaufzeit von 30 Monaten und in Zusammenarbeit von mehr als 70 Institutionen in über 30 Ländern. Die Projektmission ist, Open Science Commons für die Forschungsgemeinschaften zu erschließen, und Zugang zu bzw. Nutzung von elektronischen Infrastrukturen für die Forschung zu schaffen, und somit innovative Forschung zu fördern.

Zu diesem Zweck wird mit Forschungsinfrastrukturanbietern unterschiedlicher Disziplinen zusammengearbeitet, unter anderem mit DARIAH-ERIC.

Im Workpackage DARIAH Competence Centre (cf. EGI-DARIAH Competence Center) werden im besonderen Forscher_innen aus den Arts und Humanities angesprochen und eingeladen, elektronische Infrastrukturen zu nutzen.

Übergeordnete Ziele des Competence Centre sind

(1) die Stärkung der Zusammenarbeit zwischen DARIAH-EU und EGI,basierend auf Workflow-orientierten Gateway-Applikationen und der extensiven Nutzung von Applikationen aus den Arts und Humanities in der EGI Cloud (EGI FedCloud).

(2) die Vermehrung der zugänglichen e-Science Sercives und Applikationen für die Forscherinnen der Arts und Humanities sowie die Integration existierender NGI Ressourcen in EGI

(3) das Bewusstsein von möglichen Vorteilen (Stichwort „exzellente Forschung“) durch elektronische Infrastrukturen und e-Science Technologien für Forscher.innen der Arts und Humanities zu wecken bzw. zu stärken, indem man Konditionen schafft für nachhaltige Steigerung einer wachsenden Community von den Arts und Humanities und in den Social Sciences

(4) die Arbeit, die mit anderen Initiativen gestartet wurde wie zB. DC-NET (2009-2012), DCH-RP (2012-2014) auszuweiten und Ergebnisse für die Arts und Humanities konkret nutzbar zu machen und zu implementieren.

Für das DARIAH-CC werden Miniprojekte pilotmäßig umgesetzt. Im konkreten Paper stellen die Autor.innen den SADE Use Case (Näheres s.u.) vor, in welchem die „Datenbank der bairischen Mundarten in Österreich (DBÖ)“ basierend auf unterschiedlichen Formaten und Arbeitschwerpunkten – Archivierung, Digitalisierung (cf. Wandl-Vogt 2007-*), technische Professionalisierung und Georeferenzierung (cf. explore.AT (2015-*), semantisch-kulturelle Vernetzung (Näheres s.u.; cf. COST-ENeL 2013-*)- weiterentwickelt wird.
Im Rahmen des laufenden Nationalstiftungsprojekts exploreAT! (1. April 2015; 48-60 Monate) werden methodisch neue Wege der Lexikographie erprobt. Schwerpunkte der aktuellen Arbeit liegen in der semantisch-kulturellen Vernetzung (Linked Open Data), visuellen Analyse, der Entwicklung von Serious Games sowie einer Erarbeitung und Umsetzung von Bürgerbeteiligungsmodellen (Citizen Science) und der Implementierung von Open Science. Methodisch werden damit Schritte in Richtung eines neuen digitalen Wörterbuchs (WBÖ 3.0) gesetzt.

Die Ergebnisse werden mittels bestehender Netzwerke wie vor allem der DARIAH Arbeitsgruppe Lexical Resources und des COST Netzwerks IS 1305 zur elektronischen Lexikographie (ENeL 2013-2017; Davidovic 2015 et al.) im europäischen Kontext eingebettet, reflektiert und diskutiert. Für den SADE Use Case stellen diese Communities erste – aber nicht ausschließliche – Anlaufstellen für Zusammenarbeit dar.

Der SADE Use Case

SADE – kurz für: Storing and Accessing DARIAH content on EGI – sieht die Entwicklung eines kompletten Workflows für Forscher.innen im Bereich lexikalische Ressourcen / Lexikographie dar.

Ausgehend der Beispielsammlung der ÖAW – zunächst der Daten aus dem Projekt (2007-) – und in Kontext mit Kooperationspartnern dieses Projekts,

beispielsweise dem Naturhistorischen Museum Wien, Wikimedia.AT, Open Knowledge Foundation Österreich, Europeana, sowie der Partner aus DARIAH und COST ENeL, wird ein lexikalisches Netzwerk für kulturelle Fragestellungen erschlossen.

Ein erster Schritt für die Nutzung von Science Commons ist die Entwicklung von Science Gateways. Diese tragen nachweislich zur vermehrten Nutzung von Infrastrukturen in den jeweiligen Forschungsbereichen / Disziplinen bei (cf. Balasko et al. 2013). Ob ein Science Gateway für DARIAH CC oder für einen Forschungsbereich eingerichtet wird, ist derzeit nicht endgültig entschieden. Das Gateway beruht auf Adaptierungen und fallspezifischen Erweiterungen von gUSE/WS-PGRADE (gUSE 2009-*) sowie gLibrary (cf. gLibrary) Technologien. Es wird gemäß den Erwartungen und Bedürfnissen der Lexikograph.innen angepasst und weiterentwickelt.

Um eine Verbindung zwischen den sehr unterschiedlichen Forscher.innengruppen herzustellen, finden unter anderem Vernetzungstreffen statt, z. B. Februar 2015: COST-ENeL-meeting in Wien; Dezember 2015: DARIAH-COST-meeting zu Common Names in Wien.

Die Ergebnisse des SADE Use Case werden umgekehrt im COST-EneL-Netzwerk zur Diskussion gestellt, um User Interessen auf breiter Basis einzuholen und in die Weiterentwicklugn der Infrastruktur einzubringen.

Für die Nutzung der Cloud von essentieller Bedeutung ist das Vorhandensein einer inzwischen eingerichteten Virtuellen Organisation für die Arts und Humanities – vo.dariah.eu (vgl. EGI-VO (joined EGI 01072011).

Forschungsparadimenwechsel im Blickwinkel der Zusammenarbeit

Zusammenfassend wird das vorgestellte Projekt unter dem breiten Blickwinkel der Workplace-Innovation (Kesselring 2014 et al.) diskutiert. Dabei wird der Schwerpunkt auf folgende Punkte gelegt:

(1) Produktinnovation

Durch die Zusammenarbeit entstehen neuen „Produkte“ wie beispielsweise ein Science Gateweay für die Geisteswissenschaften.

Für die Arts und Humanities wird beabsichtigt, Produkte zu entwickeln, die Folge-innovationen bedingen und den Forschungsprozess stimulieren.

(2) Prozessinnovation

Die Zusammenarbeit im interdisziplinären, internationalen, interkulturellen, multilingualen Kontext bedingt Veränderungen laufender Forschungsprozesse bzw. zielt sogar teilweise von Beginn an auf eine derartige Veränderung und Innovation ab.

Durch die Implementierung in bestehende Forschungsnetzwerke und die Anbindung an aktuelle Forschungsfragen auf europäischem Niveau soll sichergestellt werden, dass es sich nicht nur um Projektinnovationen – z. B. eine neue Datenspeicherung für die handelt – sondern dass dadurch neue Forschungs- und Entwicklungsprozesse angestoßen werden.

(3) Organisationsinnovation

Die Einbettung in das Forschungsparadigma digitaler Infrastrukturen und Cloud Services bedingt Organisationsinnovationen, wie beispielsweise die Gründung der virtuellen Organisation für DARIAH, der Working Group Cloud Services in DARIAH u. ä.

Nach der Implementierung der Services aus dem SADE Use Case soll eine Impactmessung der Workplaceinnovation entwickelt werden.
An einer Maximierung der positiven Wirkungen der technischen Neuerungen flankierend durch soziale Innovationen am Beispielfall der digitalen Lexikographie wird gearbeitet.

Bibliographie

Balasko, Akos / Farkas, Zoltan / Kacsuk, Peter (2013): „Building science gateways by utilizing the generic WS-PGRADE/gUSE worklfow sytem.“, in: *Computer Science Journal* 14, 2: 307-325.

COST-ENeL (2013-*): *European CO-operation in Science and Technology: European Network of e-Lexicography (ENeL)*. COST Association. http://www.cost.eu/COST_Actions/isch/IS1305 [letzter Zugriff 15. Oktober 2015].

DARIAH: *Digital Research Infrastructures for the Arts and Humanities* (DARIAH). http://www.dariah.eu [letzter Zugriff 15. Oktober 2015].

Davidovic, David / Wandl-Vogt, Eveline / Skala, Karolj / Kalman, Tibor (2015): „EGI Engage – Competence Centre for DARIAH.“, in: *12th European Semantic Web Conference 2015 Project Networking Session*. http://2015.eswc-conferences.org/sites/default/files/PN-ESWC-2015_num5.pdf [letzter Zugriff 15. Oktober 2015].

DC-NET (2009-2012): *Digital Cultural heritage NETwork*. ERA-NET (European Research Area Network). http://www.dc-net.org/ [letzter Zugriff 15. Oktober 2015].

DCH-RP (2012-2014): *Digital Cultural Heritage Roadmap for Preservation*. http://www.dch-rp.eu/ [letzter Zugriff 15. Oktober 2015].

EGI (2010-*): *European Grid Infrastrucutre* (EGI). http://www.egi.eu [letzter Zugriff 15. Oktober 2015].

EGI: Dariah Competence Centre: https://wiki.egi.eu/wiki/Competence_centre_DARIAH [letzter Zugriff 15. Oktober 2015].

EGI-Engage (2015-*): *Engaging the Research Community towards an Open Science Commons* (EGI-ENGAGE). https://www.egi.eu/about/egi-engage/ [letzter Zugriff 15. Oktober 2015].

EGI-VO (joined EGI 01072011): *EGI: Virtual Organisation for arts and humanities*: vo.dariah.eu. EGI-

Engage Competence Centre https://wiki.egi.eu/wiki/EGI_Virtual_Organisation_for_arts_and_humanities:_vo.dariah.eu [letzter Zugriff 15. Oktober 2015].

ENeL (2013-2017): *European Network of e-Lexicography* (ENeL). COST http://www.elexicography.eu [letzter Zugriff 15. Oktober 2015].

exploreAT! (2015-*): *exploring austria´s culture through the language glass*. http://www.oeaw.ac.at/acdh/de/node/187 [letzter Zugriff 15. Oktober 2015].

gLibrary (o.J.): *Digital Libraries on the Grid* (gLibrary). INFN [letzter Zugriff 15. Oktober 2015].

gUSE (2009-*): *Grid and Cloud User Support Environment* (gUSE). Laboratory of Parallel and Distributed Systems (LPDS) http://guse.hu/about/architecture [letzter Zugriff 15. Oktober 2015].

Kesselring, Alexander / Blasy, Cosima / Scopetta, Anette (2014): *Workplace Innovation. Concepts and indicators*. European Commission Report http://ec.europa.eu/DocsRoom/documents/8250/attachments/1/translations/en/renditions/native [letzter Zugriff 15. Oktober 2015].

Wandl-Vogt, Eveline (ed.) (2007-*): *"Datenbank der bairischen Mundarten in Österreich"* (DBÖ), Wien 1993-2010, electronically mapped. Institut für Österreichische Dialekt- und Namenlexika, Zentrum für Sprachwissenschaften, Bild- und Tondokumentation, Österreichische Akademie der Wissenschaften (publiziert: Wien: 1. Juli 2010). http://wboe.oeaw.ac.at [letzter Zugriff 15. Oktober 2015].

Automatische Typenbestimmung in historischen Drucken

Weichselbaumer, Nikolaus

weichsel@uni-mainz.de
Johannes Gutenberg Universität Mainz, Deutschland

Christlein, Vincent

vincent.christlein@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Eine zentrale Methode der Analytical Bibliography ist die Bestimmung der Drucktype einer Inkunabel. Da im Frühdruck nur sehr beschränkt mit Schriften gehandelt wurde, ermöglicht diese Bestimmung oft die (näherungsweise) Datierung und Firmierung eines Drucks. Als Hilfsmittel dafür steht seit Jahrzehnten das Typenrepertorium zur Verfügung, bei dem der Benutzer mithilfe der Form von ›M‹(gebrochene Schriften) und ›Qu‹(Antiqua) sowie der Schriftgröße eine Vorauswahl aus den mehreren Tausend bekannten Typen trifft und dann Tafeln mit dem vollständigen Zeichensatz einer Schrift mit dem zu bestimmenden Druck abgleicht. Diese Methode ist erprobt, aber langwierig und nur für Spezialisten zu handhaben, auch wenn das umständliche Wälzen des Tafelwerks inzwischen einer Datenbank (Staatsbibliothek zu Berlin 2014) gewichen ist.

Diese Methode soll durch ein Werkzeug ergänzt werden, das auf die inzwischen in großem Umfang zur Verfügung stehenden Volldigitalisate von Inkunabeln mit Mitteln der Mustererkennung analysiert und dem Nutzer vollautomatisch die in einem vorliegenden Digitalisat verwendete Type bzw. eine Liste der wahrscheinlichsten Typen ausgibt.

Die automatische Identifikation historischer Druckschriften ist bisher nicht bearbeitet. Es kann aber auf Ansätze zur Identifikation von Schreiberhänden zurückgegriffen werden, die zwar nicht 1:1 übertragbar sind, aber viele Probleme vorwegnehmen. Konkret bedeutet das eine Merkmalsextraktion an den Konturen der Schrift, die so gewonnenen lokale Merkmale werden anschließend zu einem globalen Merkmalsvektor zusammengefasst und klassifiziert.

Ein großer Vorteil für die Umsetzung dieses Ansatzes ist die bestehende im Gesamtkatalog der Wiegendrucke vorliegende Zuordnung von ca. 15.000 digitalisierten Inkunabelausgaben den jeweiligen Schriften des Typenrepertoriums. Dieser große Bestand an Ground-Truth-Daten ermöglicht es, für bestimmte Typen zu trainieren und so die Gesamterkennungsrate signifikant zu erhöhen. Dabei sinkt durch die große Datenmenge die Abhängigkeit von variierenden Aufnahmebedingungen bei Digitalisaten. Außerdem verbessert sich die Unterscheidung von Type und Hintergrund (Bedruckstoff), der bei kleineren Beständen an Ground-Truth-Daten leicht mitklassifiziert wird.

Das fertige Werkzeug soll einerseits die Arbeit von Inkunabelforschern erheblich beschleunigen und auch nahestehenden Disziplinen die Bestimmung von Typen ermöglichen. Gleichzeitig könnte diese Methode auch die Fortschreibung des Typenrepertoriums ins 16. Jahrhundert ermöglichen, für das es bisher kein derartiges Verzeichnis gibt, weil die händische Bestimmung angesichts der im Lauf des 16. Jahrhunderts exponentiell steigenden Druckproduktion ausgeschlossen ist. Je nach erreichbarer Präzision der vollautomatischen Erkennung soll eine halbautomatische Erkennung ergänzend implementiert werden, bei der der Benutzer wenige typische Zeichen markiert und so die Zuordnungsgenauigkeit wenigstens in den Bereich der analogen Bestimmungstechnik bringt, dabei aber immer noch wesentlich schneller und einfacher zu bedienen ist.

Der Vortrag präsentiert die Ergebnisse eines Proof of Concept, der mit Einzelseiten aus 100 Inkunabeldigitalisaten, von denen jeweils zwei aus derselben Schrift gesetzt waren, die Gangbarkeit dieses Ansatzes untersucht hat. Es wurde dazu eine Methode zur Schreiberidentifizierung adaptiert, die sehr erfolgreich zeitgenössische Schriften dem richtigen Schreiber

zuordnen kann (Christlein / Bernecker / Angelopoulou 2015). Dabei handelt es sich um einen ganzheitlichen Ansatz, bei dem auf Basis einer Bilddatei Merkmale an der Schrift berechnet werden. Für diese Studie wurde der Textbereich im Bild manuell markiert und anschließend binarisiert (Sauvola / Pietkäinen 2000). Zusammenhängende Komponenten wurden anhand ihrer Flächengröße und Breiten-Höhenverhältnisses gefiltert, so dass möglichst nur Schrift extrahiert wird. An den Konturen der Schrift werden anschließend Zernike-Momente berechnet, welche anschließend mittels VLAD zu einem globalen Merkmalsvektor aggregiert werden. Diese Merkmalsvektoren dienen anschließend zur Typenbestimmung.

Erste Ergebnisse zeigen, dass diese Methode einen möglichen Weg zur Typenbestimmung darstellt: Testete man eines der Dokumente und verglich es mit den restlichen 99 Dokumenten so wurde in 45% der Fälle das Dokument mit derselben Type als wahrscheinlichstes Dokument zurückgeliefert. Betrachtet man die Liste der zehn wahrscheinlichsten Dokumente, so befand sich die richtige Type mit 77% Wahrscheinlichkeit unter ihnen. Anstatt also händisch jede mögliche Type zu überprüfen, kann eine nach Wahrscheinlichkeit sortierte Liste zur Bestimmung benutzt werden und somit den Aufwand der Typenbestimmung drastisch verringern. Dabei ist zu berücksichtigen, dass bei der Durchführung dieses Experiments keinerlei Rücksicht auf unterschiedliche Auflösungen oder anderen Störelemente (Artefakte, schlechte Bildqualität, etc.) genommen wurde und die Datenbasis noch sehr gering war. Bei Berücksichtigung von Störfaktoren und insbesondere bei höheren Datenmengen ist eine signifikante Steigerung der Erkennungsrate zu erwarten.

Damit lässt sich bereits an diesem Proof of Concept zeigen, dass die Methode im Prinzip funktioniert und eine wertvolle Ergänzung zur konventionellen Vorgehensweise darstellen kann, auch wenn noch zu klären bleibt, ob sie die analoge Bestimmung mittelfristig ersetzen kann.

Bibliographie

Christlein, Vincent / Bernecker, David / Angelopoulou, Elli (2015): "Writer Identification using VLAD encoded Contour-Zernike Moments", in: *Document Analysis and Recognition (ICDAR) 2015*. 13th International Conference 23–26 August 2015, Nancy 906-910.

Eisermann, Falk / Duntze, Oliver (2014): "Auf der Spur der seltsamen Typen. Das digitale Typenrepertorium der Wiegendrucke", in: *Bibliotheksmagazin* 3: 41–48 https://www.bsb-muenchen.de/fileadmin/imageswww/pdf-dateien/bibliotheksmagazin/BM2014-3.pdf [letzter Zugriff 29. Dezember 2015].

Haebler, Konrad (1905): *Typenrepertorium der Wiegendrucke*. Abt. I. Deutschland und seine Nachbarländer. Halle: Haupt.

Sauvola, Jaakko / Pietikäinen, Matti (2000): "Adaptive document image binarization", in: *Pattern Recognition* 33, 2: 225-236.

Staatsbibliothek zu Berlin (2014): *Typenrepertorium der Wiegendrucke digital* http://tw.staatsbibliothek-berlin.de [letzter Zugriff 29. Dezember 2015].

Das digitale Handbuch der Höfe und Residenzen im spätmittelalterlichen Reich. Eine suchoptimierte Präsentation von strukturierten und verlinkten XML-TEI Daten.

Wettlaufer, Jörg
jwettla@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

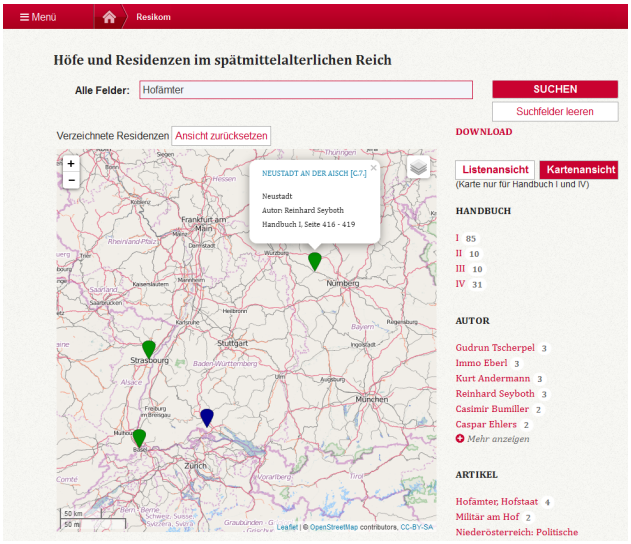
Tech, Maike
tech@sub.uni-goettingen.de
Staats- und Universitätsbibliothek Göttingen, Deutschland

Naegle, Sibylle
naegle@sub.uni-goettingen.de
Staats- und Universitätsbibliothek Göttingen, Deutschland

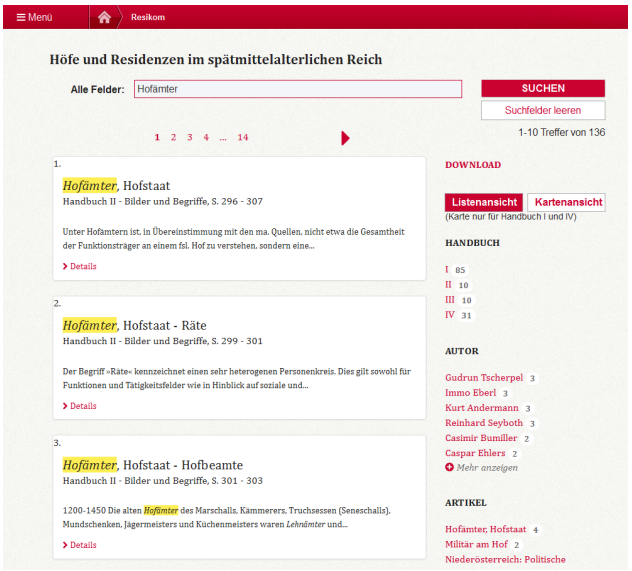
Das digitale Handbuch der Höfe und Residenzen im spätmittelalterlichen Reich ist die online-Ausgabe einer Druckpublikation, die zwischen 1998 und 2011 von Autorinnen und Autoren der Hof- und Residenzenforschung erstellt und von der Residenzen-Kommission der Akademie der Wissenschaften zu Göttingen konzeptionell und redaktionell betreut wurde. Das ca. 5000 Seiten und mehrere hundert Abbildungen umfassende Werk besteht aus über 1000 Einzelartikeln, die dynastisch, topographisch und sachlich gegliedert und vielfältig miteinander verknüpft sind. Der topographische Schwerpunkt liegt auf Höfen und Residenzen des Heiligen Römischen Reichs Deutscher Nation in der Zeit zwischen etwa 1200 bis 1650 (Hirschbiegel / Wettlaufer 1999, 2000, 2002; Wettlaufer 2005). In einer Zusammenarbeit mit der SUB Göttingen wurde ein Konzept für eine Online Präsentation der Druckdaten, die vom Thorbecke Verlag im TEI-Format zur Verfügung gestellt wurden, entworfen und umgesetzt.

In dem Projekt spielen Datenmodellierung, Vernetzung und Visualisierung eine entscheidende Rolle.

Aufbauend auf einen SOLR - Index wird ein Zugriff auf Text und Bilder über eine string-basierte Suche sowie über eine topographischen Zugriff mittels Open-Street-Map Karten und dem leaflet Javascript Framework angeboten, die die ortsgebundenen Informationen zu den Artikeln erschließen.



Eine Facettierung zu Handbuchteilen, Autoren und Artikeln erlaubt die Einschränkung der Suchergebnisse auf bestimmte Teilmengen. Die Texte und Bilder werden sowohl über eine HTML-basierte Ansicht als auch im PDF Format für die Benutzer präsentiert. Abkürzungen werden zur besseren Verständlichkeit bei mouse-over Events aufgelöst. Quellen und Literatur zu den einzelnen Artikeln können bei Bedarf ausgeklappt und angezeigt werden. Ein Zitierlink erlaubt die Referenzierung zur seitenidentischen Druckausgabe des Handbuchs. Eine besondere Herausforderung stellte die Verlinkung der Artikel untereinander dar, die über eindeutige Identifikatoren für alle Texte und Bilder miteinander verknüpft sind. Aufgrund der Datengrundlage waren die End- und Zielpunkte der zu verknüpfenden Strings, die nur durch einen Pfeil im Fließtext gekennzeichnet sind, nicht leicht zu bestimmen. Trotzdem konnte eine über 90% korrekte Verlinkung durch automatisierte string-matching Verfahren mit Hilfe des kontrollierten Lemmavokabulars erreicht werden.



Die Architektur des Projekts ist auf eine leichte Archivierbarkeit und langfristige Nachnutzung hin optimiert, da nach dem Auslaufen von Akademieprojekten in der Regel keine Mittel mehr für eine Pflege von digitalen Projektergebnissen mehr zur Verfügung stehen (Wettlaufer 2012). Über XML TEI stehen die Textdaten langfristig lesbar bereit und durch Apache SOLR können die Daten mit Hilfe einer etablierten Technologie effizient gesucht und angezeigt werden. Die Nutzung von Typo3 und der Extension "Find" erlauben die Integration in die schon bestehende digitale Infrastruktur der Göttinger Akademie. Die "Find"-Erweiterung der SUB Göttingen schafft eine Schnittstelle zu beliebigen SOLR-Indizes in Typo3 und zeichnet sich durch eine leichte Konfigurierbarkeit und erweiterte Templating-Fähigkeiten aus, mit denen komplexe Ansichten der Suchergebnisse realisiert werden können. Das Hosting an der Göttinger Staats- und Universitätsbibliothek lassen die langfristige Pflege und Verfügbarkeit der Handbücher im Rahmen einer bestehenden Kooperationsvereinbarung erwarten.

Für die Zukunft sind neben einer Erweiterung des Userinterface mit zusätzlichen Materialien und einer Verknüpfung der Artikel mit prosopographischen Datensätzen von Hofinhabern auch eine Bereitstellung der Metadaten der Einzelartikel als Linked Open Data sowie eine Verknüpfung mit weiteren Handbüchern zu einem ähnlichen Themenbereich aus einem Nachfolgeprojekt geplant. Das digitale Handbuch wird voraussichtlich im Frühjahr 2016 öffentlich zur Verfügung stehen.

Projektseite: http://adw-goe.de/forschung/abgeschlossene-forschungsprojekte-aus-dem-akademienprogramm/hof-und-residenz/

Bibliographie

Hirschbiegel, Jan / Wettlaufer, Jörg (1999): "Materialien zum Werk 'Fürstliche Höfe und Residenzen

im spätmittelalterlichen Reich. Ein dynastisch-topographisches Handbuch'. Zusammengestellt von Jan Hirschbiegel und Jörg Wettlaufer", in: *Mitteilungen der Residenzen-Kommission der Akademie der Wissenschaften zu Göttingen*. Sonderheft 3 http://resikom.adw-goettingen.gwdg.de/MRK/SH3.pdf [letzter Zugriff 15. Oktober 2015].

Hirschbiegel, Jan / Wettlaufer, Jörg (2000): "Projektdatenbank: Fürstliche Höfe und Residenzen im spätmittelalterlichen Reich. Ein dynastisch-topographisches Handbuch", in: *Mitteilungen der Residenzen-Kommission der Akademie der Wissenschaften zu Göttingen* 11, 2: 9-14 http://resikom.adw-goettingen.gwdg.de/MRK/MRK11-2.pdf [letzter Zugriff 15. Oktober 2015].

Hirschbiegel, Jan / Wettlaufer, Jörg (2002): "Fürstliche Höfe und Residenzen im spätmittelalterlichen Reich. Bilder und Begriffe", in: *Mitteilungen der Residenzen-Kommission der Akademie der Wissenschaften zu Göttingen* 12, 1: 12-18 http://resikom.adw-goettingen.gwdg.de/MRK/MRK12-1.pdf [letzter Zugriff 15. Oktober 2015].

SUB Göttingen (o. J.): *TYPO3 extension providing a frontend for Solr indexes* https://github.com/subugoe/typo3-find [letzter Zugriff 12. Februar 2016].

Wettlaufer, Jörg (2005): "Höfe und Residenzen im spätmittelalterlichen Reich. Erste Ergebnisse des Handbuchprojekts der Residenzen-Kommission der Akademie der Wissenschaften zu Göttingen", in: Pils, Susanne / Niederkorn, Jan Paul (eds.):*Ein zweigeteilter Ort? Hof und Stadt in der frühen Neuzeit* (= Forschungen und Beiträge zur Wiener Stadtgeschichte 44). Wien: Böhlau 7-26.

Wettlaufer, Jörg (2012): "Das digitale Handbuch der Höfe und Residenzen im spätmittelalterlichen Reich. Probleme und Erfahrungen einer digitalen Bereitstellung von kollaborativen Werken in Open Access nach dem Projektende", Vortrag auf dem Workshop *Rechtliche Rahmenbedingungen der Akademievorhaben der Akademie der Wissenschaften zu Göttingen und der Union der deutschen Akademien der Wissenschaften AG „Elektronisches Publizieren“*. 8. und 9. Oktober 2012, Göttingen, Historische Sternwarte http://www.digihum.de/agep/docs/wettlaufer_2012_agep.pdf [letzter Zugriff 15. Oktober 2015].

histoGraph: Graphbasierte Exploration und Crowdbasierte Indexierung

Wieneke, Lars
lars.wieneke@cvce.eu
CVCE Luxembourg, Luxemburg

Düring, Marten
marten.duering@cvce.eu
CVCE Luxembourg, Luxemburg

Guido, Daniele
daniele.guido@cvce.eu
CVCE Luxembourg, Luxemburg

histoGraph

Der Vortrag wird das im CVCE DH Lab entwickelte Werkzeug histoGraph vorstellen, das die graphbasierte Exploration von digitalisierten Quellen mit crowdbasierter Indexierung verknüpft. histoGraph basiert auf einer zu Demonstrationszwecken entwickelten Software, die Teil des FP7-geförderten Projekts CUbRIK zur Mensch-Maschine-Interaktion in der Multimediasuche war. Der Vortrag enthält neben einer Präsentation des neu entwickelten Designs und des weiterentwickelten Konzepts auch eine Live-Demo. histoGraph wird ab dem Frühjahr 2016 als open source Software frei verfügbar sein.

Mit histoGraph eröffnen wir neue Perspektiven auf die umfangreichen Bestände des Centre Virtuel de la Connaissance sur l’Europe . Gegenwärtig sind dort ca. 20.000 Texte, Bilder und Fotos online verfügbar, hierarchisch organisiert in thematischen Sammlungen (*ePublications*). Diese Sammlungen erzählen die Geschichte der europäischen Integration seit 1945 anhand von sorgfältig ausgewählten Primärquellen.

Exploration

histoGraph ergänzt diese expertenbasierten Sammlungen um einen freieren, explorativen Zugang: Nutzer entscheiden, welche Entität – in unserem Falle: welche Person, Institution oder welches Dokument für sie von Interesse ist.

Das histoGraph-Interface ist in drei vertikale Spalten gegliedert: Die erste Spalte gibt einen ersten Überblick zu seiner Biographie und kookkurrienden anderen Personen. Die zweite Spalte listet alle assoziierten Dokumente auf. Die dritte Spalte repräsentiert diese auf Kookkurrenz basierenden Beziehungen als Graph. histoGraph bietet Nutzern nun mehrere Optionen, diese Ergebnisse zu filtern oder zu sortieren. Von besonderer Bedeutung ist aber die Möglichkeit, gezielt nach Beziehungen zwischen bestimmten Personen zu suchen. Hierzu werden zwei oder mehrere Personen ausgewählt und alle Dokumente aufgelistet, in denen beide erwähnt werden. Darüber hinaus zeigt der Graph alle weiteren Personen, die gemeinsam mit den Gesuchten erwähnt werden. Diese Art der Suche ist inspiriert vom Prinzip des *shortest path*, einer gängigen Methode zur Beschreibung von Netzwerktopologien und -zentralitäten. Hierbei werden

alle Schritte gezählt die nötig sind um von einem Knoten des Netzwerks zu einem anderen zu gelangen.

Diese Abfrage funktioniert übrigens ebenso gut für Dokumente oder Institutionen, nur dass in diesem Falle ähnliche Dokumente oder häufig zusammen erwähnte Institutionen dargestellt werden. Dieser Ansatz kombiniert eine gezielte Suche mit einem freieren Finden, dass unerwartete Querverbindungen außerhalb der ursprünglichen Suche sichtbar machen kann. Der größte Unterschied zwischen den eingangs erwähnten hierarchisch organisierten thematischen Sammlungen und histoGraph ist, dass Nutzer die Freiheit haben, ihren eigenen Interessen zu folgen und selbstständig nach für sie relevanten Dokumente und Sozialbeziehungen zu forschen. In histoGraph werden damit mehrere Aspekte historischen Arbeitens aufgegriffen: (1) das genaue Studium einzelner Objekte, (2) deren Betrachtung innerhalb ihres jeweiligen Kontexts, (3) die Suche nach weiteführenden, bislang unberücksichtigten Dokumenten. Die enge Verbindung zwischen Dokumenten und abstrakter Visualisierung sorgt dafür, dass letztere mit Gewinn „gelesen“ und evaluiert werden können.

histoGraph arbeitet momentan ausschließlich mit Kookkurrenzen. Es ist mit diesem Ansatz nur sehr schwer möglich, weitergehende Aussagen über die Bedeutung einer solchen Beziehung beispielsweise zwischen zwei gemeinsam erwähnten Personen zu machen: Diese können miteinander interagiert haben, in unterschiedlichen Kontexten erwähnt worden sein oder gar mit dem Hinweis, dass sie absolut nichts miteinander zu tun hatten. Diese Beliebigkeit ist allerdings auch eine Stärke: Sie überlässt Nutzern die Entscheidung, was als eine relevante Beziehung zu gelten hat. Hierbei gilt: Je genauer Beziehungen definiert sind, desto geringer ist der Anteil an irrelevanten Beziehungen. Aber auch: Je großzügiger Beziehungen definiert sind, desto höher ist die Chance, forschungsrelevante Querverbindungen zu entdecken. Im Entwicklungsprozess versuchen wir, die Balance zwischen diesen beiden erstrebenswerten und doch entgegengesetzten Polen zu halten.

Indexierung

histoGraph eignet sich allerdings nicht nur für die Erforschung von digitalen Sammlungen sondern auch für deren Indexierung. Wir arbeiten mit einer Kombination aus unterschiedlichen Werkzeugen für die Identifizierung von *named entities* wie Personen, Institutionen, Zeitangaben und Orten. Um diese automatisch generierten Annotationen zu prüfen und gegebenenfalls zu verbessern, arbeiten wir zusätzlich mit Methoden des *crowdsourcing*. Hierbei werden einfache Aufgaben, wie etwa die Erkennung von Gesichtern in Fotos oder die Bestätigung eines Datums von so genannten generischen *crowds* übernommen. Anspruchsvollere Aufgaben, wie etwa der Umgang mit Namensvettern bleibt einer *crowd* von Experten vorbehalten. Das System eignet sich ebenso für das kollaborative Indexieren und Annotieren in Teams, etwa einer Projektgruppe.

Im Vergleich mit den bisherigen Sammlungen ermöglicht histoGraph also eine freie Exploration des Materials und das effektive Finden von potentiell relevanten Dokumenten und Beziehungen. Im Zentrum steht hierbei nicht die von Experten kuratierte Auswahl, die mit einem Museumsbesuch vergleichbar ist sondern ein mehr oder minder zielgerichtetes Stöbern, dass einem Archivbesuch näher kommt.

Bibliographie

Centre Virtuel de la Connaissance sur l'Europe (2004-2016), Luxembourg http://www.cvce.eu/ [letzter Zugriff 09. Januar 2016].

Wieneke, Lars / Düring, Marten / Silaume, Ghislain / Lallemand, Carine / Croce, Vincenzo / Lazzarro, Marilena / Nucci, Francesco u. a. (2014): “histoGraph – A Visualization Tool for Collaborative Analysis of Historical Social Networks from Multimedia Collections”, in *Proceedings of 18th International Conference Information Visualisation (IV), 2014 Conference*, Paris.

Big Babylonian Pictures. Kohärenztechniken zur konsistenten Vernetzung von Visualisierungen zu mentalen Modellen

Windhager, Florian
florian.windhager@donau-uni.ac.at
Department für Wissens- und Kommunikationsmanagement, Donau-Universität Krems, Österreich

Schreder, Günther
guenther.schreder@donau-uni.ac.at
Department für Wissens- und Kommunikationsmanagement, Donau-Universität Krems, Österreich

Smuc, Michael
michael.smuc@donau-uni.ac.at
Department für Wissens- und Kommunikationsmanagement, Donau-Universität Krems, Österreich

Mayr, Eva
eva.mayr@donau-uni.ac.at
Department für Wissens- und Kommunikationsmanagement, Donau-Universität Krems, Österreich

Methoden der Informationsvisualisierung (InfoVis) zielen auf die Unterstützung von Kognition im Angesicht von abstrakten Daten- und Themenbeständen. Visuelle Repräsentationen helfen bei der primären Synthese von multiplen Datendimensionen, sowie bei der vertiefenden Analyse und der Vermittlung relevanter Zusammenhänge. Entsprechend trägt auch in den Geisteswissenschaften seit geraumer Zeit ein wachsendes Spektrum von bildgebenden Verfahren zur Exploration und Kommunikation textlicher und thematischer Korpora bei (vgl. Sula 2013; Jänicke et al. 2015). Neben Methoden der statistischen Datenvisualisierung offerieren Techniken der Kartographie und Chronographie, der Dendrogrammatik, des Topic Modelling oder der Netzwerkvisualisierung Einsichten in die Struktur und Dynamik komplexer multidimensionaler (Meta)Datenbestände. Nachdem die zunehmende Nutzung dieser einzelnen Verfahren bereits vorausgesetzt werden kann, richtet der Beitrag seinen Fokus auf ihre synergetische Kombination und Vernetzung im Rahmen von Interfaces mit „ *multiplen views*“ (Roberts 2007).

Interfaces mit multiplen Ansichten kombinieren unterschiedliche Visualisierungsverfahren in einem parallelen Arrangement um deren komplementäre Perspektiven zu verknüpfen. Da multidimensionale Datensätze selten durch Einzelvisualisierungen erschöpfend exploriert werden können, sind multiple Ansichten als „mixed methods“-Ansätze der Visualisierung zu verstehen, die einander ergänzende Blickwinkel und Projektionen zu einem analytischen System mit erhöhter Leistungskraft kombinieren. Multiple Ansichten multiplizieren aber auch die kognitiven Anforderungen an NutzerInnen, die nicht nur die jeweiligen Eigenlogiken (i.e. Syntax & Semantik) der Einzelsichten zu entschlüsseln haben, sondern die auch ihre Einsichten aus einer bildersprachlichen Vielfalt in ein *bigger picture* zusammenführen müssen. Dies kann zu mehr oder weniger gelungenen mentalen Montagen führen.

Im Rahmen der kognitionswissenschaftlichen Reflexion eines solchen makrokognitiven *Sensemaking*-Szenarios greift Tversky (1993) auf die *Theorie mentaler Modelle* (Johnson-Laird 1980) zurück – und postuliert die Existenz eines Qualitätsgefälles von mentalen Repräsentationen. Während sie den Begriff des „ *mentalen Modells*“ für Repräsentationen reserviert die ein hohes Maß an Kohärenz, Konnektivität und proportionaler Konsistenz ihrer Teile aufweisen, werden weniger kohärente Repräsentationen „ *kognitive Collagen“* genannt, die als schnelle und partielle Skizzen von komplexen Gegenständen meist fragmentarisch bleiben und die verschiedene Referenzpunkte nur unvollständig und in verzerrter Form verknüpfen. In dieser Gegenüberstellung besitzen kognitive Collagen zwar den pragmatischen Vorteil der Schnelligkeit (*good-enough representations*), aber nur mentale Modelle erlauben dank ihrer Eigenschaften der Kohärenz und Proportionalität anspruchsvollere kognitive Anschlussoperationen, wie perzeptive und konzeptuelle Schlussfolgerungen, globale Bewertung lokaler Einsichten, sowie die Erschließung bislang unbekannter Perspektiven auf den Gegenstand. *Kohärenz* ist vor diesem Hintergrund als Desiderat von mentalen Repräsentationen zu betrachten, das gemäß den Ansätzen der distribuierten Kognition (Scaife / Rogers 1996; Liu et al. 2007; Patterson et al. 2014) aber nur dann erzielt werden kann, wenn auch schon bei *externen* Repräsentationen (InfoVis Interfaces) ein ausreichendes Maß an Kohärenz gegeben ist.

Die Bedeutung dieser Diskussion für die zukünftige Entwicklung von Visualisierung und Modellierung in den Digital Humanities ergibt sich unmittelbar aus der multidimensionalen Natur der meisten ihrer Datenbestände. InfoVis-Interfaces mit „multiplen views“ sind in solchem Kontext als Standardtechnik unverzichtbar, doch nur wenig Aufmerksamkeit wurde bislang der Frage gewidmet, wie aus einer beziehungslosen Pluralität von Ansichten ein konzeptuell wohlvernetztes und kohärentes Ensemble geformt werden kann, das dabei hilft, diverse eigenlogische Ansichten kognitiv bestmöglich zu kohärenten mentalen Modellen zu verknüpfen. Da die Möglichkeiten solcher Verknüpfungen meist schon im Rahmen einzelner Forschungsprojekte unterentwickelt bleiben, wird die Verknüpfung von Visualisierungen verschiedener Forschungsgruppen oder Communities oft gar nicht erst versucht.

Um die kollektive Aufmerksamkeit verstärkt auf dieses Defizit zu lenken und ForscherInnen der Digital Humanities auch auf der Ebene makrokognitiver Operationen methodisch zu unterstützen, präsentiert und vernetzt der Beitrag eine Reihe von *Kohärenztechniken* zu einem methodischen Rahmenwerk, das dazu imstande ist, multiple Ansichten und ihre diversen Bildsprachen unter Erhalt ihrer jeweiligen Eigenlogiken effektiver zu vermitteln.

A) Methoden die die *initiale Konstruktion* von mentalen Modellen unterstützen: *Advance Organizer* veranschaulichen die Grundstrukturen eines Datensatzes, die eine erste konzeptuelle Orientierung erlauben und durch detaillierte Lernprozesse angereichert werden (Ausubel 1960). Als *Navigatoren* dienen diese Strukturmodelle der fortgesetzten Orientierung und Navigation zwischen multiplen Ansichten. Mit Blick auf zeitorientierte Daten wird das Potential von *Raum-Zeit-Kuben* demonstriert, die die Navigation zwischen dynamischen Datenprojektionen unterstützen (Bach et al. 2014), und die als *Multiple Raum-Zeit-Kuben* für

multidimensionale Datenbestände weiterentwickelt werden (Windhager 2013).

B) Methoden die die *sequentielle Integration* multipler Ansichten in kohärente mentale Modelle unterstützen: Bei der sequentiellen Nutzung multipler Ansichten dienen *Seamless Layout Transitions* dem Erhalt von bereits vorhandenen „mental maps“ (Freire / Rodríguez 2006) während *Seamless Canvas Transitions* (Federico et al. 2011) anschaulich die Unterschiede zwischen verschiedenen dynamischen Visualisierungstechniken vermitteln. Techniken der *Narrativen Visualisierung* verknüpfen schließlich multiple Ansichten über narrative Gestaltungselemente (Segel / Heer 2010), die die Integration durch ergänzende sequentielle Passagen erleichtern.

C) Methoden die die *parallele Integration* diverser Ansichten in kohärente mentale Modelle unterstützen umfassen verschiedene Typen von *Coordinated Multiple Views* (Roberts 2007; Sedlmair et al. 2009), durch die verschiedene InfoVis-Layouts, verschiedene zeitliche Selektionen, oder verschiedene Skalierungen in Beziehung gesetzt werden. *Koordinierte Interaktionsmethoden* wie *Linking & Brushing* erlauben die vertiefende synchronisierte Exploration und Integration multipler Ansichten. Techniken des *Visual Linkings* veranschaulichen darüber hinaus Entsprechungen von visuellen Elementen und Strukturen über unterschiedliche Ansichten hinweg (Collins / Carpendale 2007). *HyperImage*-Techniken (Warnke et al. 2007) ermöglichen zusätzliche Verlinkungen paralleler Ansichten – sowie entscheidende Verknüpfungen des lokalen Interfaces mit Bildern und Modellen außerhalb des Systems.

Der Mehrwert dieses Frameworks ergibt sich aus der erstmaligen Sammlung, Systematisierung und funktionalen Vernetzung von Kohärenztechniken, deren Entwicklung im Rahmen einzelner Applikationen zumeist zur Gänze vernachlässigt wird. Gerade ihre synergetische Erschließung scheint jedoch unverzichtbar, wenn es um zukünftige Systeme mit erhöhter analytischer Auflösung, gesteigerter Nutzerfreundlichkeit, und kohärenter visueller Syntax geht. Neben der praktischen Implementierung dieser Techniken betrachten wir ihre kognitionswissenschaftlich Reflexion und Evaluation als unerlässlich für die Entwicklung von InfoVis-Interfaces der nächsten Generation, die neben den Standardoperationen der visuellen Analyse auch makrokognitive Inferenz- und Syntheseprozesse unterstützen.

Zudem betrachten wir die kollektive Erkundung und Entwicklung von inter-piktorialen Kohärenztechniken als unmittelbare Voraussetzung für die verbesserte Vernetzung und Co-Konstruktion von geteilten mentalen Modellen (*shared mental models*) zwischen unterschiedlichen Communities (Swaab et al. 2002). Ein Ausblick richtet sich vor diesem Hintergrund auf die Möglichkeit, durch die verbesserte Vernetzung von visuellen Repräsentationen auch die Abstimmung von Diskursen und Disziplinen in den Digitalen Geisteswissenschaften besser zu koordinieren.

Bibliographie

Ausubel, David P. (1960): "The use of advance organizers in the learning and retention of meaningful verbal material", in: *Journal of Educational Psychology* 51: 267-272.

Bach, Benjamin / Dragicevic, Pierre / Archambault, Daniel / Hurter, Christophe / Carpendale, Sheelagh (2014): "A Review of Temporal Data Visualizations Based on Space-Time Cube Operations.", in: *EuroVis-STARs*: The Eurographics Association 23–41.

Collins, Christopher M. / Carpendale, Sheelagh (2007): "VisLink: Revealing relationships amongst visualizations", in: *Visualization and Computer Graphics, IEEE Transactions on* 13, 6: 1192-1199.

Federico, Paolo / Aigner, Wolfgang / Miksch, Silvia / Windhager, Florian / Zenk, Lukas (2011): "A Visual Analytics Approach to Dynamic Social Networks", in: Lindstaedt, Stefanie / Granitzer, Michael (eds.): *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW), Special Track on Theory and Applications of Visual Analytics (TAVA), Graz 2011*. New York: ACM 47: 1–47: 8.

Freire, Manuel / Rodríguez, Pilar (2006): "Preserving the mental map in interactive graph interfaces", in: *Proceedings of the working conference on Advanced visual interfaces*. New York: ACM 270–273.

Jänicke, Stefan / Franzini, Greta / Cheema, Muhammad Faisal / Scheuermann, Gerik (2015): "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges", in: *EuroVis-STARs*. The Eurographics Association.

Johnson#Laird, Philip N. (1980): "Mental models in cognitive science", in: *Cognitive science* 4, 1: 71-115.

Liu, Zhicheng / Nersessian, Nancy J. / Stasko, John T. (2008): "Distributed cognition as a theoretical framework for information visualization", in: *Visualization and Computer Graphics, IEEE Transactions on* 14, 6: 1173–1180.

Patterson, Robert E. / Blaha, Leslie M. / Grinstein, Georges G. / Liggett, Kristen K. / Kaveney, David E. / Sheldon, Kathleen C. / Haviga, Paul R. / Moore, Jason A. (2014): "A human cognition framework for information visualization", in: *Computers & Graphics* 42: 42–58.

Roberts, Jonathan C. (2007): "State of the art: Coordinated & multiple views in exploratory visualization", in: *Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV'07. Fifth International Conference on* IEEE: 61–71.

Scaife, Mike / Rogers, Yvonne (1996): "External cognition: how do graphical representations work?", in:

International Journal of Human-Computer Studies 45, 2: 185–213.

Sedlmair, Michael / Ruhland, Kerstin / Hennecke, Fabian / Butz, Andreas / Bioletti, Susan / O'Sullivan, Carol (2009): "Towards the big picture: Enriching 3d models with information visualisation and vice versa", in: *Proceedings of the 9th International Symposium on Smart Graphics, SG 2009, Salamanca, Spain, May 28-30, 2009*. Berlin / Heidelberg: Springer 27–39.

Segel, Edward / Heer, Jeffrey (2010): "Narrative visualization: Telling stories with data", in: *Visualization and Computer Graphics, IEEE Transactions on* 16, 6: 1139-1148.

Sula, Chris A. (2013): "Quantifying Culture: Four Types of Value in Visualisation", in: Bowen, Jonathan / Suzanne Keene / Ng, Kia (eds.): *Electronic Visualisation in Arts and Culture*. London: Springer 25–37.

Swaab, Roderick I. / Postmes, Tom / Neijens, Peter / Kiers, Marius H. / Dumay, Adrie C. (2002): "Multiparty negotiation support: The role of visualization's influence on the development of shared mental models", in: *Journal of Management Information Systems* 19, 1: 129-150.

Tversky, Barbara (1993): "Cognitive maps, cognitive collages, and spatial mental models", in: Frank, Andrew U. / Campari, Irene (eds.): *Spatial Information Theory*. A Theoretical Basis for GIS. Berlin: Springer 14-24.

Warnke, Martin / Kuper, Heinz-Günter / Helmers, Sabine (2007): "Hyperlmage. Bildorientierte e-Science-Netzwerke", in: *cms-Journal* 29, 80–84.

Windhager, Florian (2013): "On Polycubism. Outlining a Dynamic Information Visualization Framework for the Humanities and Social Sciences", in: Fuellsack, Manfred (ed.): *Networking Networks*. Origins, Applications, Experiments. Wien: Turia + Kant.

Netzwerkanalysen als Methode in der historischen Epistemologie

Wintergrün, Dirk
dwinter@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Deutschland

Valleriani, Matteo
valleriani@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Deutschland

Lalli, Roberto
rlalli@mpiwg-berlin.mpg.de
Max-Planck-Institut für Wissenschaftsgeschichte, Deutschland

Netzwerke und Historische Epistemologie

Methoden der Netzwerkanalyse kommen immer stärker als heuristisches Tool zum Einsatz, wenn es darum geht historische Prozesse zu beschreiben und zu analysieren. Netzwerktheorie stellt hierbei eine Möglichkeit dar, um systematisch Wissensstrukturen und die Formation und Transformation von Wissenssystemen zu beschreiben. Methodische Ansätze können hierbei von den Sozialwissenschaften, der Innovationsforschung sowie aus der Informatik und der Graphentheorie übernommen werden. Zunächst machten die meisten Ansätze in der historischen Forschung hauptsächlich von den qualitativen Konzepten gebrauch und weniger von den quantitativen Methoden, die die Netzwerktheorie liefert. Ein eindrucksvolles Beispiel zeigt sich im Werk von Irad Malkin in seinen Arbeiten über die griechische Antike (Malkin 2011). Mit der verbesserten Verfügbarkeit von Tools für die qualitative Analyse verändert sich diese Situation nun deutlich. Die Anzahl der Fallstudien hat mittlerweile die Größe erreicht, dass erste Theoretisierungen dieser Methode in der historischen Forschung unternommen wurden (van den Heuvel 2015). Große Impulse gehen von der sich um http://historicalnetworkresearch.org organisierenden Gruppe aus. Netzwerktheorie hat insbesondere ein großes Potential in der Wissenschaftsgeschichte insbesondere der historischen Epistemologie, die Entwicklung von Wissen als eine Verknüpfung von sozialen, materiellen und kognitiven Wissenssystemen sieht. Diese Systeme lassen sich als Netzwerke im Sinne der Netzwerktheorie verstehen. Dazu schlagen wir drei miteinander interagierende Netzwerke vor: ein soziales Netzwerk, ein semiotisches Netzwerk, sowie das darüber liegende epistemische Netzwerk: grob gesprochen ein Netzwerk von Akteuren, ein Netzwerk der Repräsentation von Wissen, das sich in materiellen Objekten oder auch kodifizierten Verfahren darstellt und schließlich das eigentliche kognitive Netzwerk. Wir werden in unserem Beitrag die ersten Ergebnisse zweier Fallstudien vorstellen. Dabei liegt unser Schwerpunkt weniger auf den konkreten Ergebnissen, sondern darauf zu zeigen, wie sich der Prozess der Übersetzung von historischen Fakten und Annahmen in netzwerktheoretisch auswertbare Daten darstellt und damit die Digital Humanities dazu beitragen, interdisziplinäre geisteswissenschaftliche Forschung zu ermöglichen.

Fallstudien

In der ersten Fallstudie gehen wir der Frage nach, wie sich eine bestimmte Wissenstradition in der frühen Neuzeit mittels gedruckter Traktate in der Zeit von 1472 bis in das Jahr 1650 in ganz Europa verbreiten konnte. Es geht hierbei um die mit der *Sphaera des Sacrobosco* (Thorndike 1949), einem ursprünglich handgeschriebenen grundlegenden Text, verbundenen Wissenstradition. Ursprünglich ein Text über Astronomie und Kosmologie wurde dieser im Laufe seiner Editionsgeschichte immer wieder durch zusätzliche Texte erweitert und umfänglich kommentiert. Es gehörte zum verbindlichen Wissenskanon der Universitäten dieser Periode. In der Zeit von 1472 bis 1650 haben wir bisher 363 Editionen identifiziert, die in ganz Europa veröffentlicht wurden. Welches sind die Voraussetzungen, die eine solche Verbreitung ermöglichten und welche Wissensinhalte wurden verbreitet und wie veränderten sich diese über die Zeit? Wir sehen in der Netzwerktheorie einen vielversprechenden Ansatz diese Fragen zu klären. Dazu haben wir ein erstes Netzwerk von wesentlichen Akteuren, den Verlegern, identifiziert, die maßgeblich für die Verbreitung des Traktats waren. Von welcher Form die Interaktionen zwischen den Verlegern waren, können wir bisher nur in sehr begrenztem Rahmen sicher bestimmen. Gleiches gilt für Ihre Rolle in den lokalen Wissensnetzwerken, in die sie eingebunden sind. Es ist jedoch möglich, Hypothesen über die Einflussbereiche aufzustellen, die die Kanten in dem Netzwerk rechtfertigen. Diese Hyptohesen lassen sich mit Netzwerktools in unserem Falle durch den Einsatz von Skripten in iPython (cf. Pérez / Granger 2007) unter Benutzung der Pakete networkx (cf. NetworkX developer team 2014) und graph-tool (cf. Peixoto) testen und modifizieren, in unserem Beitrag werden wir die von uns angewandten Verfahren darstellen. Die Skripte werden in Zukunft auf der Webseite des Projektes veröffentlicht werden. Neben den Verlegern gibt es ein weiteres Netzwerk, das eigentliche Netzwerk der Publikationen besser der Publikationsereignisse, diese stellen im Rahmen der angerissenen Theorie der Wissenssysteme eine andere Kategorie dar, sie sind materieller Ausdruck von Wissen, gehören damit zum semantischen Wissenssystem. Auch hier stellt sich die Frage, wie diese Editionen miteinander in Beziehung stehen. Offensichtlich stehen diese Beziehungen in enger Verbindung mit dem Akteursnetz. Jedoch gibt es auch andere Beziehungen, wie zum Beispiel die in den Editionen enthaltenen zusätzlichen Traktate und Kommentare, oder Abweichungen von Vorgängern. Auch hier sind wir bisher nur in der Lage Hypothesen anzustellen, aber auch hier zeigen uns Methoden der Netzwerktheorie bereits jetzt Wege auf, wie diese sich überprüfen lassen. Schließlich ist die entscheidende Frage: Welches Wissen wird durch die Traktate in welcher Form vermittelt und wie trägt dieses zur Wissensorganisation in der frühen Neuzeit bei? Wie oben beschrieben war das Traktat über die Sphaera ursprünglich ein Traktat über Astronomie und Kosmologie, die Ergänzungen umfassen jedoch einen wesentlich erweiterten Wissensbereich. Die Themenfelder, um die das Traktat erweitert wurden, umfassen Felder der mathematischen Astronomie, Kalenderberechnungen, die Benutzung und in einigen Bereichen konkrete Anleitung für die Konstruktion von astronomischen Instrumenten, nautische Astronomie und Geographie, Kartographie, Meteorologie, Arithmetik, Geometrie, der Konstruktion und des Gebrauchs von mathematischen Instrumenten zur Ausführung arithmetischer Berechnungen, Astrologie, Literatur, angewandte Optik und Mechanik. Die Antwort auf die Frage, welche Inhalte an welchen Stellen in die einzelnen Editionen eingeflossen sind, gibt Aufschlüsse über die Veränderung der frühneuzeitlichen Wissensstruktur. Auch hier kann Netzwerktheorie nach unserer Überzeugung einen wesentlichen Beitrag dazu leisten, schlüssige Antworten aufzufinden. Wir untersuchen dazu welche Ko-Autoren in welchen Traktaten genannt werden, insbesondere Clavius und Pedro Nuñes, und wie sich diese Subnetze ausprägen. Diese Koautoren stehen jeweils für bestimme Wissensgebiete geben also ein Indiz dafür welche Wissensbereich sich neu etablierten. Auch für diese Analysen haben wir Skripte entwickelt, die helfen diese Netzwerke zu visualisieren und deren Charakteristika zu untersuchen. Erste Ergebnisse der Netzwerkanalyse lassen Rückschlüsse auf die Stabilität des Wissensnetzwerkes zu. Wir können im wesentlichen drei Phasen deutlich identifizieren: eine noch instabile radiale Verbreitung, eine relative lange Phase der Stabilität und schließlich eine Phase der Reduktion.

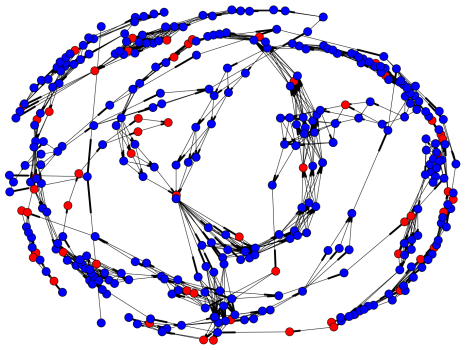


Abb. 1: Minimaler verbundener Graph der Editionen der Sphaera. Kanten spiegeln hier die chronologische Ordnung wieder. Dieser Graph legt die Randbedingungen für die weiter Analyse der Einflussbereiche der Editionen fest. Blaue Punkte stellen die lateinischen Editionen, rote die Editionen in lokalen Sprachen dar.

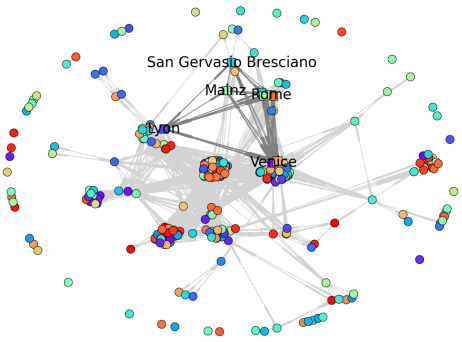


Abb. 2: The graph shows the role of editions where Clavius is given as one co-author (squares). The color indicates the publishers which are clustered by city. Only the places where Clavius is a co-author have labels. The big unnamed cluster in the middle is Paris.

Die zweite Fallstudie widmet sich einem anderen Bereich. Dem epistemischen Netzwerk der Allgemeinen Relativitätstheorie. Kurz umrissen geht es hierbei um die Analyse des Prozesses, der dazu führte, dass die ART in der Nachkriegsphase von einem Randproblem zu einem zentralen Ankerpunkt der theoretischen Physik wurde. Dieser Prozess wurde auch als "Renaissance der allgemeinen Relativität" (Will 1983) bezeichnet. Aus einem zersplitterten lose gebundenen Netzwerk von Einzelpersonen wird in dieser Zeit ein stabiles Netzwerk, dass von Institutionen getragen wird. Darauf aufbauend entwickelt sich ein Netzwerk materieller Kommunikation in Form von neuen Journalen und Konferenzreihen, sowie eine neue disziplinäre Sprache, die das Wissen über die ART kodifizierte. Die ART entwickelte sich so zur unangefochtenen Theorie von Raum und Zeit (Blum 2015). Zum jetzigen Zeitpunkt haben wir ein Netzwerk aus über 500 Akteuren identifiziert und ihre Beziehungen klassifiziert. Mit Methoden der Netzwerkanalyse können wir zeigen, welche Akteure über die Zeit an Bedeutung gewonnen und verloren und die Formation von Subclustern zeigen. Auch hier werden wir in unserem Beitrag im wesentlichen darauf eingehen, wie der konkrete Prozess der Umwandlung von historischen Fakten und Hypothesen in netzwerktheoretische Konzepte als Anwendung der DH in der historischen Forschung vollzogen wird. Insbesondere werden wir auf die Überlegungen und Schwierigkeiten eingehen, sinnvolle Verknüpfungen zwischen den Akteuren zu definieren und insbesondere zu Gewichten, d. h. in der Sprache der Netzwerktheorie, weak and strong ties zu idenfizieren. Von besonderer Bedeutung ist auch hier die Frage des Umgangs mit der dynamischen Entwicklung des Netzwerkes. Wie können diese so visualisiert werden, dass der historisch Forschende daraus Rückschlüsse ziehen kann, wie können diese Tools so zur Verfügung gestellt werden, dass sie auch direkt durch den Forscher nutzbar sind? In unserem Fall haben wir dazu unterschiedliche Ansätze verfolgt, erneut die Umsetzung mittels iPython, und die Visualisierung dann in Gephi und Cytoscape.

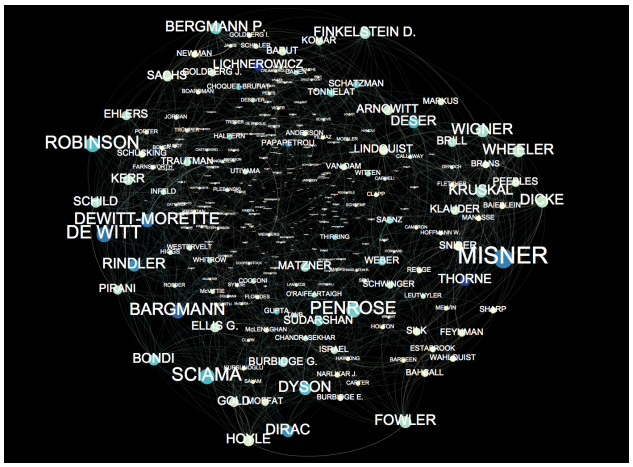


Abb. 3: Visualization of the network of collaborations of scientists who worked on general relativity in the post-war period. Node size is proportional to the degree centrality. The color of the nodes is a representation of betweenness centrality (from lighter to darker).

Bibliographie

Drucker, Johanna (2013): "Performative Materiality and Theoretical Approaches to Interface", in: *Digital Humanities Quartely* 7,1 http://www.Digitalhumanities.org/dhq/vol/7/1/000143/000143.html [letzter Zugriff 03. September 2015].

Edwards, Charlie (2012): "The Digital Humanities and Its Users", in: Gold, Matthew K. (ed.): *Debates in the Humanities* http://dhdebates.gc.cuny.edu/debates/text/31 [letzter Zugriff 03. September 2015].

Kirschenbaum, Matthew (2008): *Mechanisms*. New Media and the Forensic Imagination. Cambridge: MIT University Press.

Lauer, Gerhard (2013): "Die digitale Vermessung der Kultur. Geisteswissenschaften als Digital Humanities", in: Geiselberger, Heinrich / Moorstedt, Tobias (eds.): *Big Data*. Das neue Versprechen der Allwissenheit. Berlin: Suhrkamp.

Manovich, Lev (2001): *Language of New Media*. Cambridge: MIT Press.

Warwick, Claire (2012): "Studying users in digital humanities, in: Warwick, Claire / Terras, Melissa / Nyhan, Julianne (eds.): *Digital Humanities in Practice*. London: Facet Publishing.

dariahTeach - Freizugängliche Plattform für DH Lehrmaterialien

Wissik, Tanja
Tanja.Wissik@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Durco, Matej
Matej.Durco@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

In den letzten Jahren ist auf Grund der steigenden Sichtbarkeit der Digital Humanities auch die Frage der Vermittlung von Digital Humanities, im universitären und außeruniversitären Bereich, immer wichtiger geworden (vgl. Thaller 2015). Dies zeigen die steigende Anzahl an Studiengängen für Digital Humanities in der einen oder anderen Form an Universitäten (vgl. *Digital Humanities Course Registry*, Schmeer 2014), die Liste von DH-Studiengängen im deutschsprachigen Raum (vgl. Sahle 2011) und Sahle (2013) , v. a. mit der im Anhang befindlichen Auflistung von DH-BA- und MA-Studiengängen, außrdem Initiativen und Arbeitsgruppen zu Referenzcurricula (vgl. z. B. Thaller 2015) sowie Vorträge und Artikel zum Thema Lehre in den Digital Humanities oder DH Curricula (vgl. z. B. Sula et al. (2015); Thaller (2015); die Ausgabe "Digital Humanities Pedagogy" der Zeitschrift *CEA*, vgl. Iantorno 2014).

Die meisten dieser Vorhaben sind stark auf den universitären Kontext ausgerichtet. Sula et al. (2015) untersucht 34 universitäre DH Curricula in den USA, Australien, Kanada und UK. Die DHd Arbeitsgruppe „Referenzcurriculum Digital Humanities“ beleuchtet Fallbeispiel von DH Curricula an Universitäten im deutschsprachigen Raum. Auch in der Dariah *Digital Humanities Course Registry* sind mehr Studiengänge als Weiterbildungskurse wie z. B. Summer Schools aufgelistet (vgl. Schmeer 2014). Auch auf der *coursera* Plattform für kostenlose Online-Kurse gibt es unter dem Suchbegriff „Digital Humanities“ nur zwei Treffer (Coursera 19.08.2015).

Das dariahTeach Projekt, ein Projekt im Rahmen der Erasmus+ Strategische Partnerschaften Förderung, hat sich zum Ziel gesetzt, modulare Lehrmaterialien für die Digital Humanities zu entwickeln, die in einer Vielzahl von Szenarien – in universitären Studiengängen aber vor allem auch in Workshops bis hin zu informellen individuellen Lernsituationen – eingesetzt werden können und allen Interessierten über eine spezielle Open-Access und Open-Source-Web-Plattform zugänglich sein werden.

Das Poster wird einen Überblick über das dariahTeach Projekt und seine Ziele geben. Die bereits zur Verfügung stehenden Projektresultate werden vorgestellt unter anderem der Prototyp der Plattform zur Bereitstellung der Lehrmaterialien mit einigen exemplarischen Lehrmaterialien. Die freizugängliche Plattform wird nach den speziellen Anforderungen der zukünftigen User entwickelt sowie auf die unterschiedlichen Lehr- und Lernsituationen abgestimmt.

Das dariahTeach Projektkonsortium, bestehend aus *Maynooth University* (Koordination), *Aarhus Universitet*, *Athena Research and Innovation Center in Information Communication & Knowledge Technologies*, *Belgrade Centre for Digital Humanities*, *Erasmus Universiteit Rotterdam*, *Österreichische Akademie der Wissenschaften* und *Université de Lausanne* erarbeitet zumindest 5 Kernmodule zu den folgenden Themen: Einführung in Digital Humanities; Textkodierung; Multiliteralität und audio-visuelle Medien; Ontologien & Wissensrepräsentation; Retrodigitalisierung von Wörterbüchern.

Neben den englischen Lehrmaterialien werden auch Versionen, in den Sprachen der Partnerländer, die diese Module erstellen, vorliegen. dariahTeach wird eine Sammlung von mehrsprachigen Lehrmaterialien für Digital Humanities bereitstellen um mehrsprachige Aus- und Weiterbildung in den Digital Humanities zu stärken und um die Internationalisierung und Mehrsprachigkeit in den Digital Humanities zu fördern.

Die Web-Plattform wird auch nach dem Projektende weiter im Rahmen der DARIAH Infrastruktur zugänglich sein und es wird möglich sein, fortlaufend Material auf der Plattform bereitzustellen oder bereits bestehende Inhalte zu aktualisieren oder mehrsprachige Lehrmaterialien hinzuzufügen.

Bibliographie

Coursera (19. August 2015): *Coursera Inc.*. Mountain View, CA, USA https://www.coursera.org/ [letzter Zugriff 09. Januar 2016].

Schmeer, Hendrik (2014): *Digital Humanities Course Registry* https://dh-registry.de.dariah.eu/ [letzter Zugriff 08. Januar 2016].

Iantorno, Luke A. (2014): "Introducing Digital Humanities Pedagogy", in: *CEA Critic* 76, 2: Special Issue: Digital Humanities Pedagogy 140-146.

Sahle, Patrick (ed.) (2011): *Digitale Geisteswissenschaften*. Cologne Center for eHumanities (CCeH), Universität zu Köln http://www.cceh.uni-koeln.de/Dokumente/BroschuereWeb.pdf [letzter Zugriff 09. Januar 2016].

Sahle, Patrick (2013): *DH Studieren! Auf dem Weg zu einem Kern- und Referenzcurriculum der Digital Humanities* (= DARIAH-DE Working Papers 1). Georg-August-Universität Göttingen: GOEDOC http://resolver.sub.uni-goettingen.de/purl/?dariah-2013-1 [letzter Zugriff 09. Januar 2016].

Sula, Chris Alen / Cunningham, Phillip / Hackney, Sarah (2015): *A Survey of DH Curricula at the Present Time. Keystone Digital Humanities Conference, Kislak Center for Special Collections*. Philadelphia, USA, July 22–24, 2015 http://sceti.library.upenn.edu/KeystoneDH/abstracts.html [letzter Zugriff 09. Januar 2016].

Thaller, Manfred (2015): "Panel: Digital Humanities als Beruf – Fortschritte auf dem Weg zu einem Curriculum", in: *Digital Humanities als Beruf*. Fortschritte auf dem Weg zu einem Curriculum. Akten der Dhd Arbeitsgruppe „Referenzcurriculum Digital Humanities“, Jahrestagung 2015, Graz 3-5.

Gegenwärtige dialektspezifische Daten und deren Anwendung in der Dialektometrie

Zhekova, Desislava
desi@cis.uni-muenchen.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

Krefeld, Thomas
thomas.krefeld@lmu.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

Herteis, Simeon
simeon.herteis@gmail.com
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

Einleitung

Die Datenverarbeitung innerhalb der Geisteswissenschaften ist sehr eng mit den gegenwärtigen technologischen Entwicklungen verbunden und dementsprechend auch stark davon abhängig. Ein sehr gutes Beispiel dafür ist das Gebiet der Dialektologie / Dialektometrie. Klassische Dialektometrie ist eine Forschungsrichtung innerhalb der Linguistik, die sich mit der Erforschung möglichst hochrangiger Ordnungsstrukturen in sprachgeographischen Netzen beschäftigt. Diese Aufgabe wurde bislang hauptsächlich durch die Analyse gesprochener Sprache (z. B. akustische Aufnahmen) oder der sogenannten Fragebögen (z. B. gezielt abgefragte, schriftliche Daten) bewältigt. Ein Nachteil dieser ist allerdings, dass die erhobenen Daten stark beeinflusst oder nicht schriftlich sind. Durch die gegenwärtigen Entwicklungen in der Informationstechnologie sind Sammlungen von neuartigen Dialektdaten erreichbar (die ohne äußeren Einfluss, gesammelt wurden und darüber hinaus in schriftlicher Form als Datensatz vorhanden sind), womit in der Dialektometrie neue Wege gegangen werden können. Ein Beispiel dafür sind neue Medien, wie z. B. Wikipedia, Twitter, digitale Zeitschriften, etc., in denen außerdem Veränderungen in der Gesellschaft schnell abgebildet werden.

Allein in Wikipedia ist eine große Anzahl an Dialekten vertreten, wie zum Beispiel die italienischen Dialekte Lombardisch (31.986 Artikel) [1] , Sizilianisch (25.273 Artikel), Neapolitanisch (14.346 Artikel) etc., die fortlaufend mit neuen Artikeln erweitert werden, die nicht nur von einem, sondern von mehreren Autoren editiert werden. Aus diesen Artikeln kann eine bisher nicht vorhandene Art Korpus erstellt werden, dessen Untersuchung die Beantwortung völlig neuer Fragestellungen möglich werden lässt.

Die Größe dieser neuen Korpora ermöglicht nicht nur neuartige Fragestellungen in der Dialektometrie, sondern auch einen zeitgenössischen und automatisierten Vergleich für die Analyse von Dialekten und ihren linguistischen Eigenschaften (basiert auf statistische Ansätze). Für solche Verfahren ist allerdings nicht nur die vorhandene Datenmenge wichtig, sondern auch die leichte Erreichbarkeit von qualitativen Annotationen und Analysetools. Diese wurden bislang hauptsächlich für die Standardsprachen entwickelt, für Dialekte existieren diese bis jetzt nur in wenigen Ausnahmefällen.

Ein solches Analysetool für die Standardsprache Italienisch ist AnIta (Tamburini / Melandri 2012), ein morphologisches Finite-State-Analysetool, welches bisher nur für das Italienische verwendet werden kann. In AnIta können aber auch viele empirische Belege für Dialekte integriert werden, sodass die maschinelle Bearbeitung vieler italienischer Dialekte möglich wird. Die neuen Dialektwikipedias ermöglichen auch einen halb automatisierten Ansatz dafür.

SiMoN

U#berblick

In unserer Softwaredemonstration mo#chten wir eine vorla#ufige Erweiterung von AnIta vorstellen, die mit vielen regelma#ßigen Verbparadigmen des sizilianischen Dialekts erweitert wurde - SiMoN (Sizilianische Morphologie fu#r NLP-Anwendungen). Die Version der Softwaredemonstration ist schon online erreichbar. Aus Eintra#gen der sizilianischen Wikipedia wurden Verblemmata (368 sizilianische Lemmata) fu#r das Lexikon von AnIta automatisch extrahiert anhand von dem Auftreten regula#ren sizilianischen Verbendungen und einer Liste von Verben im Italienischen. Da sich die Verben des Sizilianischen in nur zwei Typen aufteilen

(statt wie im Italienischen in drei), sind nur Verbeinträge mit Endungen auf *-ari* und auf *-iri* vorhanden. Die gesamte Zahl, der durch Flexionsparadigmen erfassten Verbformen beläuft sich auf ca. 24.700. Damit bietet SiMoN einen ersten Grundstock für die Entwicklung einer computergestützten, sizilianischen Morphologie.

Dokumentierte Paradigmen

Der Fokus der zu untersuchenden Paradigmen liegt in dieser Arbeit auf den Konjugationsmustern regelmäßiger Verben. Das vorderste Ziel ist es hier, eine Grundlage für die Verbanalyse für Sizilianisch zu schaffen. Im Gegensatz zum Italienischen gibt es für einige Verben eine große Zahl an Wahlmöglichkeiten für Endungen konjugierter Formen, die regional unterschiedlich verbreitet und gleichermaßen gültig sind. Bonner und Cipolla (2001) dokumentieren für die regelmäßigen Verben einiger Zeiten und Modi alternative Formen, die wir verfolgen. Diese Alternativformen gehören alle zum selben Paradigma. Daher gibt es im jeweiligen Lexikon der beiden Verbtypen in SiMoN teilweise mehrfache Einträge zur Konjugation der ersten, zweiten oder dritten Person. Eine vorläufige Analyse des gewonnenen Wikipedia-Korpus zeigte ebenfalls, dass die verschiedenen Varianten der Verben in der Praxis verwendet werden. Stammveränderungen in der sizilianischen Verbgrammatik existieren ebenfalls, diese Fälle werden allerdings mit SiMoN im Moment noch nicht abgedeckt.

	Indikativ: Präsens				Indikativ: Imperfekt			
	parrari		*battiri*		*parrari*		*battiri*	
	Konj.	Var.	Konj.	Var.	Konj.	Var.	Konj.	Var.
1s	parru	-	batti	-	parrava	-avu	battia	-iu -eva -evu -iva -ivu
2s	parri	-	batti	-	parravi	-	battivi	-evi
3s	parra	-	batta	-	parrava	-	battia	-eva -iva
1p	parramu	-	battemu	-	parràvamu	-	batta	-evamu ivamu
2p	parrati	-	battiti	-	parràvavu	-	battiavu	-evavu -ivavu
3p	parranu	-unu	battinu	-unu	parràvanu	-àvunu	battianu	-evanu -ivanu

	Indikativ: Partizip Perfekt				Indikativ: Präteritum				Imperativ	
	aviri + parrari		*aviri + battiri*		*parrari*		*battiri*		*parrari*	*battiri*
	Hilfsverb	Part.	Hilfsverb	Part.	Konj.	Var.	Konj.	Var.	Konj.	Konj.
1s	aiu		aiu		parrai	-avi -aiu -avu	battivi	-ii -iu -ivu	-	-
2s	ai		ai		parrast	-	battisti	-	parra	batti
3s	avi	parratu	avi	battutu	parrau	-ò	battiu	-	parrassi	battissi
1p	avemu		avemu		parrammu	-amu	battemmu	-emu	parramu	battemu
2p	aviti		aviti		parrastivu	-astu	battistivu	-astu -istu	parrati	battiti
3p	annu		annu		parrarunu	-aru	batterunu	-eru	parrassiru	battissiru

	Gerundium				Imperfekt				Futur	
	stari + parrari		*stari + battiri*		*parrari*		*battiri*		*parrari*	*battiri*
	Hilfsverb	Ger.	Hilfsverb	Ger.	Konj.	Var.	Konj.	Var.	Konj.	Konj.
1s	staiu		staiu		parrassi	-	battissi	-	parrirò	battirò
2s	stai		stai		parrassi	-	battissi	-	parrirai	battirai
3s	sta	parrennu	sta	battennu	parrassi	-	battissi	-	parrirà	battirà
1p	stamu		stamu		parrassimu	-	battissimu	-	parriremu	battiremu
2p	stai		stai		parrassivu	-	battissivu	-	parririti	battiriti
3p	stannu		stannu		parrassiru	-assinu	battissiru	-issinu	parrirannu	battirannu

Tabelle 1: Die regelma#ßigen Konjugationsformen, die in SiMoN integriert wurden.

In Tabelle 1 sind die regelmäßigen Konjugationsformen (die in SiMoN vorhanden sind) am Beispiel der sizilianischen Verben *parrari* (Deutsch - reden) und *battiri* (Deutsch - schlagen) aufgeführt. Die Formen beider Verbtypen in den Flexionskategorien Indikativ, Imperativ und Subjunktiv, sowie Konditional und Gerundium sind jeweils vorhanden. Die Paradigmen der unregelmäßigen Hilfsverben *essiri* (Deutsch - sein) und *aviri* (Deutsch - haben) sowie das sehr häufig verwendete *fari* (Deutsch - machen) wurden ebenfalls in SiMoN in die Liste der Lemmata aufgenommen, um Partizipkonstruktionen u. ä. zu erkennen.

Ausblick

Unserer Ziel ist vorerst anhand den Texten der Wikipedia für Standard Italienisch und alle andere Dialektwikipedias weiterhin automatisch dialektspezifische Verben zu extrahieren und damit SiMoN zu erweitern. Damit können zusätzliche Dialekte auch behandelt und entwickelt werden. SiMoN würde dann eine automatisierte morphologische Analyse für reguläre italienische Dialektparadigmen ermöglichen, was wir bis jetzt nur für Sizilianisch anbieten können. Weiterhin ist es geplant auch irreguläre Dialektparadigmen manuell zu integrieren.

Notes

1. Die Zahlen sind von Wikipedia entnommen worden (Stand: August 2015).

Bibliographie

Bonner, J. K. "Kirk" / Cipolla, Gaetano (2001): *Introduction to Sicilian Grammar*. Brooklyn, NY: Legas.

Tamburini, Fabio / Melandri, Matias (2012): „AnIta: A Powerful Morphological Analyser for Italian“, in: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey 941-947.

Kollaboratives Schreiben gestern, heute und morgen: Nutzen und Grenzen eines Visualisierungs- und Analysemodels aus der digitalen Literaturforschung

Zimmermann, Heiko
heiko.zimmermann@uni-trier.de
Universität Trier, Deutschland

"Journale sind eigentlich schon 'gemeinschaftliche' Bücher. Das Schreiben in Gesellschaft ist ein interessantes Symptomn — das noch eine große Ausbildung der Schriftstellerey ahnden läßt. Man wird

vielleicht einmal in 'Masse' schreiben, denken, und handeln — Ganze Gemeinden, selbst Nationen werden Ein Werck unternehmen." (Novalis 1965: 645).

Sieht man vom gemeinschaftlichen Schreiben aktueller 'Wirklichkeiten' in Facebook ab, ist Novalis' Prophetie weit entfernt von kreativen Schreibprozessen der Gegenwart. Dennoch hat es unterschiedlichste Formen kollaborativen Schreibens seit jeher gegeben. Dieses Schreiben hat die Literaturkritik und -wissenschaft oft vor Probleme gestellt (vgl. Ede / Lunsford 1990). In den letzten Jahren haben die Möglichkeiten des vernetzten Schreibens am Computer neue Formen und Dimensionen kollaborativer Schreibenprozesse gefördert. Werke wie die Enzyklopädie *Wikipedia*, den aus Fanfiction entstandenen Bestseller-Roman *Shades of Grey*, das multimediale Universum um die MTV-Serie *Teenwolf* oder auch das bisher größte digitale Romanprojekt *A Million Penguins* wären ohne den Rechner im Internet unmöglich gewesen.

Zur selben Zeit ist in der englischsprachigen Welt das Genre der digitalen Literatur aufgekommen, welches die Literaturwissenschaft ebenfalls vor große Herausforderungen stellt. Ein Hauptproblem ist das der Rekonfigurationen von Autor- und Leserschaft, das mittels poststrukturalistischer Metaphern (Landow 2006; Simanowski 2002; Winko 1999) nicht hinreichend beschrieben werden konnte. Auch Zwischenwesen wie das Modell des Wreaders, also des schreibenden Lesers, konnten die Abweichungen von tradierten Rollen in der Literaturproduktion und -rezeption nicht sinnvoll modellieren. Aus den selben Gründen funktionieren auch buchgeschichtliche Modelle des Literaturmarktes wie das von Robert Darnton (1982) nur bedingt, um die Wirklichkeit digitaler Literatur zu beschreiben.

Um das Problem der Autor- und Leserschaft und unzureichender tradierter Modellierungen zu lösen, wurde das visuelle Beschreibungs- und Analysemodell des textuellen Handlungsraums entwickelt (Zimmermann 2015a). Es basiert auf dem Texton-Skripton-Modell von Espen Aarseth (1997: 62-65) und ordnet allen am Text handelnden Akteuren einen eindeutigen Platz im Handlungsraum zu, der abhängig ist von der Art und Weise und vom Zeitpunkt ihres Handelns am Text im Kontinuum von Produktion und Rezeption (vgl. Abbildung 1). Anwendungen dieses Modells waren bisher auf englische digitale Literatur beschränkt und haben in diesem Feld ergeben, dass es bestimmte Konstellationen von Handelnden in der Literaturproduktion und -rezeption, beispielsweise Foucaults Idee einer beherrschenden Stellung der Autorfunktion in literarischen Diskursen, in Frage stellt (Zimmermann 2015b).

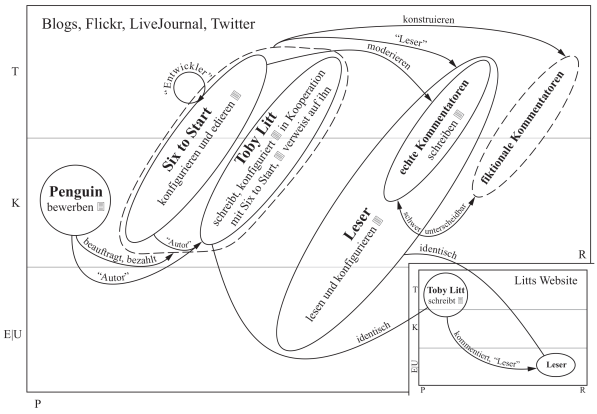


Abb. 1: Beispiel eines TeKEU-Diagramms eines textuellen Handlungsraums: Toby Litts Werk *Slice* (Zimmermann 2015a: 186).

Der vorgeschlagene Vortrag soll gleichsam mehrere Brücken zwischen akademischen Disziplinen und literarischen Traditionen schlagen. Das Modell des textuellen Handlungsraums, das aus dem Feld der elektronischen Literaturforschung - und damit aus einem Kerngebiet der digitalen Geisteswissenschaften, sofern diese nicht allein über Methoden und Werkzeuge definiert werden - stammt, soll nicht nur auf digitale englischsprachige Literatur angewendet werden, sondern auch auf nicht-digitale deutsche und englischsprachige literarische Texte der Gegenwart und des 20. Jahrhunderts. Damit werden Verbindungen zwischen verschiedenen Literaturen (zeitlich, sprachlich) und akademischen Feldern (digitale Literaturforschung, digitale Geisteswissenschaften, traditionelle Literaturwissenschaft) hergestellt.

Nachdem das Modell im Vortrag kurz vorgestellt wurde, fragt dieser nach den Formen von Autor- und Leserschaft ausgewählter kollaborativ geschriebener Texte, nach Möglichkeiten solches Schreiben sinnvoll zu klassifizieren und danach, ob sich Rückschlüsse auf die (kommerzielle) Verwertbarkeit ebendieser Literatur ziehen lassen. Flankierend wird damit eine Fallstudie für die Permeabilität traditioneller Literaturanalyse für Modelle aus dem Bereich der digitalen Literaturwissenschaft vorgestellt, und es werden die Potentiale und Grenzen einer derartigen Visualisierung literarischen Schaffens aufgezeigt.

Bibliographie

Aarseth, Espen J. (1997): *Cybertext*. Perspectives on Ergodic Literature. Baltimore: Johns Hopkins UP.

Cumming, Charles (2008): *The 21 Steps*, in: Six to start (ed.): *We Tell Stories: Six Authors, Six Stories, Six Weeks*.
http://www.wetellstories.co.uk/stories/week1/ [letzter Zugriff 12. Februar 2016].

Darnton, Robert (1982): "What Is the History of Books?" in: *Dædalus* 111, 3: 65-83.

Dilke, Christopher / Forster, Edward M. / Coppard, A. E. / Laver, James (1975): "Three Courses and a Dessert: Being a New and Gastronomic Version of the Old Game of Consequences" in: Forster, Edward Morgan: *The Life to Come and other Stories*. Harmondsworth: Penguin 235-74.

Ede, Lisa S. / Lunsford, Andrea A. (1990): *Singular Texts / Plural Authors*. Perspectives on Collaborative Writing. Carbondale: Southern Illinois UP.

Flores, Leonardo Luis (2010): *Typing the Dancing Signifier: Jim Andrew's (Vis)Poetics*. Diss. University of Maryland. Digital Repository at the University of Maryland http://drum.lib.umd.edu/bitstream/handle/1903/10799/Flores_umd_0117E_11445.pdf?sequence=1&isAllowed=y [letzter Zugriff 12. Februar 2016].

Foucault, Michel (2003): "Was ist ein Autor?", in: Jannidis, Fotis / Lauer, Gerhard / Martinez, Matias / Winko, Simone (eds.): *Texte zur Theorie der Autorschaft*. Stuttgart: Reclam 198-229.

Hettche, Thomas / Hensel, Jana (2000): *Null: Literatur im Netz*. Ostfildern: DuMont.

James, E. L (2011): *Fifty Shades of Grey*. London: Vintage.

Krajewski, Markus (2001): "Ver(b)rannt im Fahlen Feuer: Ein Karteikartenkommentar", in: *Der gerissene Faden: Nichtlineare Techniken in der Kunst*. Themenheft von *Kunstforum International* 155: 288-92.

Landow, George P. (1994): *Hyper / Text / Theory*. Baltimore: Johns Hopkins UP.

Landow, George P. (2006): *Hypertext 3.0*. Critical Theory and New Media in an Era of Globalization. Baltimore: Johns Hopkins UP.

Le Lionnais, François (1984): "Über experimentelle Literatur" in: Queneau, Raymond / Le Lionnais, François (eds.): *Hunderttausend Milliarden Gedichte*. Frankfurt am Main: Zweitausendeins.

Litt, Toby (2008): *Slice*, in: Six to start: *We Tell Stories: Six Authors, Six Stories, Six Weeks*.

Manovič, Lev (2002): "Models of Authorship in New Media", in: Manovič, Lev: *Homepage* http://manovich.net/index.php/projects/models-of-authorship-in-new-media [letzter Zugriff 12. Februar 2016].

Mason, Bruce / Thomas, Sue (2008): *A Million Penguins Research Report*. Institute of Creative Technologies, De Montfort University http://www.ioct.dmu.ac.uk/documents/amillionpenguinsreport.pdf [letzter Zugriff 12. Februar 2016].

Nabokov, Vladimir (2000): *Pale Fire*. London: Penguin.

Novalis (1965): *Schriften*. Herausgegeben von Paul Kluckhohn und Richard Samuel. Bd 2: Das philosophische Werk I. Darmstadt: WBG.

The Penguin Blog (2006): *The Official Blog of Penguin Books UK* http://penguinblog.co.uk/ [letzter Zugriff 17. Februar 2016].

Simanowski, Roberto (2000): "Einige Vorschläge und Fragen zur Betrachtung digitaler Literatur" in: Jäger, Georg / Simanowski, Roberto (eds.): *IASL Diskussionsforum online: Netzkommunikation in ihren Folgen*. LMU München.

Simanowski, Roberto (2001): "Autorschaft in digitalen Medien: Eine Einleitung" in: *Text + Kritik* 152: 3-21.

Simanowski, Roberto (2002): *Interfictions*. Vom Schreiben im Netz. Frankfurt am Main: Suhrkamp.

Stelling, Anke / Dannenberg, Robbie (2003): *Nimm mich mit*. Berlin: Fischer.

Winko, Simone (1999): "Lost in Hypertext? Autorkonzepte und neue Medien", in: Jannidis, Fotis / Lauer, Gerhard / Martinez, Matias / Winko, Simone (eds.): *Rückkehr des Autors*. Zur Erneuerung eines umstrittenen Begriffs. Tübingen: Niemeyer 511-33.

Zimmermann, Heiko (2015a): *Autorschaft und digitale Literatur*. Geschichte, Medienpraxis und Theoriebildung. Trier: WVT.

Zimmermann, Heiko (2015b): "Electronic Literature as a Means to Overcome the Supremacy of the Author Function.", in: *ELO 2015 - The End(s) of Electronic Literature*, Bergen, Norwegen.

Modellierung von Forschungsdaten durch Annotation

Zinsmeister, Heike

heike.zinsmeister@uni-hamburg.de
Universität Hamburg, Deutschland

Annotationen im Sinne von „Content" (vgl. Agosti / Ferro 2007) modellieren Forschungsdaten anhand dreier Dimensionen: der Konzepte, der Repräsentationsformate und zusätzlich auch durch den Annotationsprozess.

Der Schwerpunkt dieses Beitrags liegt auf der konzeptuellen Modellierung textueller Daten durch Abstraktionen und die Hervorhebung charakterisierender Eigenschaften bis hin zum Ergänzen assoziativer Bezüge. Der Beitrag spannt dabei einen Rahmen von intuitiven Randkommentaren (vgl. Blustein et al. 2011) bis hin zu wohldefinierten Annotationsprojekten (z. B. Stede / Neumann 2014) und gibt einen Ausblick auf die Umsetzung in Repräsentationsformaten (Dipper 2005; Piez 2011).

Das Besondere an der Modellierung eines Textes oder einer Textsammlung durch Annotation ist, dass die Auszeichnungen und Kommentare nicht einem holistischen Ganzen zugewiesen werden, wie es

normalerweise bei beschreibenden Metadaten der Fall ist, sondern dass sie mit bestimmten Bestandteilen des dekomponierten Textes verknüpft werden und damit eine gewissen Distribution im Text aufweisen, die der Modellierung grundsätzlich einen quantitativen Aspekt verleiht und für weiterführende Auswertungen und Visualisierungen verwendet werden kann.

In der geisteswissenschaftlichen Tradition bestehen unterschiedliche Zugänge zur Modellierung von Forschungsdaten durch Annotation. Beispielhaft stellt der Beitrag den prozessorientierten, hermeneutischen Ansatz der Literatur- und Kulturwissenschaft (Bradley / Vetch 2007, Bradley 2008) dem produktorientierten, deduktiven Ansatz der Korpus- und Computerlinguistik (Leech 1997, Pustejovsky / Stubbs 2012) gegenüber und schlägt eine Synthese der Modellierung in Anlehnung an die zusammenfassenden Darstellungen in Burghardt (2014) und Gius / Jacke (2015) vor. Zusätzlich gibt er einen Ausblick auf Repräsentationsmöglichkeiten von ambigen Annotationen (vgl. Barteld et al. 2014).

Bibliography

Agosti, Maristella / Nicola Ferro (2007): "A Formal Model of Annotations of Digital Content", in: *ACM Transactions on Information Systems* 26, 1.

Barteld, Fabian / Ihden, Sarah / Schröder, Ingrid / Zinsmeister, Heike (2014): „Annotating descriptively incomplete language phenomena", in: Levin, Lori / Stede, Manfred (eds.): *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop* 99–104.

Blustein, James / Rowe, David / Graff, Ann-Barbara (2011): „Making Sense in the Margins: A Field Study of Annotation", in: *Research and Advanced Technology for Digital Libraries*. Berlin: Springer 252–259.

Bradley, John (2008): "Pliny: A model for digital support of scholarship", in: *Journal of Digital Information (JoDI)* 9, 1.

Bradley, John / Vetch, Paul (2007): "Supporting Annotation as a Scholarly Tool # Experiences from the Online Chopin Variorum Edition", in: *Literary and Linguistic Computing* 22, 2.

Burghardt, Manuel (2014): *Engineering Annotation Usability-Toward Usability Patterns for Linguistic Annotation Tools*. Doktorarbeit. Universität Regensburg.

Dipper, Stefanie (2005): „XML-Based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation", in: *Proceedings der Berliner XML-Tage 2005 (BXML 2005)* 39–50.

Gius, Evelyn / Jacke, Janina (2015): „Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse", in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten der Digital Humanities* (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). http://www.zfdg.de/sb001_006 [letzter Zugriff 09. Januar 2016].

Leech, Geoffrey (1997): "Introducing Corpus Annotation", in: Garside, Roger / Leech, Geoffrey / McEnery, Tony (eds.): *Corpus Annotation*. Linguistic Information from Computer Text Corpora. London / New York: Longman 1–18.

Piez, Wendell (2010): „Towards Hermeneutic Markup. An Architectural Outline", in: *Digital Humanities 2010. Conference Abstracts, präsentiert auf der Digital Humanities Conference 2010 (DH 2010)* 202–205.

Pustejovsky, James / Stubbs, Amber (2012): *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.

Stede, Manfred / Neumann, Arne (2014): "Potsdam Commentary Corpus 2.0: Annotation for Discourse Research", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* 925–929.

Index der Autorinnen und Autoren